ISLA
LISBOA

8th TEACHING AND LANGUAGE CORPORA CONFERENCE

Lisbon, Portugal, 03-06 July 2008

Organizing Committee

Dr Ana Frankenberg-Garcia

Dr Tawfiq Rkibi

Ms Maria do Rosário Braga da Cruz

Mr Ricardo Carvalho

Ms Cristina Direito

Mr Diogo Santos-Rosa

Título

8th TEACHING AND LANGUAGE CORPORA CONFERENCE

Autor: Vários

Data: 2008

Editor

Associação de Estudos e de Investigação Científica do ISLA-Lisboa

ISBN: 978-989-95523-1-9

Depósito Legal: ******/**

Impressão: Offsetmais

PROGRAMME COMMITTEE

Austria	
Bernhard Kettemann	University of Graz
Hong Kong SAR, China	
Lynne Flowerdew	Hong Kong University of Science and Technology
Italy	
Guy Aston	University of Bologna
United Kingdom	
Chris Tribble	King's College London
Lou Burnard	Oxford University
Martin Wynne	Oford Text Archive
Poland	
Agnieszka Lenko-Szymanska	Warsaw University
Portugal	
Ana Frankenberg-Garcia	ISLA Lisbon

8TH TEACHING AND LANGUAGE CORPORA CONFERENCE
ISLA, 4th of July 2008

INTRODUCTION

Ana Frankenberg-Garcia

Teaching and Language Corpora (TaLC) conferences are now a well-established biennial event. TaLC began at the University of Lancaster sixteen years ago and, after being held there twice, and then successively at the universities of Oxford, Graz, Bologna, Granada, and Paris 7, we are delighted that, for the present occasion, we have been asked to bring the 8th Teaching and Language Corpora (TaLC 8) conference to the Instituto Superior de Línguas e Administração in Portugal.

The use of corpora in teaching has been growing steadily in the past couple of decades. This increased interest is reflected in the 110 proposals from 28 different countries that we received for TaLC 8. The present volume is a compilation of the 3 invited talks, 48 papers, 22 posters, 4 software demonstrations and 4 workshops that were finally presented at TaLC 8, Lisbon, between 3 and 6 July 2008.

Just by looking at their titles, it is easy to see how diversified and widespread the domain of teaching and language corpora has become. We have presentations about teaching advanced learners as well as beginners, and this includes university students, school children and even pre-schoolers. If in the beginning corpora was used mainly to teach English, on this occasion we also have presentations that draw on corpora of Portuguese, Mandarin, Cantonese, Spanish, Greek, Russian, Ukranian, Persian, Japanese, Czech, French, German, Italian, Lithuanian and Romanian. And it is not just different languages that are represented here, but also different types of languages: academic discourse, classroom discourse, youth language, learner language, translated language and even the discourse of diplomacy and of subtitles.

Whereas corpora used to be mostly about the written medium, in TaLC 8 there is no shortage of presentations about corpora and speech. In addition to grammar and lexis, many of the papers in these proceedings are about phraseology, translation, literature and culture.

Another point to be made is that this conference is not just about putting existing corpora to use in the classroom, but also about compiling different types of corpora for teaching, exploring novel types of pedagogical tagging, developing accessible corpus tools for education, using corpora for language assessment and training novice users how to use corpora. Of course, not everyone needs to learn how to use corpora. We have here both presentations that focus on the direct use of corpora by learners and presentations about developing data-driven materials which people who have never heard of corpora can benefit from.

The content of this volume is also a measure of the direction in which we are heading. It is perfectly clear that language corpora are not being used as an end in itself, but as a means of achieving different educational goals in different educational settings.

Our common interest in corpora and teaching has brought together researchers, practitioners and theorists in an unprecedented way. It is not just a question of establishing links with delegates from different countries. TaLC is an opportunity for corpus compilers to meet corpus users, for people developing corpus tools to exchange ideas with people developing corpus-based materials and people using corpora directly in the classroom. It is a chance for researchers working with written corpora to meet those working with spoken corpora, for teachers using corpora of one particular language to talk to teachers using corpora of other languages, for people interested in teaching literature to meet people interested in error analysis. These proceedings are a reflection of the fruitful exchange of ideas that will have taken place during TaLC 2008.

CONTENTS

KEYNOTE SPEAKERS

PAPERS

POSTERS

SOFTWARE DEMOS

WORKSHOPS

KEYNOTE SPEAKERS

FREQUENCY IS IMPORTANT - AND CHALLENGING:
A PRESENT-DAY CORPUS PERSPECTIVE

Geoffrey Leech
Lancaster University, UK

I begin my lecture with a brief survey of how frequency - in particular, word frequency - had a role in language learning in the days before electronic corpora existed. Then I consider how the 'corpus revolution' made available frequency information about language use in a totally unprecedented way from the 1960s onward. We now live in an age where frequency dictionaries and frequency-based grammatical information are becoming more and more available - for example, a new corpus-based frequency dictionary of Portuguese has just been published. New sources of frequency information from the Web are being tapped. Various kinds of knowledge found in present-day learners' dictionaries (grammatical, collocational, semantic) are getting to be frequency-based.

But how far is this useful for the language teacher and learner? Is the right kind of frequency knowledge being captured? In the second half of the presentation, I will consider the equation "more frequent = more important", what questions of frequency we really need to ask, and how far they can be answered in the present state of corpus linguistics.

WORKING WITH DIFFERENT CORPORA IN TRANSLATION TEACHING

Natalie Kübler
Université Paris 7, France

Corpus use in translation teaching has established itself for some time now. Several types of corpora have been taken into account in this field, such as parallel (also called translation) corpora, comparable corpora, monolingual corpora, disposable corpora, specialised vs "general" corpora etc.

Depending on the translation type -- literary or pragmatic translation -- corpus use can vary very much and offers several approaches to help learners with the act of translating. This paper however will focus on pragmatic translation, i.e. translation that is based on communicative, rather than literary criteria.

We will present the different possible approaches that can be applied for translation and translation teaching, and the different types of corpora used in this respect. Learner translation corpora will be presented to illustrate how such a corpus can be used in teaching translation. This type of corpus (which could be defined as a sub-type of learner corpora), is still quite rare. Suggestions for combining different approaches to obtain better results will be shown.

TALC IN ACTION: RECENT INNOVATIONS IN CORPUS-BASED ENGLISH LANGUAGE TEACHING IN JAPAN

Yukio Tono
Tokyo University of Foreign Studies, Japan

In this talk, I will present recent innovations in English language teaching in Japan with a special emphasis on the creation of the world's first corpus-based TV English conversation program. The program ran from 2003 to 2006, a hundred units featuring 100 keywords selected based on BNC. Each unit focuses on useful collocation patterns of the keywords, with model skits videotaped in UK, USA and Australia, and ample exercises. A special CG character called "Mr Corpus" introduced the corpus ranking.

The impact of the program was significant. More than a million people watched the program and the word "corpus" became a familiar term. Various corpus-based teaching materials have been published since then. There is also a growing demand among English teachers to know more about corpora and corpus-based ELT. I will report on some of these recent developments in Japan and share some useful tips, guidelines, and a lesson I learned from these experiences.

PAPERS

RAISING LANGUAGE AWARENESS THROUGH INVESTIGATION OF A LEARNER CORPUS OF ONLINE COMMUNICATION

Katherine Ackerley[1]

Fiona Dalziel[2]

Francesca Helm[3]

Abstract

This paper will describe the use of a learner corpus of online written communication. Over the past few years, English language courses at the University of Padova, Italy, have made extensive use of software for computer-mediated communication (CMC) for a wide range of language learning activities. The texts produced by these learners represent a wide range of genres and include: personal profiles, learner diaries, online debates, formal reports and compositions, peer interaction, teacher/student interaction. While this emphasis on learner production and interaction is believed to foster second language acquisition, it is also important for learners to focus on form. By building 'quick and dirty' local learner corpora the authors have explored learners' patterns of language use and identified some problem areas. With the compilation of a reference corpus it has also been possible to investigate the overuse, underuse and avoidance of forms such as modal verbs and expressions of agreement. The authors have developed a variety of learning activities, from worksheets for classroom use to data-driven activities where the learners access both reference corpora and those made up of their own work. The piloting of these materials appear to confirm that this can be a stimulating way for learners to become aware of patterns of their own language use and to "notice the gap" between their output and the target input. The success of these activities with the learners together with the realization that we have at our disposal a wealth of electronic learner production has led us to start a more ambitious project of compiling a more systematic learner corpus of online communication, the Padova Learner Corpus. This could constitute a highly original diachronic learner corpus containing texts learners produce during the 3 years of their university careers and consisting of different genres.

Keywords: computer mediated communication, online genre, learner corpus, reference corpus, data-driven learning

The Learning Context

The learner corpora presented contain the written production of undergraduates studying on the two language degree courses, *Lingue, letterature e culture moderne* and *Discipline della mediazione linguistica e culturale* at the University of Padova. In the 2007-2008 academic year, over 500 students enrolled on these courses. Students come from a variety of language learning backgrounds and their level of language competence at entry ranges from A2 to B2 level according to the *Common European Framework of Reference for Languages* (Council of Europe, 2001). At the beginning of the first year, students sit a placement test to divide them into reasonably homogenous groups for their English classes. The English courses adopt a blended approach to language learning, with students attending lessons in the classroom and in the multimedia language laboratory, and carrying out individual and group assignments online.

CMC and language learning

Computer-mediated communication (CMC) has been used for language learning for many years at Padova University. Although CMC can take on various forms, in this paper we use CMC to refer to asynchronous written

[1] Katherine Ackerley is an English language teacher and researcher at the Faculty of Arts and Humanities, University of Padova. She is the co-ordinator of online language learning for the University Language Centre. Her research interests include the application of corpus linguistics to language learning, online language learning and computer-mediated communication.

[2] Fiona Dalziel is an English Language teacher and researcher at the Faculty of Arts and Humanities University of Padova. She is the co-ordinator of the CercleS European Language Portfolio (ELP) project for the University Language Centre. Her research interests include corpus-based approaches to language learning, learner corpora, the ELP and online language learning.

[3] Francesca Helm is an English language teacher and researcher at the Department of International Studies, University of Padova. Her research interests are in the areas of computer-mediated communication and the use of technology in language learning; intercultural communication and telecollaboration; learner corpora and corpus-based approaches to language learning.

production and interaction mediated by networked computers, allowing for “many to many” communication. CMC is seen not only as a tool to enhance language learning but as a part of everyday communication and information practices in educational, professional, recreational and interpersonal realms (Thorne 2008: 417). The ability to communicate in a foreign language through CMC is deemed to be a fundamental requirement for our learners to become active, ‘literate’ members of the Information Society. As Warschauer (2000) writes, “if our goal is to help students enter into new authentic discourse communities, and if those discourse communities are increasingly located online, then it seems appropriate to incorporate online activities for their social utility as well as for their perceived particular pedagogical value.”

In their computer-conferencing environment, FirstClass, learners engage in a variety of written tasks which involve production, such as writing personal profiles, diaries, reports and interaction with other learners, participating in online debates, peer review and surveys. CMC has been embraced enthusiastically by both teachers and learners because learners become more active and confident users of the target language, they engage in collaborative learning and are highly motivated, corroborating research findings (Chun 1994, Kern 1995, Warschauer 2000) about the value of CMC for language learning. The use of CMC in language teaching has been based mainly on sociocultural and interactionist theories of Second Language Acquisition (Thorne 2008) whereby it is through the construction and negotiation of meaning that learners develop their foreign language competence. SLA researchers have sought to demonstrate how learners’ language production and interactional skills improve as they are pushed to produce more output and negotiate meaning in interaction (Pellettieri 2000).

CMC and focus on form

Although the focus in sociocultural and interactionist approaches to online language learning tends to be on communication and the negotiation of meaning, CMC practitioners have recognised that “learners need reflective activities to develop language awareness, as well as productive activities, in order to become effective and autonomous learners” (Levy and Kennedy 2004: 53). In our particular context, the emphasis on communication and meaning is at times felt to take time away from a focus on form; in end-of-course questionnaires learners have revealed a desire for more personalised feedback and correction of mistakes, but with one language instructor per 100 students, individualised feedback is a huge burden on the teacher and not always possible.

The SLA literature regarding ‘focus on form’ looks at how to balance attention to fluency and form specifically in communicative contexts. As Doughty and Williams (1998: 3) write, “focus on form entails a prerequisite engagement in meaning before attention to linguistic features”. One of the advantages of CMC is that a permanent record remains of all production and interaction and these records are valuable for post-activity discussion and reflection. It is interesting, however, that to date few practitioners have considered using a corpus approach to focus on form in post-task reflective activities, with the exception of Belz (2004).

A ‘quick and dirty’ local learner corpus of online debates

In an attempt to meet our learners’ needs we have started to compile what Seidlhofer (2002) describes as “ ‘quick and dirty’ local learner corpora” to identify problem areas across large groups of learners. Small corpora of learner productions on particular tasks, for example online debates, have been compiled to explore problem areas for our students and identify mistakes across groups of learners. For instance in the academic years 2006/7 and 2007/8 online debates were carried out on topics of social interest, such as the death penalty, euthanasia, immigration and integration and the legalization of soft drugs. The aims of the debate were to improve general language skills through a motivating task and to develop sensitivity to register differences. The students contributed to the debates using an informal register as they were interacting with their peers, debating social issues in a forum intended to be similar to public discussion forums found on the Internet, for instance those related to news sites, such as the BBC. For each debate the students were given a prompt and a series of Internet links to related articles. The students knew that they would not be assessed on the quality of their work, but that the aim of the activity was to communicate in English and to interact with their peers by expressing their opinions on various controversial issues.

Compiling a reference corpus

The identification or compilation of appropriate reference corpora in learner corpus studies is a complex affair. How appropriate is it to use a large corpus like the BNC as a reference corpus? Small learner corpora tend to consist of samples of just one genre (usually the academic essay) while large reference corpora have a whole array of text types. While some features of learner production may not vary across different modalities and genres, other features may be genre- and mode-specific (Barlow 2005: 336). Our debate corpus represents an online genre, but most of the large reference corpora contain few or no examples of CMC; the texts included conform mainly to 'predigital epistolary conventions' (Thorne 2008: 417). It was thus decided that it would be more appropriate to compile a reference corpus that represented, as far as possible, the same genre of text, written in the same mode by proficient target language users. We compiled a reference corpus which consists of contributions to public discussion forums on the same topics that the learners had discussed with their peers. The forums we selected came from the ‘BBC Have your Say’ website and were similar to the type of online debate our learners engaged in, that is discussions about contemporary social issues (migration and euthanasia) stemming from a prompt or news story.

Creating activities

As mentioned above, the design of teaching activities involved a “quick and dirty” approach to corpus compilation and analysis. The aim was to identify problem areas that could be dealt with during the course, while the task on which the learner corpus was based was still fresh in the minds of learners. The authors chose first to look at the debate on immigration conducted in the 2006/7 academic year. The learner corpus consisted of 111 texts written by 65 learners (the learners had contributed between one and ten messages to the debate). The total number of words was 17,905. After selecting a suitable reference corpus, as described above, it was decided to start by looking at learners’ use of modal verbs: these are known to be a challenging area of English grammar for learners, and their use has in fact already been investigated in a number of learner corpus studies (see for example Aijmer 2002; Hyland and Milton 1997). Moreover, it was also thought that modals, an important way of expressing stance, could be particularly relevant to the genre of online debates, where students are expected to take an ideological standpoint on various issues.

An initial comparison between the learner and reference corpus of debate messages on the topic of immigration and the reference corpus revealed, for example, that must appeared more often in the learner corpus (35 times as opposed to 23 in the reference corpus) whereas should was used less frequently by learners (63 times as opposed to 85 in the reference corpus). Moreover, all occurrences of must regarded its deontic/root rather than epistemic meaning, and of these, 12 referred to the obligations of immigrants in the host country. Such use was rare in the reference corpus, occurring only three times, with most instances of must relating to the obligations of the host community. One set of activities, designed for classroom use, was aimed at helping learners to reflect on the strong obligation expressed by must, which in some circumstances could be considered inappropriate. The learners were first given a printed sheet containing a number of concordance lines from the reference corpus from which the verb should had been deleted and were asked to decide which word was missing.

```
29 enviously the achievements of others. Migration        be strictly controlled.
30                                         Immigrants        be helped and encouraged. Otherwise
31two or more years. This is outrageous! Migration        be encouraged as we bring skilled
32 ght those countries to that position. So migration        not be stopped. To live away from your
33 nk all those people who have suggested migrants        adopt the cultures and traditions of their
34 eekers or otherwise) and also believed the Maori        be denied any financial help
35 y you will become a citizen of our country, but it        be fully up to Americans to decide who
36 migration should be afforded equally to all, yet it        not grow to a point where it creates a
37 lly be absorbed into the culture. Yes immigration        be reviewed often and as necessary.
38                                  Mass immigration        be stopped altogether in the UK and
39 trategies. Secondly, more political stability issues        be addressed with more emphasis on l
40 ts legal form, should be encouraged. Immigration        be afforded equally to all, yet it should
41 To integrate takes time so the rate of Immigration        be proportionate to that.
42  media coverage. Asylum seekers and immigrants        not be used
```

Exercise based on concordance lines from immigration reference corpus

1. Learners were then shown concordances of *must* from the learner corpus and asked to identify the subject of the modal verb, to decide in which cases *must* could be replaced by *should* and what difference this might make to meaning.

abits somehow. On the one hand, every country **must** welcome immigrants, on the other hand,
e iene" made a report on this problem!!! I do **must** say that Senegaleses are very good and
ll together is important cooperate: everybody **must** have a dignified job, everybody have to
hasn't changed, I still think that foreigners **must** be accepted (they obviously must behave
migrant comes to the country where I live, he **must** respect the rules we have...so: no temp
in conclusion, if an immigrant comes here he **must** treat this country as his second home,

far as the wall in Via Anelli is concerned I **must** admit that some immigrants are really v
I think those who enter Italy **must** adapt their needs to Italian's rules. I
foreigners must be accepted (they obviously **must** behave in a polite and correct way!). T
ened to Enrico, and I think that those people **must** be punished... And, Valentina, I don't
ly here. I feel really sorry for peopole that **must** escape, leaving their own houses and la
PEOPLE WHO LIVE IN IT... BUT THERE **MUST** BE COLLABORATION
th you about criminals who come to Italy.They **must** be all sent to jail. But concerning beg
ing the beggers that cheat and criminals they **must** be punished according to the law.No pit
t welcome immigrants, on the other hand, they **must** conform themselves to usages and
nally I would never give them any money. They **must** go and work as everyone else.The only
efully in our nation!! If they come here they **must** have the true intention to work and to
o recognize people by seeing their face, they **must** let us see who they are. It's also a po

Some examples of *must* from the learner corpus

2. Modals were also the subject of hands-on data-driven corpus activities (Johns 1993) presented in the language laboratory, but with focus on the verb *may*. This time, the debate had been carried out in the 2007/8 academic year and the corpus consisted of 15,694 words. First of all, students were asked to fill in a table comparing the frequency of modal verbs in the learner and the reference corpus, using the wordlist function of Wordsmith Tools. The biggest difference concerned *may*, appearing only once in the learner corpus but 19 times in a reference corpus of the same size.

The learners were then asked to look in detail at how *may* was used in the reference corpus, where it appeared 8 times with its function as an indicator of concession. The learners were not aware of the latter use, and this led to reflection on how conceding a point in a discussion or debate can be an effective way of bringing one's audience around to one's own point of view.

Another laboratory task involved hands-on work using learner (17,086 words) and reference corpora of debates on the issue of euthanasia. A preliminary analysis of the corpora had investigated not only the use of modals but of other epistemic devices, using a list compiled by McEnery and Kifle (2002). This revealed overuse on the part of the learners of the noun possibility, probably due to the influence of the L1, in this case Italian for the majority of the learners. The learners themselves were asked to look at the use of the word possibility in the two corpora. They discovered that it occurred 18 times in the learner corpus but only once in the reference corpus. They were then required to look at the occurrences of possibility in the learner corpus and to think of possible alternatives. The students came up with various answers, including the use of modals and semi-modals (can, could, be able to) or the word opportunity.

One classroom task looked at expression of agreement and disagreement, once again with reference to the online debate topic of immigration and integration (2007/8). Working with printed concordance lines, first of all, the learners identified some inappropriate uses of the verb agree in their own work. They then went on to see how in both corpora (learner and reference), emphasis could be given to expressions of agreement (I agree wholeheartedly, I couldn't agree more etc.) and how hedging was often used in order to put forward a difference of opinion without offending a potential reader. Interestingly enough, in the majority of cases in the reference corpus when the verb agree was used, the writer then went on to express disagreement (I do agree that … but …).

Padova University Learner Corpus

The use of these mini learner corpora, created ad hoc for the needs of a particular class, proves both useful and stimulating for students, but the need for the more systematic collection of the vast amount of materials stored on FirstClass has been recognised. Although extensive research has been carried out on existing learner corpora (Granger, Hung and Petch Tyson 2002; Prat Zagrebelsky 2004), it has been pointed out that these are currently restricted to general and academic English (such as argumentative essays) generally produced by advanced learners and that longitudinal corpora are rare (Barlow 2005; Tono 2003; Prat Zagrebelsky forthcoming).

The wealth and variety of texts written by students of varying levels of competence throughout their 3 years of study (or five in the case of those doing a post-graduate course) provides exciting opportunities for both research and materials development. Work has therefore recently begun on the creation of the Padova University Learner Corpus. This involves obtaining students' permission for their work to be used for research purposes and gathering and recording data about individual learners and tasks. The work of consenting students is then retrieved from FirstClass and archived in sub-corpora, which allow the researcher to study the language produced by a given group of students in a specific text type (see table below). These sub-corpora vary in register from personal presentations and the informal messages exchanged between peers in the Student Bar, an area created to allow online socialising, to the language produced in formal reports. The degree of spontaneity or planning with which each text is produced also affects the kind and quality of language. Messages in the Student Bar tend to be totally spontaneous, while the Personal Profiles, though written in informal English, undergo a process of peer revision and are eventually assessed, therefore a certain degree of planning and editing goes into their writing.

Sub-corpora	Text type	Year of study	Level	No. words
Student Bar	Informal messages between students	1st year	A2-B1	30,000
Personal Profile	Informal written presentations	1st -2nd year	A2-B2	52,900
Learner diary	Diachronic collection of weekly messages	1st -2nd year	A2-B2	315,000
Formal reports	Reports summarising online debates	1st year	A2-B2	32,700
Debate corpus	Informal online debate on social issues	1st year	B1-B2	150,000
Total				**580,600**

Table 1: Sub-corpora in the Padova University Learner Corpus

Conclusions

We are still at the early stages of corpus compilation: texts have been gathered but decisions still need to be made regarding issues such as treatment of spelling mistakes and error tagging, how best to organise the data and how to make it accessible. To achieve this it is necessary to ensure with the systematic retrieval of and archiving of texts, taking care to include significant metadata about both text types and students. The future objectives of this research include not only a broader and more systematic approach to corpus design, but also the further analysis of the various sub corpora with a view to continued development of materials that are relevant to students' needs. Another, more ambitious aim would be to analyse the language produced by individual students across different texts and over time, allowing the researcher to observe a learner's progress longitudinally, over his/her three years of university study. A short-term objective is to integrate data-driven learning based on learner corpora more extensively in order to see which types of tasks are successful and which need further development. It is hoped that these activities will increase students' awareness of the usefulness of corpora in their own language learning with the final aim of enabling them to become more autonomous and successful learners.

References

Council of Europe. 2001. *Common European framework of reference for languages.* Cambridge: Cambridge University Press.

Aijmer, K. 2002. "Modality in advanced Swedish learners' written interlanguage." In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching,* S. Granger, J. Hung and S. Petch-Tyson (eds.). Amsterdam: Benjamins, 55-76.

Barlow, M. 2005. "Computer-based analyses of learner language." In *Analysing Learner Language,* G. Barkhuizen and R. Ellis (eds.). Oxford: Oxford University Press, 335-357.

Belz, J. 2004. "Learner corpus analysis and the development of foreign language proficiency." *System* 32/4: 577-591.

Chun, D. 1994. "Using computer networking to facilitate the acquisition of interactive competence." *System* 22/1: 17-31.

Doughty, C. and **Williams, J.** 1998. *Focus on form in classroom second language acquisition.* Cambridge: Cambridge University Press.

Granger, S. Hung, J. and **Petch-Tyson, S.** 2002. *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam Philadelphia: J. Benjamins.

Hyland, K. and **Milton, J.** 1997. "Qualifications and certainty in L1 and L2 students' writing." *Journal of Second Language Writing* 6/2: 183-205.

Johns, T. 1993. "Data-driven learning: an Update." *TELL&CALL* 2: 4-10.

Kern, R. 1995. "Restructuring classroom interaction with networked computers: Effects on quantity and characteristics of language production." *The Modern Language Journal* 79/4: 457-476.

Levy, M. and **Kennedy, C.** 2004. "A task-cycling pedagogy using stimulated reflection and audio-conferencing in foreign language learning." *Language Learning & Technology* 8/2: 50-69.

McEnery, T. and **Kifle, N. A.** 2002. "Epistemic modality in argumentative essays of second- language writers." In *Academic Discourse*, J. Flowerdew (ed.). London: Longman, 182-195.

Pellettieri, J. 2000. "Negotiation in cyberspace." In *Network-based language teaching: Concepts and practice*, M. Warschauer and R. Kern (eds.). Cambridge: Cambridge University Press, 59-86.

Prat Zagrebelsky, M. T. 2004. *Computer learner corpora.* Alessandria: Edizioni dell'Orso.

Prat Zagrebelsky, M. T. Forthcoming. "Learner Corpora at the Crossroads of Computer Corpus Linguistics, Foreign Language Pedagogy and Second Language Acquisition Research." In *Corpora for University Language Teachers*, C. Taylor Torsello, K. Ackerley, and E. Castello (eds.). Bern: Peter Lang.

Seidlhofer, B. 2002. "Pedagogy and local learner corpora: Working with learning-driven data." In *Computer learner corpora, second language acquisition and foreign language teaching*, S. Granger, J. Hung and S. Petch-Tyson (eds.). Amsterdam: John Benjamins, 213-234.

Thorne, S. L. 2008. "Mediating Technologies and Second Language Learning." In *Handbook of Research on New Literacies*, D. Leu, J. Coiro, C. Lankshear and M. Knobel (eds.). Mahwah, NJ: Lawrence Erlbaum, 417-449.

Tono, Y. 2003."Learner corpora: design, development and applications." In *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003). Technical Papers 16*, D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). Lancaster University: University Centre for Computer Corpus Research on Language, 800-809.

Warschauer , M. 2000. "Online learning in second language classrooms: An ethnographic study." In *Network-based language teaching: Concepts and practice*, M. Warschauer and R. Kern (eds.). Cambridge: Cambridge University Press.

WHAT DO ANNOTATORS ANNOTATE? AN ANALYSIS OF LANGUAGE TEACHERS' CORPUS PEDAGOGICAL ANNOTATION

José María Alcaraz[4]
Pascual Pérez-Paredes[5]

Abstract

One of the most neglected areas of research in corpus-based language teaching is the pedagogical annotation of corpora. The studies that deal with this emerging area are scarce, dealing mainly with theoretical aspects (Braun 2005, 2006, 2007) and practical implementation issues (Pérez-Paredes and Alcaraz 2007).

The two case studies we report involved professionals of different background and training experience. Our research discusses quantitative as well as qualitative data that emerges from the annotations analyzed. The former include an analysis of different levels of annotation relevant in pedagogical corpus annotation (corpus, text and section levels) and different measures concerning the categories and keywords annotated. The qualitative information we have considered incorporates background data of the annotators as well as insights into the mediation role underlying the annotating task of the teachers in the study. Despite the explorative scope of our study, the results show that, while retaining significant differences in terms of the annotation items targeted, the annotations share a common understanding of the role and scope of pedagogic annotation in language teaching. These results reveal that pedagogical annotation may become an important tool in a situation where teachers organize language learning experiences around the use of ad-hoc, teacher-led corpus-based materials that focus on the specific needs of learners, as opposed to situations which make use of either raw texts, where no annotation is available, or POS tagged corpora. The authors suggest that further research will be necessary to establish more solid links between pedagogical annotation and needs-driven selection of corpus-based materials in the language classroom. The authors propose the use of two annotation density measures to gain further insight into idiosyncratic annotation behaviour. Despite the limitations of the case study methodology, this research shows that pedagogical annotation can be measured and analysed.

Keywords: corpus annotation, youth language, corpus-based teaching, material development, software-aided analysis of language

Pedagogical annotation in the field of corpus-based studies: from possibilities to feasible corpus-based materials

Different researchers have turned their attention to the role of corpora in the foreign language classroom (Fligelstone 1993, Barlow 1996, Aston 1997, Hunston 2002). Römer (forthcoming) has examined these contributions in the light of the distinction between direct and indirect applications of corpora to language education. The latter benefit from insights into the descriptive nature of language to inform language teaching. Examples of this application are the selection of lexis based on frequency or on the patterning features of language, including phraseology and collocation studies. Revisiting commonly taught grammar such as modality or verb tenses has also been in the researchers' agenda. By describing the language based on language corpora, new evidence is gathered on the ways language works and, most significantly, on the way language is used by speakers.

Indirect applications, on the contrary, bring existing language corpora to the classroom. As Römer (forthcoming) puts it, while indirect approaches "centres on the impact of corpus evidence on syllabus design [...] and is

[4] José M. Alcaraz-Calero is a Computer Science Engineer at the Computer Science School, University of Murcia. He has widely contributed to Corpus-based language research and thus has implemented solutions for different projects at UMU, including automatic semantic tagging of Spanish based of word sense disambiguation (WSD) and intelligent tagging algorithms for the CUMBRE corpus of Spanish. At the moment he is completing his PhD in autonomic computing, networks and ontologies.

[5] Pascual Pérez-Paredes started his collaboration with the English Department in the University of Murcia, Spain in 1996. After a research stay in the University of Texas at Austin, he completed his PhD in Applied Linguistics in 1999. He currently teaches CALL and Applied Linguistics. He is also a Sworn/Official Translator. His main interests are quantitative research of register variation, the compilation and use of language corpora and the implementation of Information and Communication Technologies in Foreign Language Teaching/Learning. He is a member of the Research Group Lingüística Aplicada Computacional, Enseñanza de Lenguas y Lexicografía (LACELL). Pascual Pérez-Paredes directs a research project funded by the SENECA agency on orality. He is also the coordinator of a MINERVA project funded by the European Commission: SACODEYL. At the moment, he is involved in corpus-based international projects such as LINDSEI (UCL) [http://cecl.fltr.ucl.ac.be/] and ICCI (TUFS) [http://www.tufs.ac.jp].

concerned with corpus access by researchers", direct approaches are "more teacher and learner-oriented". In this context, the learner becomes a researcher, a detective, and the corpus is the field where students are exposed to the reality of language use. The main learning activity of direct applications is the examination of concordance lines and, consequently, the examination of nodes in a given context. This activity requires that learners induce meanings and formulate hypothesis about how language works. In general terms, corpora are seen here as complements to traditional teaching resources.

In the two approaches above we can appreciate that corpus-based linguistic research is predominant and, therefore, applications to other fields are possible but, positively, did not motivate the design of the original corpus. Both direct and indirect approaches belong to what we may call the possibilities scenario. This is characterized by the effort of language educators and researchers to apply existing work in the language research-oriented paradigm to the wealth of resources that can be used in FLT. Figure 1 represents the so-called possibilities scenario:

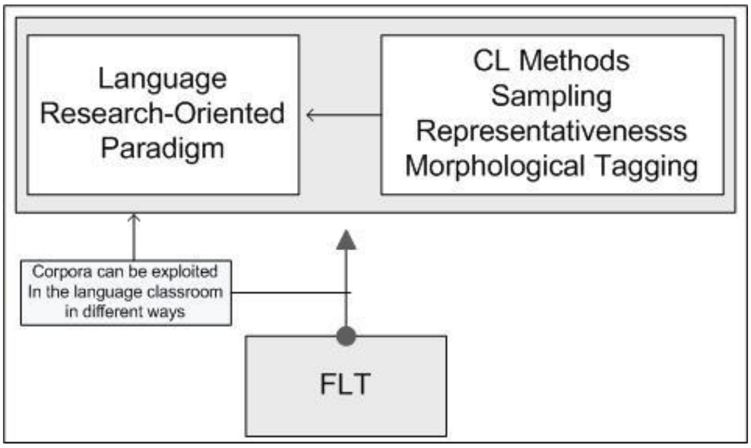


Figure 1: The possibilities scenario

In this scenario, the corpus is principled and, in most of the cases, claims to be representative of a language, language community, register or specialised use. It is possible to use these research-motivated corpora in the language classroom but, reasonably, there has to be an effort to adapt their primary research orientation in the language classroom. Pérez-Paredes (2007) has discussed some of the limitations to the adaptation of general, principled corpora to the language classroom. One of the most relevant is the high cognitive demand which is put on the learner, who is asked to search for linguistic patterns in a corpus which has been most certainly designed to serve research purposes. In this context, the learner will have to interpret accumulative concordance lines, which are usually extracted from texts of a very different nature and, for most learners, totally unrelated to their learning experiences. This poses a high demand on learners, who are invited to refine their search precisely because of the complexity that is presented before their eyes in terms of genres and language. Besides, learners will have to discriminate what is relevant and what is not in terms of language use and weigh down the influence of the context and cotext in the results. Unfortunately, evaluations of corpus-based learning in non-tertiary institutions are scarce (Braun 2007).

Researchers and learners certainly have different goals when approaching a language corpus. However, in the possibilities scenario learners have been someway driven to apply research methods in their learning routines. While this seems to be a positive asset for university language students (Bernardini 2004), it definitely appears as a hassle (Mauranen 2004) for mainstream young learners in secondary education or non-tertiary contexts.

Customising corpora for the language classroom: the potential of teacher-driven annotation

Customised corpora in the language classroom

It is unquestionably possible to use language corpora in the language classroom. However, the primary goals of research-oriented corpora make it essential that teachers adapt these resources to the needs and everyday demands of language learners. Bernardini (2004:32) sees a very significant potential in discovery learning, but she is cautious about the technological limitations and, even more important, about the training and background of students: “ [...] learners require guidance and heightened awareness to learn from corpora and much of their potential (for strategic learning, serendipity, reasoning-gap, as well as for stimulating communicative activities) would be lost if learners did not have a chance to carry put relatively complex analyses, requiring them to observe phraseological regularities and restrictions and the functions associated with them”.

These complex analyses are difficult to implement in mainstream language teaching and learning, where there is a high pressure on communicative goals and not so much on mastering the complexities of the lexico-grammatical interface. However, there is a chance for learners to carry out the type of discovery learning discussed above if the language corpora that are brought to the language classroom are not in conflict with the syllabus and general communicative orientation of mainstream, non-specialised tertiary, language students’ goals. Here there must be some room for topic-driven corpora (Braun 2007) that are geared towards the integration of corpus-based materials and general, mainstream language learning, especially secondary education language learning. In the framework of SACODEYL6, we believe that the professional, teacher-led, pedagogical annotation of resources may play a significant role in bringing language corpora to the language classroom. If we want learners to become discoverers and researchers, it may make sense to think of teachers as guides and pathfinders. This is where annotation is crucial. Instead of a possibilities scenario, pedagogical teacher-driven annotation of corpus resources may facilitate a paradigm which is focused on language learning and mainstream language learners’ needs. This scenario is capable of being accomplished in the context of instructed mainstream language learning:

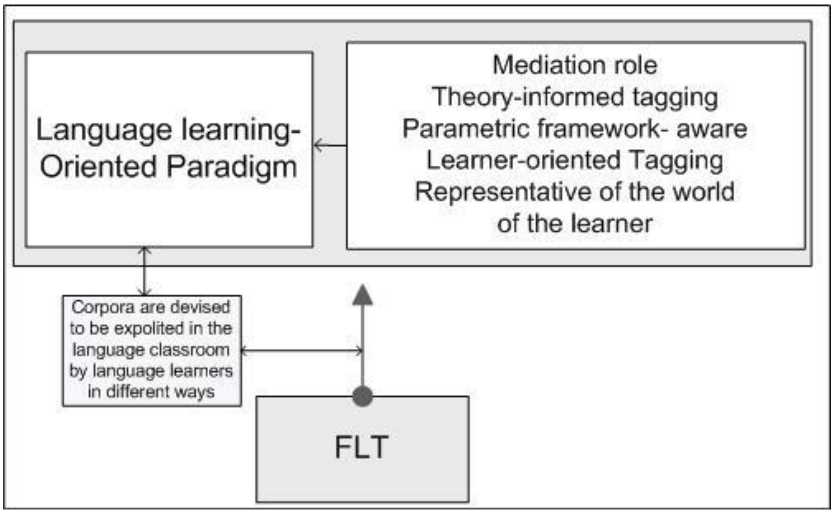

Language learning-Oriented Paradigm
Mediation role
Theory-informed tagging
Parametric framework- aware
Learner-oriented Tagging
Representative of the world of the learner
Corpora are devised to be expolited in the language classroom by language learners in different ways
FLT

Figure 2: The feasibility scenario.

This scenario is based on Widdowson’s (2003) theory of the role of applied linguistics in language education and, in particular, on five milestone ideas. First we find the mediation role that applied linguistics plays in informing language learning and teaching on those disciplinary points of reference that are found to be of relevance in the field of linguistics. Second, the notion that pedagogical corpora can be annotated, that is, enriched in many different ways and following different theoretical stands, not necessarily just one. In this sense, in potential, pedagogical annotated corpora are not proactively biased towards any particular theory, something which is not possible in automatic tagging. Teachers annotating a corpus from a pedagogical perspective “define problems as explicitly as possible so that they are amenable to solution reactively in the teaching and learning process (Widdowson 2003:19). Third, any pedagogically-motivated corpus should acknowledge the parametric framework where the learning and teaching experiences take place. The numbers of influences, variables, personal and institutional, which play a role in language teaching make every single language learning experience different. This uniqueness is neglected in the possibilities scenario. Fourthly, as a consequence of the ideas expressed above, annotation should be performed with the learner in mind and, if possible, bearing as many of those parametric factors above in mind as possible. Finally, it is the authenticity issue. It is unquestionable that the English we find in the BNC is authentic. However, how authentic are the genres in the BNC for a 15-year old? Widdowson (2003:

[6] System Aided Compilation and Open Distribution of European Youth Language: http://www.um.es/sacodeyl

126) believes that the world to replicate in the learning experience is that of the language user and this is where the worlds of the L1 community users and those learning it are different.

We believe that when teachers become annotators it is more feasible to put language corpora resources to good, realistic and authentic use in the language classroom.

What do annotators annotate? An analysis of language teachers' corpus pedagogical annotation

If pedagogic annotation is to play an active role in bringing corpora to the mainstream, non-tertiary education language classroom, we need to develop a deeper understanding of how teachers annotate a corpus. This is only possible if real FLT teachers are confronted with the annotation process itself and are given a hands-on, practical framework that makes this possible. In order to provide teachers with such a framework, we have made use of the tools and products developed under the SACODEYL initiative.

In particular, our research will try to gain insight into both quantitative and qualitative data that will inform our process of analysis on the ways in which FLT teachers annotate a corpus. Some of the issues we want to discuss include the number of categories, subcategories, sub-subcategories, etc. annotated on the different sections of the corpus; the number of new, teacher-implemented categories, subcategories, sub-subcategories, etc. annotated on the different sections of the corpus, as well as different quantitative data concerning the selection of exponents associated to annotated categories. From the above, and especially, from the comparison between the annotations of our informants, we set out to present an analysis of the variables and questions for further research that should be taken into consideration to implement more ambitious schemes of teacher-driven pedagogical annotation and research.

Methodology

Method

A case study methodology has been used in order to gain insight into the nature of FLT teachers' pedagogical annotation. The case study (Bronwyn et Al. 2005) is a "form of qualitative descriptive research which examines a small participant pool and which draws conclusions only about that participant or group and only in that specific context". Researchers using case studies do not focus on cause-effect relationships, instead emphasis is placed on exploration and description.

Subjects

Two female teachers of English as a Foreign Language graduated in English Studies were asked to annotate a small part of the English corpus of SACODEYL. This is a profile of both individuals:

	Subject A	Subject B
Number of years as English Language Teacher	10 hours (as a trainee)	4 years
Courses taught	Secondary	Primary and Secondary
Courses taught at the moment of research	Secondary	Secondary
Degree finished in	2007	2004
Materials currently used and publishers	Spotlight 2, OUP Targets 2, OUP	New Thumbs Up, Longman Valid, Burlington

Table 1: Training and professional background of informants.

The subjects differ significantly in their prior teaching experience. While Subject B has taught English for 4 years in primary and secondary levels, Subject A is a trainee with very little face-to-face experience with learners. They were also asked to rate (1-5, not familiar at all to expert) their familiarity with important concepts in corpus linguistics. These are the results:

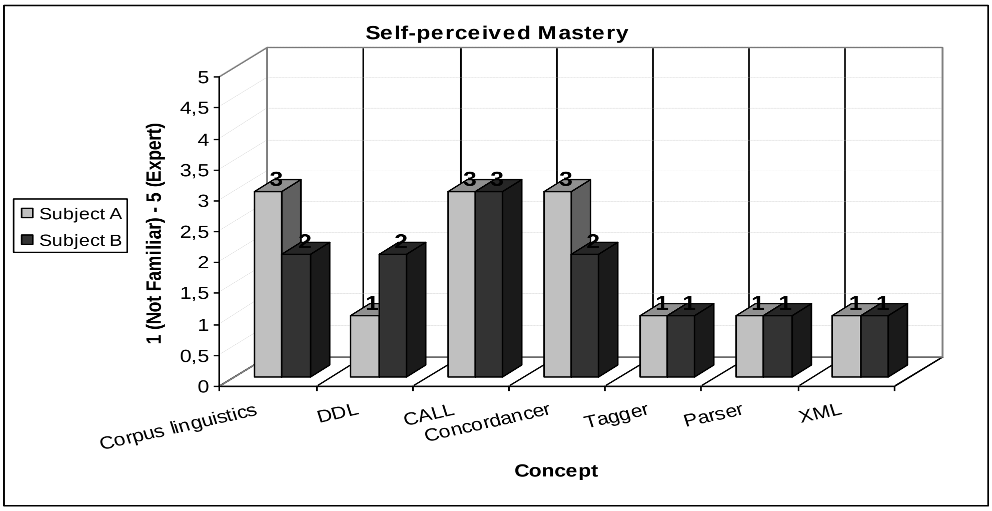


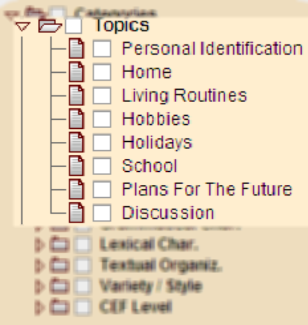
Chart 1: Self-perceived mastery in different corpus-related fields.

Figure 3 shows that both teachers share a very similar training background in the field of CL, with minor differences (+/- 1) in the CL, DDL and concordancer items.

Research condition and data-gathering

Both individuals were given a 90-minute training session which was divided in three parts. First, the individuals were introduced to the rationale behind corpus linguistics, the role of annotation in corpus linguistics and the relevance of annotation in the context of pedagogically-relevant corpora[7]. After this, they were shown a video tutorial of SACODEYL Annotator[2]. This tutorial ran for about 12 minutes. Finally, the individuals were given the task on which we have based our research. The two teachers were invited to annotate those aspects which they considered of pedagogical relevance for the learning/teaching of English as foreign language in Spain. They were prompted to watch a fragment of the corpus first and were given a full transcript of it[2]. Although they were already familiar with the notion of section (Braun 2005, 2006, Pérez-Paredes et Al. 2007), they were told again that the fragment they were going to annotate had been divided into segments that had been considered by the corpus compilers as being of pedagogic relevance. Apart from this, they were given total freedom as to what to annotate and the categories of annotation to apply or even create. They were presented with a basic framework for the annotation of the English SACODEYL corpus that comprised 6 major categories: topics, grammatical characteristics, lexical characteristics, textual organization, variety and style and CEF level for the section.

Each of these categories comprised further subcategories. As an example, we may take topics, which comprised 6 items:



Categories
- □ Topics
  - □ Personal Identification
  - □ Home
  - □ Living Routines
  - □ Hobbies
  - □ Holidays
  - □ School
  - □ Plans For The Future
  - □ Discussion
- □ Lexical Char.
- □ Textual Organiz.
- □ Variety / Style
- □ CEF Level

Figure 3: Topics sub-categories supplied with the English corpus.

[7] http://www.um.es/docencia/perez-paredes/2008/05/what-do-annotators-annotate-analysis-of.html

Both teachers read the same instructions and, particularly, both were made aware of the similarities between annotating with a view on the language classroom and the creation of FLT materials.

After a 10-minute break, both individuals were given 90 minutes to annotate the same fragment of the English corpus of SACODEYL on different laptops. The resulting annotated XML corpus was retrieved from each computer and the annotation processed for further analysis.

The fragment provided for annotation had already been segmented into 10 sections. Table 2 offers the number of words in each section:

Section #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Words	105	181	104	194	268	101	144	377	106	126

Table 2: Number of words in each of the sections.

Results

In this section a description of the results provided after the annotation process carried out by the two subjects is offered and analysed. This analysis has been focused on a pedagogical context so as to enrich our understanding of the ways in which language educators approach the annotation process of a pedagogically-relevant corpus.

After extracting the annotations of both informants, we plotted the following multi-dimensional information:

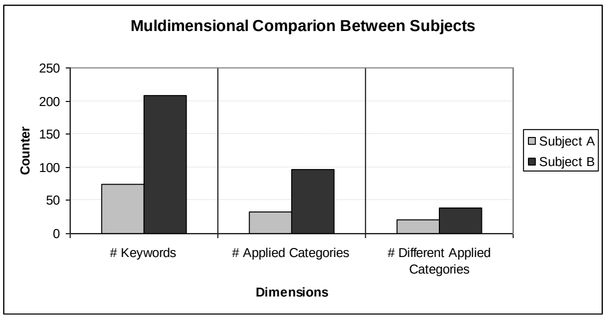
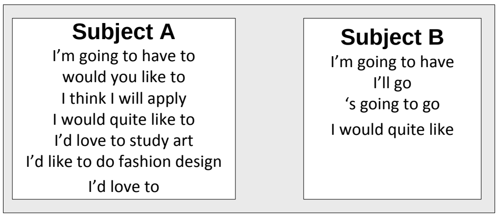


Chart 2: Multidimensional analysis of the texts annotated by the informants.

As seen in Figure 5, Subject B presents higher values than Subject A in all three dimensions. The first dimension, *number of keywords*, corresponds with the total amount of keywords annotated in the whole text. In our annotation process, the user applies a category to a section, and then, optionally, she may decide to link a keyword to this same category and section. A keyword can be any stretch of language, from one word to n-words that may account for the application of a given category, or from a more methodological perspective, keywords could be considered as language exponents (Perez-Paredes et Al. 2007, Alcaraz et Al. 2007). For example, both informants categorized section #8 of the text as plans for the future, which, according to the instructions provided implies that this particular section was considered by both annotators as being useful for the learning and teaching of this language notion. Interestingly enough, both informants linked this category to different keywords:



Subject A
I'm going to have to
would you like to
I think I will apply
I would quite like to
I'd love to study art
I'd like to do fashion design
I'd love to

Subject B
I'm going to have
I'll go
's going to go
I would quite like

Figure 4: Keywords linked to plans for the future in section #8.

While Subject A found more exponents in the section and was very careful and consistent in the boundaries of what a keyword meant for her, Subject B decided to be more selective and occasionally was inconsistent in including a syntactic subject in the formulation of her keywords. However, this proliferation of keywords in Subject A is more the exception than the rule according to our data. In total, Subject B annotated three times as many keywords as Subject A. The reader should not be influenced only by this dimension as it does not show a direct proportion with respect to the quality or usefulness of a text in pedagogical contexts. This metric is greatly influenced by the kind of categories applied to a text. i.e. in the "present simple" category the average of keywords assigned is over 5. However, in the "relative" category the average number is exactly 1. For the "plans for the future" category discussed above the average is 6.

The second dimension in Figure 5. is the *number of applied categories*. This dimension refers to the number of the categories applied to the whole text. For reasons stated earlier, this is closely linked with the number of keywords dimension. It is worth mentioning that the quantitative difference between subjects has been reduced. The third dimension, *different applied categories*, shows the different categories applied to the whole text by the annotators. The quantitative difference between the subjects has been further reduced. We should notice that one thing is the number of categories and that quite another is the categories types. Subject B has repeated the same categories throughout the text more often than Subject A. Furthermore, the former has also associated more keywords than the latter.

A more grained analysis of the second dimension would reveal the annotation behaviour of our informant across the different sections in which the text was originally split. Figure 7a shows how this behaviour evolves by revealing the evolution of the number of applied categories across all ten sections:

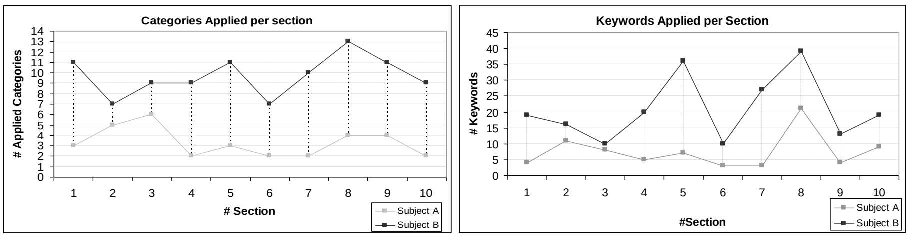
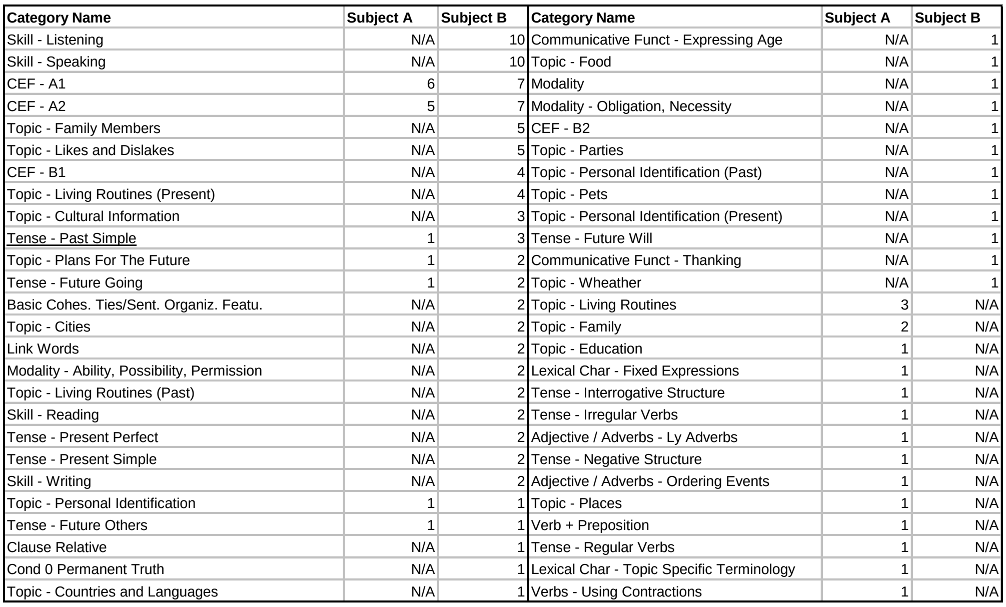


Figure 5: Applied categories (7a) and keywords (7b) carried out by the two subjects throughout all 10 sections.

Both annotators follow the same behaviour pattern with respect to the number of applied categories throughout all the sections, with the exception of sections #2 and #3, where a reduction in the difference between the subjects can be observed. The amount of applied categories to a section is closely related to the weight of this section in terms of number of words, i.e. section #8 and section #5 are most extensively annotated by both subjects and also those with a heavier weight (Table 2). Figure 7b shows how this behaviour evolves by revealing the evolution of the number of keywords applied to all ten sections. We can observe a regular pattern or annotation behaviour except for sections #3, #5 and #7. It is worth noting that sections #5 and #7 are extremely divergent, whereas section #3 is convergent. In sections #5 and #7 the number of keywords assigned by Annotator B doubles that of Annotator A, which surprisingly is unrelated with the weight of the sections. A case in point is section #8, by far the heaviest section in the document, which is annotated by both subjects in a very predictable patterned way. This finding is of interest in the field of pedagogical annotation as it may be indicative of the type of qualities and characteristics that really determine the language educational value of a text or a corpus.

The quantitative gap in terms of number of keywords between annotators, 75 vs 209, mirrors the results for the gap in terms of categories. Although the initial multidimensional study in Figure 5 is focused on quantitative grounds, it does not take into account the type, sense and nature of the different categories applied to the annotation. The categories actually annotated by the subjects can be seen in Figure 8, where the level of coincidence between both annotators can be better appreciated. This intersection represents only 13.5 %, i.e., only 7 out of the 52 applied categories appear in both annotations. Additionally, this percentage is further influenced by the methodology of the experiment, which demanded that the annotators apply a CEF level for each section. If we detract this mandatory category from the pool, only 9.2% represent a common ground. This common group is made up of the following categories: Tense: past simple, future others, future going; Topic: personal identification, plan for the future and, additionally, CEF: A1, A2. In table 8 we can see that categories rarely coincide:



Category Name	Subject A	Subject B	Category Name	Subject A	Subject B
Skill - Listening	N/A	10	Communicative Funct - Expressing Age	N/A	1
Skill - Speaking	N/A	10	Topic - Food	N/A	1
CEF - A1	6	7	Modality	N/A	1
CEF - A2	5	7	Modality - Obligation, Necessity	N/A	1
Topic - Family Members	N/A	5	CEF - B2	N/A	1
Topic - Likes and Dislakes	N/A	5	Topic - Parties	N/A	1
CEF - B1	N/A	4	Topic - Personal Identification (Past)	N/A	1
Topic - Living Routines (Present)	N/A	4	Topic - Pets	N/A	1
Topic - Cultural Information	N/A	3	Topic - Personal Identification (Present)	N/A	1
Tense - Past Simple	1	3	Tense - Future Will	N/A	1
Topic - Plans For The Future	1	2	Communicative Funct - Thanking	N/A	1
Tense - Future Going	1	2	Topic - Wheather	N/A	1
Basic Cohes. Ties/Sent. Organiz. Featu.	N/A	2	Topic - Living Routines	3	N/A
Topic - Cities	N/A	2	Topic - Family	2	N/A
Link Words	N/A	2	Topic - Education	1	N/A
Modality - Ability, Possibility, Permission	N/A	2	Lexical Char - Fixed Expressions	1	N/A
Topic - Living Routines (Past)	N/A	2	Tense - Interrogative Structure	1	N/A
Skill - Reading	N/A	2	Tense - Irregular Verbs	1	N/A
Tense - Present Perfect	N/A	2	Adjective / Adverbs - Ly Adverbs	1	N/A
Tense - Present Simple	N/A	2	Tense - Negative Structure	1	N/A
Skill - Writing	N/A	2	Adjective / Adverbs - Ordering Events	1	N/A
Topic - Personal Identification	1	1	Topic - Places	1	N/A
Tense - Future Others	1	1	Verb + Preposition	1	N/A
Clause Relative	N/A	1	Tense - Regular Verbs	1	N/A
Cond 0 Permanent Truth	N/A	1	Lexical Char - Topic Specific Terminology	1	N/A
Topic - Countries and Languages	N/A	1	Verbs - Using Contractions	1	N/A

Figure 6: Categories annotated by Subject A and Subject B.

Each section was applied a title and a CEF level (Table 9):

Section and CEF Levels	Subject A	Subject B
1	Origins A1	biography A1
2	Holidays A1/A2	countries and languages A1
3	The Past A1	last holidays A1/A2
4	Routines A1	likes and dislikes A1/A2/B1
5	My Family A1	food A2/B1
6	My Home Town A2	my city A1/A2
7	My free time A2	other places I've been A1/A2
8	Plans for the future A2	future plans A2/B1
9	Future with -ing and going to A2	plans B1/B2
10	Contractions A1	my family A1/A2

Table 3: Section titles and CEF levels.

While annotators agree on CEF levels in 8 out of 10 sections, disagreements in sections 5 and 9 are just one level up or down the CEF scale. Annotator B tends to see a wider spectrum of exploitation than Annotator A, who contrarily finds that all of the sections can be used in either A1 or A2 CEF levels. Disagreements in section titles, or to be more precise in section focus, may be attributed to the complex nature of oral, spontaneous and unprepared language. The SACODEYL corpora have all been contributed by young Europeans aged 13-18 who were not given time or indications whatsoever on the nature and topics of the conversations that were recorded and transcribed. This fact may account for the multi-topical nature of these texts and highlight the importance and need for the use of this type of resources in the foreign language classroom.

Conclusions

Different annotators find different categories in the same sections of a text, almost three times as many in the case of Subject B (97 vs. 33). Similarly, they make use of a different repertoire of categories across the text, Subject B presenting again a richer display (38 vs. 21). They assign different keywords to the same categories and the number of keywords assigned to all ten sections is again almost three times bigger for Subject B (209 vs 75). Interestingly, the mean keyword word-length is only slightly dissimilar (2.14 for Subject A vs. 2.87 for Subject B). Our findings therefore show that the pedagogical annotation of a corpus or a text is greatly influenced by what we may call idiosyncratic annotation behaviour. Our case studies point out to the fact that the more experienced teacher is a more prolific annotator, but this view is rather simplistic and may override other interesting findings.

The annotation behaviour of a teacher in our feasibility scenario is influenced by a very rich parametric framework. Our research shows that there is agreement on the annotation when the object of the annotation process is norm-referenced. A good example of this is the CEF Levels, where both annotators found the text sections of similar level for further exploitation in language learning. However, annotation behaviours differ when the individuals are given the chance to annotate the text according to their own criteria. In the experiment both subjects were instructed in the notion that “annotating is the same as to decide which contents you want to see in your textbook, for example, grammar points, communicative functions, lexis, etc. [...]What you annotate can be used for teaching and learning purposes[8]. From this perspective, the resulting taxonomy tree is a reflection of the pedagogy of a given annotator, while the resulting annotated text is a projection of that particular pedagogy which integrates the mediation role played by the annotator/teacher in her interaction with the variables that condition her teaching. The mediation role in the feasibility scenario is therefore conditioned by the annotator’s representation of the learners she is tagging for and the uses that a section, text or corpus will be given in that context. As an illustration, the data we gathered show that, despite the lack of experience in annotating textual resources, both annotators created their own categories in their taxonomy trees, which indicates that they actually took an active role in the assignation of taxonomies. It is interesting to nite that most of the new categories that were created by our informants are concerned with classroom practice and that Subject B is particularly productive here. Besides, the section titles provided by both subjects (Table 9) show that sections 3 and 10 represent different opportunities for language learning, with a traditional grammar focus in the case of Annotator A as opposed to a more topic-oriented bias in Annotator B.

These case studies indicate that the annotation behaviour of teachers may be highly dissimilar and very idiosyncratic. Therefore, we believe that this behaviour can be fully understood only if annotation is profiled. This profile can be achieved by combining some of the measures discussed in the previous paragraph and, in particular, the number of categories applied, the words per section and the number of keywords applied. Thus, we may develop two different Annotation Density (AD) measures: *Category AD* and *Keyword AD*. As it is difficult to establish a comparison among texts of different length, a metric of density can be used to obtain data in which the length of a text does not distort the real sense of the measured value. *Category AD* offers the *weight* which the annotator has given to the categories in a section, irrespective of the length of this section. In Table 10 we can see that the annotators display double CA density in section #3 as compared to those of section #2. Curiously enough, section #2 is longer than section #3. *Keyword AD* is an analogous metric which focuses on the keywords applied to a section, providing the weight which an annotator has given to keywords in a section irrespective of its length. The mathematical description of these metrics could be the following:

$$Category\ AD = \frac{\#of\ Categories\ Applied\ /text}{\#of\ Words/text} \qquad Keyword\ AD = \frac{\#of\ Keywords\ Applied\ /text}{\#of\ Words/text}$$

[8] http://www.um.es/docencia/perez-paredes/2008/05/what-do-annotators-annotate-analysis-of.html

Table 10 illustrates these measures with the data from our research:

Keyword Density (KAD)			**Category Density (CAD)**		
Section	**Subject A**	**Subject B**	**Section**	**Subject A**	**Subject B**
1	0.04	0.18	1	0.03	0.10
2	0.06	0.09	2	0.03	0.04
3	0.08	0.10	3	0.06	0.09
4	0.03	0.10	4	0.01	0.05
5	0.03	0.13	5	0.01	0.04
6	0.03	0.10	6	0.02	0.07
7	0.02	0.19	7	0.01	0.07
8	0.06	0.10	8	0.01	0.03
9	0.04	0.12	9	0.04	0.10
10	0.07	0.15	10	0.02	0.07
Mean value	0.04	0.13	Mean value	0.02	0.07

Table 4: Annotation Density measures.

On average, Subject B applies 0.07 categories per word, while Subject A applies 0.02; Subject B applies 0.13 keywords per word, while Subject A needs 4 words to assign a keyword. These density measures may be a necessary complement to understand the pedagogic quality of corpus-based resources in the language classroom and their potential uses by peer teachers in an environment of teacher collaboration. This is a very interesting area where the social network values attached to expressions such as folksonomies or social tagging (Al-Khalifa and Davies 2006) may converge into the notion of teacher-led pedagogical annotation. By profiling the annotation behaviour of teachers in this way, we may approach the exploitation of corpus-based resources in an informed way and gain insight into (a) the specific mediation role played by a particular annotator and (b) her annotation behaviour.

The aim of pedagogical annotation escapes the kind of automatic tag assignation which is found in morphological tagging. The density measures offered above do not point out per se to better annotation habits. On the contrary, they indicate different approaches to annotation. Becoming aware of these differences is probably a first step towards a future situation where corpus-based resources are shared by the community of professionals in much the same way as other learning resources are tagged and shared in many other fields (Al-Khalifa and Davies 2006). But this sharing effort must be vaccinated against the virus of subjectivity or, in other words, we must make the effort to recognize that subjective appreciations on the uses of language corpora are part of the mediation role played by teachers in bringing corpus resources to the language classroom.

The results of our research confirm that pedagogical annotation is feasible, that the annotation tools that SACODEYL have developed can be used with a very low learning curve, and that annotation behaviour can be profiled. It will take further research to gain insight into the ways in which this profiling may contribute to uses in the language classroom by both learners and teachers. In particular it will be necessary to accomplish further research using a larger group of teachers with a different background and submit their annotations and profiles to the scrutiny of peer teachers.

Widdowson (2003:102) raises a very interesting issue when he uses Michael McCarthy's remark that using a corpus is not just "dumping large loads of corpus material wholesale into the classroom", and goes on to state: "What, then, it is a matter of?" Our feasibility scenario is closer to the mediating corpus advocated by Widdowson (2003) than those in the line of the possibilities scenario. The role of teachers here is crucial. If teachers become annotators, language learners may stand a better chance to become discoverers.

References

Alcaraz, J.M., Pérez-Paredes, P., Mercader, A. and Tornero, E. 2007. "A generic tool for annotating TEI-compliant corpora". Paper presented at the *1st International Conference on Corpus-Based Approaches to ELT*. Universitat Jaume I, Castellón, November 2007.

Al-Khalifa, H. S. and Davis, H. C. 2006. “FolksAnnotation: A semantic metadata tool for annotating learning resources using folksonomies and domain ontologies”. In *Innovations in Information Technology*, November 2006: 1-5.

Aston, G. 1997. “Small and large corpora in language learning”. In *Practical applications in language corpora.* Lodz: Lodz University Press (eds).

Barlow, M. 1996. “Corpora for theory and practice”. In *International Journal of Corpus Linguistics* 1:1-37.

Bernardini, S. 2004. “In the classroom: Corpora in the classroom: An overview and some reflections on future developments”. In *How to Use Corpora in Language Teaching*, Sinclair, J. McH. (Ed), 15–36.

Braun, S. 2005. “From pedagogically relevant corpora to authentic language learning contents”, *ReCALL* 17/1:47-64.

Braun, S. 2006. “ELISA - a pedagogically enriched corpus for language learning purposes”. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, Frankfurt M: Peter Lang. (eds) 25-47.

Braun, S. 2007. “Integrating corpus work into secondary education: from data-driven learning to needs-driven corpora”. *ReCALL* 19/3: 307-328.

Bronwyn B., Dawson, P., Devine, K., Hannum, C., Hill, S., Leydens, J., Matuskevich, D., Traver, C. and Palmquist, M. 2005. *Case Studies. Writing@CSU*. Colorado State University Department of English. Retrieved from http://writing.colostate.edu/guides/research/casestudy/

Fligelstone, S. 1993. “Some reflections on the question of teaching, from a corpus linguistics perspective”. *ICAME Journal*, 17: 97-110.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Mauranen, A. 2004.” Spoken - general: Spoken corpus for an ordinary learner”. In *How to Use Corpora in Language Teaching*, Sinclair, J. McH. (Ed), 89–105.

Pérez-Paredes, P. and Alcaraz, J.M. 2007. “Developing annotation solutions for online data-driven learning”. Paper presented at the *EUROCALL Conference*. University of Ulster at Coleraine, September 2007.

Pérez-Paredes, P., Alcaraz, J.M., Mercader, A., Tornero, E. 2007. “Extracting data from xml annotated corpora: not so mysterious ways into data driven learning (DDL)”. Paper presented at the *1st International Conference on Corpus-Based Approaches to ELT*. Universitat Jaume I, Castellón, November 2007.

Römer, Ute. (Forthcoming). “Corpora and Language Teaching”. In *Corpus Linguistics. An International Handbook*, Lüdeling, Anke & Merja Kytö (eds.). Berlin: Mouton de Gruyter.

Widdowson, H.G. 2003. *Defining issues in English Language Teaching*. Oxford: Oxford University Press.

DDL: REACHING THE PARTS OTHER TEACHING CAN’T REACH?

Alex Boulton[9]

Abstract

The potential applications of electronic corpora in language teaching and learning have received considerable attention in recent years, but their direct application to the classroom has not become part of mainstream practice. Indeed, the majority of published research in data-driven learning (DDL) focuses only on advanced learners using sophisticated equipment for complex language points. But there seems to be nothing essential about including all these ingredients from the start. In particular, DDL in early stages can eliminate the computer from the equation by using prepared materials on paper – considerably easier for the novice learner to deal with.

This paper reports on a simple experiment to see how lower-level learners cope with such paper-based corpus materials and a DDL approach compared to more traditional teaching materials and practices. Pre- and post-tests show both are effective compared to control items, with the DDL items showing the biggest improvement. A questionnaire shows a favourable reaction to the activities.

It is hoped that a larger body of empirical evidence will help DDL break out of its current research confines and into more mainstream teaching practices, especially if it can be shown that DDL has benefits for lower-level learners without expensive resources or extensive training. Simple materials of this sort, if they are seen to be effective, might counter a number of frequent objections to DDL, and contribute to greater awareness of its potential as they require little training, are easily shared, and can be incorporated into published materials.

Keywords: data-driven learning, worksheets, concordance print-outs, lower levels

Background

Electronic corpora have made their mark in many areas connected with language teaching and learning. They can even be used directly by teachers and learners in what Johns (1991a) calls data-driven learning or DDL; but it is notable that such uses have not crossed over into mainstream practice or been taken up by major publishers (Boulton 2008b). There are certainly many barriers to the implementation of DDL, and it may be that work to date has been insufficiently convincing, showing frustratingly small effects (Boulton & Tyne 2008) and concentrating on a minority audience: a survey of fifty empirical DDL studies (Boulton 2008a) found only four conducted outside higher education, and only four with beginning or low level learners. It may even be that current research encourages the belief that DDL is only useful for advanced learners in a computer laboratory, and with experts (i.e. the researchers) devoting considerable time to developing corpora and training learners in small groups.

One particular problem is that researchers have to keep one eye on publishable output by pushing things ever further, so tend to focus on an ambitious hands-on approach to corpus manipulation. It is unsurprising that learners find it difficult to get to grips with new material (the corpora), new technology (the software) and a new approach (DDL) all at once – especially at lower levels of language ability. The methodology itself is “revolutionary” enough (McCarthy 2004: 16) to warrant keeping other things simple, and one way to do this is to take the computers out of the equation at the start: much research has found the technological aspects to be a substantial source of frustration (e.g. Farr 2008), and students may even be “technophobic” (Bernardini 2002).

[9] I’ve been a full-time lecturer at Nancy Université since 1999 and a member of the CRAPEL (Centre de Recherches et d’Applications Pédagogiques en Langues) since 2004, now also part of the ATILF-CNRS. My fascination with lexis in language learning goes back to my PhD in 1998, with increasing emphasis on ICT since then. Corpus linguistics and data-driven learning provide an excellent opportunity to combine the two and, I’m convinced, have enormous potential. I’m particularly concerned with empirical evaluation and experimentation to assess the efficiency of DDL for different learners in different circumstances, and so on. I always try to keep things as simple as possible and look at practical applications for regular teachers and learners – i.e. using free tools and simple techniques. Homepage: http://arche.univ-nancy2.fr/course/view.php?id=967 .

The use of prepared, paper-based materials has other benefits too: no need for a computer laboratory, computers that go wrong, websites that crash, unexpected findings, and so on. Such materials ought to be useful in themselves, and provide a gentle lead-in to direct interaction with the corpus (e.g. Estling Vannestål & Lindquist 2007; Turnbull & Burston 1998). Although this was part of Johns' (1991b) original vision of DDL, there have been only occasional attempts to promote prepared concordance print-outs from publicly available resources (e.g. Chambers 2007); Boulton's (2008a) survey found only eight examples.

Although there is comparatively little research to date with lower level learners and using paper-based materials, what does exist is encouraging (see Boulton 2008a for a summary). This paper reports an empirical study combining the two.

Method

The learners in the study were second-year architecture students in France with no prior experience of DDL. The questionnaire results are based on 71 sets of data and the experimental results on 62, some students being absent at crucial stages. The median age at the time was 19½; most were women (38/62); all but six had French as a mother tongue. Most had been studying English for eight years, though levels are not high: in a start-of-year levels test based on the TOEIC, the average score was only 52.85% overall, corresponding to approximately 450 on the official TOEIC scale, towards the lower end of the "intermediate" band (405-600).10

Prior to the experiment, fifteen common problems were selected from students' own written productions for greater perceived relevance (cf. Seidlhofer 2000). The focus was on grammar/usage as this tends to lend itself to a corpus approach rather than, say, purely grammar (though see Boulton 2007) or formal problems as DDL is generally more suited for depth of knowledge rather than for learning new items (Cobb 1999). The experimental session was conducted during normal class time with the regular teachers. Students were first given a five-minute introduction to corpora and their potential applications based on a specially-prepared booklet. This print document then presented ten of the items (the remaining ones being used as a control): five using corpus data and DDL techniques, five using dictionary entries and traditional teaching methods – dictionaries providing an obvious point of comparison (Yoon & Hirvela 2004). Different items were given different treatment in each group: this was considered the most reliable control as the same students are involved, thus eliminating a number of variables (e.g. Stevens 1991).

The sources used are all available free on line, and easy to use for regular teachers and learners. The corpus was the BYU interface to the British National Corpus (Davies n.d.); the materials mainly comprised selected but unedited KWIC concordances, usually of between five and thirty lines, as well as some information on register, frequency and collocation. The dictionary entries were taken from the monolingual Collins COBUILD English Dictionary for Advanced Learners (2003) and Collins English-French Electronic Dictionary (2005), both available via the Reverso website; the entries were presented in the same layout as the original. Every attempt was made to produce equivalent materials: one page for each language item, the information being interspersed with questions to focus attention on particular points.

Prior to the experiment, teachers had a one-hour training session on DDL, at the end of which not all were convinced it would work. For the DDL treatment, they were encouraged to stick to the format by allowing students to discuss the questions and data in small groups to reach their own conclusions on each item before class feedback. For the traditional presentation, the teachers were allowed to intervene as they saw fit on a traditional "knowledge transmission" model. The experimental session itself lasted one hour of a 90-minute class. It took between five and ten minutes to go through each item, although all teachers were surprised that the DDL treatment did not take substantially longer than the traditional treatment.

Knowledge of the ten language items, along with five others as a control, was assessed the week before the experimental session and three weeks later to test for recall. The test of thirty questions (two for each item) was in a familiar format based on the TOEIC part V: gap-fill sentences with four possible choices each, the contexts mainly derived from dictionaries and other teaching materials. Additionally, to assess reactions to the materials and methods used, students were asked at the end of the experimental session to complete a short questionnaire in

[10] The students' objective is an official score of 700 points by the end of their third year as a precondition to obtaining their architecture diploma.

French combining closed questions on a five-point Likert scale, and open questions to be completed in their own words.

Results and discussion

Test results

The overall scores are fairly low (only 14.3/30 in Test 1) even though the test questions were based on clear answers. This suggests the items in question are, as intended, problematic for these learners. The highest scores were 25 out of 30 in Test 1 and 27 in Test 2; the lowest were six and eight respectively. The average scores increased from 14.6 in Test 1 to 17.4 in Test 2 – an improvement of 2.8 points, or 19.4% (Table 1).

	DDL /10	Traditional /10	Control /10	Total /30
Test 1	4.9	4.6	5.1	14.6
Test 2	6.4	5.7	5.3	17.4
difference	+1.5	+1.0	+0.3	+2.8
change	+31.6%	+22.2%	+5.1%	+19.4%

Table 1. Average scores, by treatment.

There are two main ways to compare the data: in Table 2, the horizontal arrows show changes between tests; the vertical arrows compare different treatments. Taking the first of these, a two-tailed paired t-test shows there to be a significant improvement overall between tests (p<0.0001). One possibility is that there may simply have been a "test effect", with students scoring higher the second time simply as they become more used to the test design and what was required. There was indeed a small improvement in the control items of 5.1%, but this is not significant (p>0.5). This means that the significant improvement must derive from the other items: a 22.2% increase in score for the traditional items (p<0.01), 31.6% for the DDL ones (p<0.0001). The first conclusions therefore are that the test effect is minimal, while both kinds of presentation do have a significant effect.

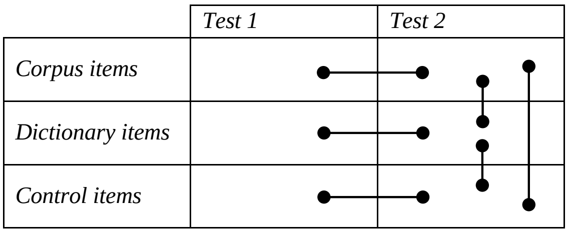


Table 2. Points of comparison.

The key question however lies in the vertical comparisons of Table 2, i.e. between the different types of presentation. Again using t-tests, there is no significant difference between the control items and the dictionary items (p>0.01), nor between the dictionary items and the corpus items (p=0.15); but there is a significant difference between the corpus and control items (p<0.001). Although the DDL treatment was more effective than the traditional treatment, there is a 15% likelihood that this could be due to chance alone.

Another point of comparison can be made between students' level, as measured by the start-of-year TOEIC scores, and the test results. Pearson's product-moment coefficient shows a strong positive correlation with both: 0.82 with Test 1 and 0.77 with Test 2. It is also possible to compare levels against the performance on the three types of items in Test 2. Unsurprisingly, the correlation was strongest (0.76) for the control items: as these were not explicitly covered in class, the students could only draw on their previous knowledge of the language. The coefficient is lower but still substantial (0.54) for the traditional items; in other words, it can be inferred that more advanced students gained greater benefit from using dictionaries and traditional teaching. On the other hand, the correlation is virtually non-existent for the DDL items (-0.13), suggesting that all levels benefited as much as each other from this type of information and approach.

Questionnaire results

The first four items on the questionnaire were closed questions, asking the students to compare the two approaches they had just experienced on a five-point Likert scale. Looking only at the positive results (agree or strongly agree), 30 of the 71 students found the dictionary work easy compared to 54 for the corpus work; 31 found the dictionary work useful, compared to 59 for the corpus work. 37 thought the dictionary work would help them avoid certain errors in the future (suggesting they felt they had learned something from the work), rising to 58 for the corpus work. Clearly the traditional treatment was less positively received overall: this is reflected in the final pair of questions, as only 28 students would like to do more dictionary activities in the future, while 51 would like to pursue the DDL work.

To see how favourably DDL was received at different levels, the learners were divided into three bands according to their start-of-year TOEIC test: the upper level corresponds to a TOEIC average of 68%; the middle level to 51%; the lower level to 39%. The responses are largely very positive for all categories; the highest band is perhaps slightly more receptive than the others, but the patterns of differences are not significant.

Two open questions allowed the students to say what they felt were the respective advantages of dictionaries and corpora. Dictionaries were considered most useful for new or unknown words (26) and for meanings or definitions (26), while 19 simply wanted translations. 20 were interested in usage information; for some this was best presented in the form of “rules”, while others preferred looking at the examples – although one particularly wanted meanings “independent of any context”.

Corpora, on the other hand, were felt to be most useful for the contexts and “concrete examples” which highlight usage and grammar (58), and to represent “practical English”, “frequent usage”, the “language of today”. Only six mentioned “formulae” or “idiomatic expressions” as such, though allusion to context and, more specifically, “words that go together” reveals a certain sensitivity to this. Most responses seem to refer to corpus use for productive purposes, although some explicit reference was also made to comprehension (13). Some were extremely enthusiastic, including the following:

Very interesting, an experience to repeat several times with other usage difficulties.

I’d never heard of corpora. Thank you!

It’s the first time I’d done this type of exercise – but none too soon! Thank you! I’ll assimilate things better this time! (Now go and kick out the teachers in high school!!!)

The final closed question asked the students if they would prefer to explore corpora on their own on computer rather than via the intermediary of paper-based materials. Although the students had no experience of hands-on computer-based DDL, they showed comparatively little interest for this: only 20 of the 71 students (i.e. less than 30%) agreed or strongly agreed. It is worth noting that the highest of the three ability levels was least keen on such an approach: 17%, compared to 38% of the middle group and 35% of the lowest. 55 of the students took the opportunity to explain why: nearly half (25) believed the prepared exercises would get straight to the point and avoid time-wasting, and teacher guidance would be essential to avoid drawing wrong conclusions from the mass of data. As two students pointed out, they would need to try hands-on DDL first, but two others simply found the possibility “unattractive”. Two felt that talking about things was a useful part of the activity rather than just sitting in front of a computer, while eight thought that “doing it themselves” would be more relevant, motivating and lead to more effective learning. More generally, many stressed the importance of context and felt that the numerous samples would help to “visualise” or get a “feel” for the items under study, whether via prepared materials or on their own.

Conclusions

The experiment reported here with learners at lower levels of language ability found that, with no prior training, they managed to gain significant benefit from prepared, paper-based DDL materials. In particular, they performed better with this approach than they did using dictionary entries and traditional teaching methods.

On the whole, these results seem to contradict the received wisdom that DDL is best reserved for more advanced, sophisticated learners, despite a number of counter-examples (e.g. Boulton in press; Yoon & Hirvela 2004). It certainly needs further exploration, though it does corroborate some of the findings of our earlier research (e.g. Boulton forthcoming). The interpretation offered then was that more advanced learners have reached their current level by traditional means – in other words, they are comparatively good with the system currently in place. Learners who have been through the same system but who come out with lower levels are, by definition, not as good at learning through traditional methods.

In fairness, the objections to DDL at lower levels are usually within the context of hands-on exploitation of corpora, and do not necessarily extend to the use of paper-based DDL materials. This is a potentially crucial point: if the findings here are confirmed, it suggests not only that DDL can be successful using paper-based materials, but that these can be used with a wide range of levels, and may thus serve as a stepping-stone to hands-on corpus exploration (Johns 1997: 113).

Most current research publications seem to throw learners in at the deep end, requiring them to master the concept of corpora, the software and DDL techniques all at once. However, it is still DDL even if learners initially work only with paper-based materials and not directly on a computer (Breyer 2006; Frankenberg-Garcia 2005); this makes the learners' task considerably easier as it reduces a number of methodological obstacles by, among other things, reducing the amount of data and limiting the range of possible answers (Thompson 2006) – not to mention technical, logistical and financial obstacles for the teacher. A proven track record for such paper-based materials might also help to convince a wider public that surprisingly little investment is required for rapid and substantial returns.

Paper-based materials, we have seen here, can bring benefits in themselves (Chambers 2005: 121). They can also serve as an introduction before moving on to more autonomous corpus exploration: clearly learners need to understand the nature of corpus data and analysis before they can explore on their own, and need guidance in their use – autonomy does not come automatically to all (O'Sullivan 2007). As Sun (2003: 609) points out, "the learning curve… is arduously steep, in that students tend to get confused easily about the concordancer outputs; thus, they need either a stronger degree of teacher involvement, or to learn in a more structured environment"; using paper print-outs may reduce some difficulties in early stages. Furthermore, learners such as ours may initially feel paper-based resources are more relevant or efficient and, as Whistle (1999: 77) puts it, simply have difficulty seeing "why the concordances could not be prepared in advance and handed out in class".

DDL materials such as those used here are extremely time-consuming to produce: each of the items here required half a day's work – compare to Johns' (1991a) eight hours for a single handout. Clearly such investment cannot be expected in normal teaching contexts, and yet there are virtually no published materials available (Boulton 2008b): of the eight empirical studies using paper-based materials reported in Boulton (2008a), all but one had to create these materials themselves. Even downloadable worksheets remain scarce, are not necessarily transferable to new contexts, and are dependent on researchers' goodwill. Greater research interest producing positive results might inspire publishers to produce materials in the area in the form of books or paying websites; integrating DDL activities into more general works; or including corpora and interactive tools on websites or DVD-ROMs which accompany their publications.

References

Bernardini, S. 2002. "Exploring new directions for discovery learning." In Teaching and Learning by Doing Corpus Analysis, B. Kettemann & G. Marko (eds.). Amsterdam: Rodopi: 165-182.

Boulton, A. 2007. “DDL is in the details… and in the big themes.” In Proceedings of Corpus Linguistics 2007, M. Davies, P. Rayson, S. Hunston, & P. Danielsson (eds.). http://www.corpus.bham.ac.uk/corplingproceedings07/ [Access date 05/2008]

Boulton, A. 2008a. “Evaluating corpus use in language learning: state of play and future directions.” Paper presented at the American Association of Corpus Linguistics. Provo, March 13-15 2008.

Boulton, A. 2008b. “‘Off-the-peg’ materials for data-driven learning.” Paper presented at the New Trends in Corpus Linguistics for Language Teaching and Translation Studies: in honour of John Sinclair, Granada, September 22-24 2008.

Boulton, A. forthcoming. “Testing the limits of data-driven learning: language proficiency and training.” ReCALL.

Boulton, A. in press. “Looking for empirical evidence of data-driven learning at lower levels.” In Practical Applications of Language and Computers: PALC07, B. Lewandowska-Tomaszczyk (ed.). Frankfurt: Peter Lang.

Boulton, A. and Tyne, H. “Learning with corpora: changing learning practices.” Paper presented at the 4th Inter-Varietal Applied Corpus Studies (IVACS) group Conference: Applying Corpus Linguistics. Limerick, June 13-14 2008.

Breyer, Y. 2006. “My Concordancer: tailor-made software for language learners and teachers.” In Corpus Technology and Language Pedagogy: new resources, new tools, new methods, S. Braun, K. Kohn & J. Mukherjee (eds.). Frankfurt: Peter Lang, 157-176.

Chambers, A. 2005. “Integrating corpus consultation in language studies.” Language Learning & Technology 9/2: 111-125. http://llt.msu.edu/vol9num2/chambers/ [Access date 02/2006]

Chambers, A. 2007. “Language learning as discourse analysis: implications for the LSP learning environment.” ASp 51-52: 35-51.

Cobb, T. 1999. “Breadth and depth of lexical acquisition with hands-on concordancing.” CALL 12/4: 345-360. http://www.er.uqam.ca/nobel/r21270/cv/Breadth.htm [Access date 03/2006]

Collins Cobuild English Dictionary for Advanced Learners, 4th edition. 2003. HarperCollins. http://dictionary.reverso.net/english-cobuild/ [Access date 01/2008]

Collins English French Electronic Dictionary. 2005. HarperCollins. http://dictionary.reverso.net/french-english/ and http://dictionary.reverso.net/english-french/ [Access date 01/2008]

Davies, M. n.d. BYU-BNC: British National Corpus. http://corpus.byu.edu/bnc/x.asp [Access date 01/2008]

Estling Vannestål, M. and Lindquist, H. 2007. “Learning English grammar with a corpus: experimenting with concordancing in a university grammar course.” ReCALL 19/3: 329-350.

Farr, F. 2008. “Evaluating the use of corpus-based instruction in a language teacher education context: perspectives from the users.” Language Awareness 17/1: 25-43.

Frankenberg-Garcia, A. 2005. “Pedagogical uses of monolingual and parallel concordances.” ELT Journal 59/3, 189-198.

Johns, T. 1991a. “Should you be persuaded: two examples of data-driven learning.” In Classroom Concordancing, T. Johns & P. King (eds.). English Language Research Journal 4: 41-16.

Johns, T. 1991b. “From printout to handout: grammar and vocabulary teaching in the context of data-driven learning.” In Classroom Concordancing, T. Johns & P. King (eds.). English Language Research Journal 4: 27-45.

Johns, T. 1997. “Contexts: the background, development and trialling of a concordance-based CALL program.” In Teaching and Language Corpora, A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (eds.). Harlow: Addison Wesley Longman, 100-115.

McCarthy, M. 2004. Touchstone: from corpus to coursebook. Cambridge: Cambridge University Press. http://www.cambridge.org/us/esl/Touchstone/teacher/images/pdf/CorpusBookletfinal.pdf [Access date 04/2008]

O’Sullivan, I. 2007. “Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy.” ReCALL 19/3: 269-286.

Seidlhofer, B. 2000. “Operationalizing intertextuality: Using learner corpora for learning.” In Rethinking Language Pedagogy from a Corpus Perspective, L. Burnard & T. McEnery (eds.). Frankfurt: Peter Lang, 207-223.

Stevens, V. 1991. “Concordance-based vocabulary exercises: a viable alternative to gap-filling.” In Classroom Concordancing, T. Johns & P. King (eds.). English Language Research Journal 4: 47-61.

Sun, Y-C. 2003. “Learning process, strategies and web-based concordancers: A case-study.” British Journal of Educational Technology 34/5: 601-613.

Thompson, P. 2006. “Assessing the contribution of corpora to EAP practice.” In Motivation in Learning Language for Specific and Academic Purposes, Z. Kantaridou, I. Papadopoulou & I. Mahili (eds.). Macedonia: University of Macedonia. http://www.reading.ac.uk/internal/appling/thompson_macedonia.pdf [Access date 05/2008]

Turnbull, J. and Burston, J. 1998. “Towards independent concordance work for students: Lessons from a case study.” ON-CALL 12/2: 10-21. http://www.cltr.uq.edu.au/oncall/turnbull122.htmll [Access date 04/2006]

Whistle, J. 1999. “Concordancing with students using an ‘off-the-web’ corpus.” ReCALL 11: 74-80.

Yoon, H. and Hirvela, A. 2004. “ESL student attitudes toward corpus use in L2.” Journal of Second Language Writing 13/4: 257-283.

DISPOSABLE CORPUS IN TRANSLATOR TRAINING: TRANSLATING MEDICAL ABSTRACTS INTO L2

Adauri Brezolin[11]

Abstract

It is widely accepted that translator training lacks research on the teaching of translation into L2. The situation seems to stem from a consensus among several theorists who firmly oppose to such directionality, claiming that, in broad terms, translating into L1 is the norm. However, it is well known that, in practice, professional translators and translation students render texts into L2. This paper presents an activity based on the compilation of a disposable corpus (Varantola 2003), involving the translation of scientific abstracts into L2 (Portuguese-English).

The activity with pedagogical and methodological implications is divided into two parts. In the first, students are exposed to the typical structure of a scientific abstract in English to get familiarized with its main sections (Koltay 1996), by paying close attention to their lexical, syntactic, semantic and collocational features. Medical abstracts are carefully selected from highly-ranked journals indexed at the PubMed Central.

In the second part, students are required to: i) compile a disposable corpus of 10.000 to 15.000 words (an average of 50 abstracts with around 200/300 words), using the same selection criteria and from journals of the PubMed Central; ii) run the corpus using AntConc, which provides word, keyword, and concordance lists, among other utilities; iii) prepare concordance lists with terms representative of each sectiont of an abstract in order to check the possibilities offered by the corpus; and iv) render an abstract from Portuguese into English, incorporating as much knowledge as possible acquired along the process.

In sum, students have the opportunity to i) improve their translational competence, which will help them in future situations when they have to act strategically before a problem, for instance; and ii) develop their translational performance by using appropriate tools when they have to manage all types of knowledge they have acquired.

Keywords: Corpus Linguistics, disposable corpus, Portuguese-English translation, reverse translation, translation education.

Introduction

Applied Translation Studies seems to have given little attention to the teaching of translation into L2. This lack of research in the area of translator training must be the result of a consensus among several theorists who firmly oppose to such directionality, claiming that, in broad terms, translating into L1 is the norm.

This situation apparently arises from two different premises: professional translators invariably work into L1 and most of them are not proficient to produce acceptable texts in L2; both applying to translator trainees as well. According to St. John (2003), “there is an unwritten golden rule that translations need ideally be done by native speakers of the TL culture and consequently, the vast majority of professional translation is probably done in this way”. Jeffcote (2005) also comments on the situation, “(t)ranslating into L2 has always been considered a risky, if not undesirable activity. For easy to comprehend reasons, translators are usually trained, and employed, to work into their mother tongue”. About the second possible reason for L1 being the norm, Stewart (2000) claims that “negative attitudes to translating into L2 spring mainly from the premise that translators do not have sufficient linguistic expertise to work into a language which is not their own”, this is corroborated by Jeffcote (2005), who reinforces the idea that “there is a sense of taboo surrounding any attempt to translate into a language that a translator does not possess a native or near native ability”.

It is well known; however, that, in practice, both professional translators and translation students render texts into L2 (Harvey, 1996; Stewart, 2000; St. John, 2003; Jeffcote, 2005). In Brazil, the situation is not different: professional translators have to translate into L2 and training translators are often exposed to reverse translation activities. This paper, then, aims to present an activity based on the compilation of a disposable corpus (Varantola, 2003), involving the translation of scientific abstracts into L2 (Portuguese-English). The activity with pedagogical

[11] Adauri Brezolin holds a master’s and a doctor’s degree in Languages (USP, São Paulo) and has been teaching translation at undergraduate courses for over 20 years, and for many years also taught English in language centers. At present, teaches translation and terminology at Universidade Metodista (São Paulo, Brazil). He is also the co-author of Pequeno Dicionário de Expressões Idiomáticas e Coloquialismos (Fiúza, 2001) and Whatchamacallit? - Portuguese-English dictionary of idioms and colloquialisms (Disal, 2006). In his spare time, he does freelance translations from English into Portuguese and proofreading in Portuguese. The teaching of translation, terminology and Corpus Linguistics are of great interest to him.

and methodological implications is divided into two parts: 1. students are exposed to the typical structure of a scientific abstract in English; 2. students are required to: i) compile a disposable corpus of 10.000 to 15.000 words ii) run the corpus using the freeware concordance program AntConc; iii) prepare concordance lists with terms representative of each part of an abstract and iv) render an abstract from Portuguese into English. Ultimately, this activity intends to compensate for such linguistic paucity and to improve students' proficiency in translation.

Exposing students to the structure of a scientific abstract

As previously stated, in the first part of the activity, students are exposed to the typical structure of a scientific abstract in English, so that they can get familiarized with its main sections. Before analyzing real examples of abstracts in English, the topic is elucidated by Koltay's ideas (1996) about what each section of an abstract - Introduction, Objectives, Methods, Discussion, Results, Conclusion – should contain. Then, nearly ten abstracts are presented to students, when they are expected to recognize each part and to pay close attention to the lexical, syntactic, semantic and collocational features proper to each section. We make sure that this short collection of abstracts is collected from top-notch quality journals.

Due to the idiosyncrasies of the Portuguese language, students are reminded of certain "traps" they may easily be caught in. In lexical terms, for instance, they may be misled by the item "trabalho", which is not "work" (a prima facie translation), but "paper", "article", "study" etc. Syntactically, they may observe if active rather than passive voice is preferred in certain parts of the abstract. From the semantic point of view, they may refresh their memory on false cognates, such as "pretender" (to intend) and not "to pretend". As to collocational features, they may have the chance to get in touch with more idiomatic cohesive mechanisms, such as "Overall", "In conclusion", among others.

Compiling a disposable corpus of medical abstracts

Here, we assume that Corpus Linguistics, "as a methodology which focuses on the identification of recurrent patterns of linguistic behaviour in actual performance data, provides the appropriate tool to test hypotheses about norms and regularities in translated texts." (Bernardini, Stewart & Zanettin 2003); abstracts in general seem to be a very suitable text type to apply such methodology on.

Disposable corpora or ad hoc comparable corpora, as Varantola (2003: 55) puts it, "are typically collected for a single translation assignment, i. e. to help in the translation of particular texts. Since these ad hoc corpora are collected to satisfy a transitory need, they are not primarily aimed at forming a part of a permanent text corpus".

The disposable corpus used for the present study was compiled from abstracts carefully selected from highly-ranked journals indexed at the PubMed Central (the U.S. National Institutes of Health free digital archive of biomedical and life sciences journal literature). Our selection criteria included: i) author or authors of abstracts had to be affiliated to a North-American university; ii) abstracts should be a single paragraph only, with no explicit indication of their sections, and iii) abstracts should contain from 200 to 300 words. When students are asked to compile their own corpora, similar criteria are recommended.

In order to compile our disposable corpus, we decided on five key words deemed as representative of subjects of interest in recent times. From these five key words; namely breast cancer, dengue, lung cancer, obesity and stem cells; we selected 10 abstracts of each subject. This way our corpus came to 50 texts and amounted to 11.468 word tokens.

The program used to run the corpus is AntConc, a freeware concordance program for Windows, Macintosh OS X, and Linux, developed by Laurence Anthony (Faculty of Science and Engineering, Waseda University, Japan). Like other programs of this kind, AntConc provides concordance lists, concordance plots, clusters, collocates, word lists, keyword lists, among other utilities. Unlike other programs that offer only demo versions at no cost, AntConc is freely downloadable at http://www.antlab.sci.waseda.ac.jp/software.html. It is important to mention that, at least for this specific activity, students felt AntConc user-friendlier than WordSmith Tools, which they had used before. Even those who have low software literacy were capable of coping with the tasks they were expected to perform. To our understanding and experience; however, WordSmith Tools provides a wider range of resources when compared with AntConc.

Using a disposable corpus to render medical abstracts

After compiling the corpus, its wordlist was obtained by using AntConc. The program automatically generates a list in frequency order. Following is a sample of the wordlist:

1	476	of
2	404	the
3	386	and
4	275	in
5	219	to
6	187	a
7	144	with
8	131	for
9	112	cells
10	104	that
11	92	was
12	81	by
13	81	is
14	75	cancer
15	73	cell
16	72	virus
17	70	were
18	68	or
19	64	The
20	61	dengue

Wordlist sampling (20 out of 2786)

As expected, the program produces a list containing both grammar and content words; however, for a study like ours, where phraseology rather than terminology is under scrutiny, subject-specific content words and some grammar words were removed from the wordlist. A sampling of the resulting list is below:

11	92	was
12	81	by
13	81	is
17	70	were
19	64	The
25	43	are
28	42	have
29	39	this
31	34	we
33	33	these
39	29	In
44	27	studies
52	22	We
57	20	results
64	18	can
66	18	has
73	17	This
75	17	used
88	15	data
101	14	study

Sampling of wordlist after the removal of some words (20 out of 2786)

Even after the removal of some words, we can notice that some content words are still on the list. This decision is based on the fact that those content words are commonly found in any type of abstract; as we see it, they are specific to certain sections of abstracts in general, and not subject-specific. We believe that, whenever a corpus is primarily analyzed, intuition is still very important to move on to statistical evidences. This classroom activity aims to stimulate "corpora awareness" in students. As a rule, a corpus offers a large amount of overt information, but it also suggests another large amount of covert information. It is up to the analyst to explore this wealth of untapped information; translation teachers/trainers can help their students develop this kind of awareness.

Checking for recurrent patterns

Then, students are asked to produce concordance lists by using, as search term, words representative of the sections of an abstract. In order to illustrate this step of the activity, four items will be used: "we", "were", "this", and "results".

```
     As an alternative, we developed functional assays
          triole amplification. We evaluated 22 missense mutati
          n both cell types. Here we examine these parameters in one
         eir normal counterparts. We found that the human colon cancer
        e to anastrozole treatment, we have used an intratumoral aromatase
       weeks of treatment. We therefore investigated whether
       n human breast tumors. Here, we show that CA12 is robustly regulated by
   molecular G-quadruplex. Here, we demonstrated that the G-rich strand in
     engue virus immune complexes, we generated native and signaling-incompet
     assay and flow cytometry. We found that both receptors mediated
```

Concordance lists with "we" (10 out of 56 hits)

```
NA binding domain (DBD) that were identified in multiple breast cancer
D-18Co did not. When these lectins were tested for their effects on cell viability in
ll viability in culture, both cell lines were affected by the lectins but at 6, 48 and
ity, 272 mothers of the NHSII participants were asked to report information on their daughters'
confidence intervals; all statistical tests were two sided. Intake of isoflavones was associ
hemokines by human mast cells were examined. Elevated levels of secrete
nd MIP-1?, but not IL-8 or ENA-78, were observed following infection of KU812 or
 degranulation. Chemokine responses were not observed when mast cells were treated
were not observed when mast cells were treated with UV-inactivated dengue virus
 of such dengue virus-permissive cells, were compared for their influence on the infectivity
```

Concordance lists with "were" (10 out of 70 hits)

Considering the widespread idea that, in scientific texts, passive voice is preferred to active voice, we can analyze the concordance lists with "we" and "were", and see that both voices are, in fact, used. For example, the verb "to examine" appears in both cases ("Here we examine"/"human mast cells were examined"). Students can be asked to look deeper into these occurrences in order to check the sections in which they are most frequent; if they appear more in Objectives or in Methods, for instance. This can help translators decide to whether change or not a structure that may be more used in a language than in another; that is, to decide on a more acceptable structure from the viewpoint of a certain type of discourse in a certain language. It seems that abstracts in Brazilian-Portuguese and American-English share more similarities than differences.

```
airplane travel. In this paper we describe the
tralization sites. In this study the ED3 epitopes
   response. In this study, we sought to
  sponse. In this study, we sought
 ote exposures. This study determined the
74 (0.45-1.20). In this meta-analysis of case-
ion to the EGFR. This approach highlights a
pathways in liver. This review will focus upon
nd liver-enriched NRs. This review will also highlight the
```

```
tes. The objective of this study was to quantify the
elerated fashion. This protocol also includes
genes into NOD mice. This protocol also discusses imp
```

Concordance lists with "this" (10 out of 56 hits)

By asking students to analyze a word as simple as "this", we can provide them with some collocations for the piece of scientific writing in question. Depending on the type of investigation, on the original word in the source language, and on the section of the abstract, we can see that "paper", "study", "approach", "review", "protocol" can be used; and most importantly, we can see what verbs collocate with each noun, for example, "This review will focus upon", "This approach highlights"; and so on.

```
cer cell line. The results suggest that it 3
        r growth. The results showed that the best
vestrant alone. These results suggest that blocking
MDA-MB-231. Collectively, our results provide evidence that specific
ruplex stabilization. Our results also provide further su
o dengue virus. These results suggest a role for ma
n also induced NF-?B. These results indicate a role for the dengu
es viral genomic sequences. These results suggest a novel role for YB-1 as a
irect ELISA and the results indicate that all se
 replication rates. Our results suggest that enhanc
unoprecipitation assay. These results demonstrate for the first time
ffinity-MRM method and the results were comparable with thos
st 6 months. Our results indicate that wit
MDSCs. Although these results raise questions as to w
```

Concordance lists with "results" (10 out of 20 hits)

Again, by examining the concordance lists with "results", we can see what sort of company the word keeps, that is, the words that typically co-occur with it. This can also be obtained from another utility provided by AntConc: collocates.

Rank	Freq.	Freq.(L)	Freq.(R)	Collocate
1	20	0	0	results
2	5	5	0	These
3	5	0	5	suggest
4	3	3	0	the
5	3	3	0	The
6	3	3	0	Our
7	3	0	3	indicate
8	2	2	0	these
9	2	2	0	our
10	1	0	1	with

This way, we can learn that we can use "these/the/our results", that "our results suggest/indicate", among other combinations.

Concluding remarks

By the time students are asked to translate an abstract from Portuguese into English, they naturally incorporate much of the knowledge acquired along the process, when they have the chance to compensate for certain deficiencies not only in lexical, syntactic and semantic, but also in collocational terms. In general, students have the opportunity to develop and to improve their translation proficiency. Besides that, translation trainers should bear in mind the following: i) never take for granted that students know what they are expected to do; so, involve them in an activity which presupposes step-by-step guidelines (corpus compilation included); ii) though students may feel discouraged by a reverse translation activity when most of them lack linguistic proficiency in L2, an activity as such may serve the purpose of testing his/her knowledge and of discovering what their weaknesses are; iii) though the use of corpora in teaching can be criticized by promoting the standardization of translated texts and by inhibiting creativity; in reality, our focus is on recurrent structures of a particular text type. Finally, “L2 translation is a widespread and often indispensable activity world-wide” (Stewart, 2000), and “the knowledge of how to compile and use corpora is an essential part of modern translational competence and should be dealt with in the training of prospective professional translators”. (Varantola 2003: 56)

References

Bernardini, S.; Stewart, D. & Zanettin, F. 2003. “An Introduction” In: Corpora in Translator Education, F. Zanettin; S. Bernardini & D. Stewart (eds.). Manchester/Northampton: St. Jerome, 01-13.

Harvey, M. 1996. “A Translation Course for French-speaking Students” In: Teaching Translation in Universities – Present and Future Perspectives, P. Sewell & I. Higgins (eds.). London: CILT, 45-65.

Jeffcote, C. 2005. “Teaching Specialised Translation into L2 - Standing the Pyramid on its Head”, International Conference on Translation and Interpretation, September 9-11, 2005, Monterey, California Theme Professional Education of 21st Century Translators and Interpreters

http://gsti.miis.edu/conference/cabcj.htm [Access date 10/15/2005]

Koltay, T. 1996. “Professional documentation for students of English translation: approaches and methods”, Folia Practico-Linguistica. XXV-XXVI.: Theoretical and practical aspects of training translators, 139-147.

St. John, E. 2003. “Translating into L2 during Translator Training”,

http://64.233.169.104/search?q=cache:ygTf6E5VBGoJ:isg.urv.es/cttt/cttt/research/stjohn.doc+%22translator+training%22+%2B+reverse+translation&hl=pt-BR&ct=clnk&cd=1&gl=br [Access date 10/09/2007]

Stewart, D. 2000. “Supplying Native Speaker Intuitions or Normalising Translation? Translating into English as a foreign language with the British National Corpus.” http://www.art.man.ac.uk/SML/ctis/events/Conference2000/corpus1.htm [Access date 10/15/2005]

Varantola, K. 2003. “Translators and Disposable Corpora”. In: Corpora in Translator Education, F. Zanettin; S. Bernardini & D. Stewart (eds.). Manchester/Northampton: St. Jerome, 55-70.

THE ECPC ARCHIVE: A GATEWAY TO THE MERGING OF CORPUS BASED TRANSLATION STUDIES AND CRITICAL DISCOURSE ANALYSIS[12]

María Calzada Pérez[13]

Abstract

The present paper introduces the ECPC research group and the electronic archive of parliamentary speeches (from the European Parliament, The Spanish Congreso de los Diputados and the British House of Commons) that the group is compiling and developing. This archive may be seen as valuable material to delve into the "jigsaw puzzle" of the parliamentary genre and of its three main pieces: qualitative, quantitative and pedagogical. Hence, the paper puts forward a CDA methodology to approach (mono and bi-, multilingual) data in a qualitative manner. Then a corpus-based methodology is advocated and briefly sketched to add quantitative rigor to qualitative results. Finally, DDL is chosen to convey the quantitative and qualitative potential of the ECPC Archive within the translation class. Following Johns (1991, quoted in McEnery, Xiao and Tono, 2006), the paper provides pedagogical exemplifications that may be arranged in three consecutive stages: observing, classifying and generalizing.

Keywords: describing paper: CDA, Corpus-based work, DDL, parliamentary genre, translation studies.

The ECPC Group: putting together a jigsaw puzzle

The European Comparable and Parallel Corpora (ECPC) is a relatively new research group working on the compilation, analysis and pedagogical exploitation of an archive of speeches from three different parliaments across Europe: The European Parliament (EP), the Spanish Congreso de los Diputados (CD) and the British House of Commons (HC). ECPC is a multi-national, multi-disciplinary group spread across Europe with members based in the United Kingdom (Mona Baker, Gabriela Saldanha, Marion Winters), Ireland (Dorothy Kenny, Saturnino Luz) and Spain (María Calzada Pérez, Rosa Agost, Pilar Jara, Noemí Marín, José Manuel Martínez). During the past three years, the group has been busy putting together the main ECPC archive consisting so far in:

ECPC_EN: English version of all (contextually-marked-up) speeches delivered at the European Parliament in 2005.

ECPC_ES: Spanish version of all (contextually-marked-up) speeches delivered at the European Parliament in 2005.

ECPC_CD: All (contextually-marked-up) speeches delivered at the Spanish Congreso de los Diputados in 2005.

ECPC_HC: All (contextually-marked-up) speeches delivered at the British House of Commons in 2005.

As may be seen, all corpora have been automatically XML-marked-up in order to sign-post:

The exact date when speeches were uttered before their respective parliaments (<date>Lunes 10 de enero de 2005 </date>);

[12] The following paper has been carried out within the framework of the EuropeanComparable and Parallel Corpora (ECPC) research project, funded by the Spanish Ministry of Education and Science (HUM2005-03756)"

[13] María Calzada Pérez is a lecturer in translation and English at Jaume I University (Castellón, Spain). She studied at the University of Granada and the University of Essex and completed her PhD at Heriot-Watt University (Edinburgh, UK). She is author of the following monographs: La aventura de la traducción: Dos monólogos de Alan Bennett (2001, on translating Alan Bennett's Talking Heads); Transitivity in Translating. The Interdependence of Texture and Context (2007); and El espejo traductólogico. Teorías y didácticas para la formación del traductor (2007, a review of theoretical and descriptive approaches to the practice and teaching of translation). She is editor of Apropos of Ideology: Translation Studies on Ideology – Ideologies in Translation Studies. She has also published widely in prestigious journals such as The Translator, Target, Meta, Text, Babel, Perspectives: Studies in Translatology, etc. Finally, she has translated extensively, including Jonathan Fine's (2006) Language in Psychiatry. A Handbook of Clinical Practice (El lenguaje en psiquiatría. Un Manual de práctica clínica, 2007). Her main research interests revolve around translation and ideology, corpus studies and European Parliamentary texts, advertising and pedagogy.

The topics discussed as part of the agenda (<indexitem number="1">Reanudación del período de sesiones</indexitem>);

The types of speakers (<chair>: President; <intervention ref…>: other participants).

Political affiliations (<affiliation EPparty="UEN"/>)

The post of each speaker (<post>President</post>)

The gender of each speaker (<gender> male </gender>)

The status of each speaker (<status>Dr</status>)

Comments about the parliamentary sessions, which are not part of the actual speeches uttered (<omit> applause</omit>)

Apart from its four main XML-marked-up corpora, the ECPC archive also comprises 4 additional untagged corpora which were compiled as transition material and which include EP speeches in English and Spanish from 1996 to 2003.

All in all, the ECPC archive has a twofold purpose. On the one hand, it has served as "raw material" to develop a free, online, monolingual and bilingual parallel concordancer (ConcECPC 1.0, which will be updated in subsequent years), with which not only this archive but other corpora may be queried. The present paper will not discuss this first purpose and, for more information on it, see Calzada Pérez and Luz (2006). On the other hand, the ECPC archive may be seen as fundamental material to study (and pedagogically exploit) the (original and translated) genre of parliamentary speeches. This study necessarily means providing (and discussing) qualitative, quantitative and pedagogical data. In effect, this study puts together three fundamental research pieces that may enlighten the "jigsaw puzzle" of the parliamentary genre – the qualitative piece through Critical Discourse Analysis (CDA), the quantitative piece through Corpus-Based Translation Studies (CTS), and the pedagogical piece through Data Driven Learning (DDL).

ECPC and CDA: the qualitative piece

Drawing qualitative information about political / parliamentary genres and subgenres has been done from different standpoints, amongst which Critical Discourse Analysis (CDA) is to be counted. Due to space restrictions, explaining in detail the main premises of this relatively new linguistic approach to texts is beyond the aims of this paper. For clear and precise maps of CDA (or CDS as is known nowadays), see Bloor and Bloor (2007), Caldas-Coulthard and Coulthard (1995), Fairclough (1995), Kress (1990), Wodak (1989), Wodak and Chilton (2005) etc. We could, nevertheless, summarise this method of analysis with a quote by Norman Fairclough (1985: 747), one of its main propounders, according to whom CDA consists in

> The adoption of critical goals [which] means, first and foremost, investigating verbal interactions with an eye to their determination by, and their effect on, social structures.

Adopting "critical goals", as stated by Fairclough above, entails performing a description and an explanation of the text or texts under analysis proceeding from context to texture and vice versa and, hence, ultimately resorting to linguistic tools to delve into structures of power. Quintrileo (2005) and Van Dijk (2002), for instance, propose guidelines to dissect the contextual (macrotextural) components of the parliamentary / political discourse.

Other scholars ─such as Bärenreuter (2005a), Bärenreuter (2005b), De Goede (1996), Kreppel (2000), Mehan (1997), Martín Rojo and Van Dijk (1997), Risse (1999), Van Dijk (2000) Van Dijk (2000), Wodak (2002), Wodak and Weiss (2004), ─ put their emphasis upon the semantics (i.e. central concepts) exchanged within contexts and focus on what they call "**topical networks**" (Wodak and Weiss 2004: 235) or "**nodal points**" (Bärenreuter 2005a: 198) since, as Bärenreuter (2005b: 198) argues:

> political struggles can be understood as the attempt of political actors to promote their respective understandings of these central concepts and make it the dominant one

To give but one revealing example, Kreppel (2000) revolves around "the grand coalition", a particular theory whereby European MEPs are presented as more conciliatory, hence less confrontational, than their counterparts at national parliaments. The author, however, disagrees with the indisputable acceptance of the actual existence of the grand coalition ─and the semantic compromise it generates within the EP─ but does not actually illustrate her disagreement with textural evidence.

There are yet other critical analysts who do precisely this. They concentrate on the linguistic evidence for ideological standpoints. Partington (2003) ─where the author reviews textural features of political discourse─ is a clear example of this microstructure-oriented, bottom-up approach. Other clear instances of this microestructural emphasis are Muntigl (2002), Chilton and Ilyn (1993), Elpass (2002).

Indeed going from context through nodal points to texture (or vice versa) is a particularly fruitful qualitative strategy to research into the genre of (original / translated) parliamentary speech. Calzada Pérez (2007), for example, comes to link together the (pragma-semiotic) realms of the EP, through the semantic nodes of the "grand coalition" to the textural translational shifts identified in (English and Spanish) speeches delivered before the EP on 9th March 1993. In this fashion, the study presents evidence that suggest that, as far as transitivity features are concerned, translated speeches are more conciliatory, less confrontational than their original "equivalents". Nevertheless, since the corpus manually analysed in this study includes one day of parliamentary speeches, its value is indeed more qualitative than quantitative and researchers could do worse than validate it with more ample quantitative data.

ECPC and CTS: the quantitative piece

In order to compile and analyse quantitative data, there is nothing within Translation Studies like corpus-based work. And this is certainly the case when dealing with the genre of parliamentary speeches. Hence, merging qualitative CDA with quantitative corpus-based studies (CTS) may be expected to have significant results. And, in fact, as Garzone and Santulli (2004: 353) have stated:

> Although rarely attempted so far, in the case where it [merging CDA with corpus-based studies] has actually been applied this integration has been impressive

One of the various studies that have brought together CDA and CTS methodologies in order to examine parliamentary speeches is Bayley, Bevitori and Zoni (2004). This piece of work uses electronic corpora to examine particular lexical choices within various parliamentary houses in Europe. The study, for example, gathers data regarding the nodal points of "danger" and "reaction to danger". As a result, it manages to offer a very clear and precise depiction of the different and similar uses of lexical items in English such as "fear", "concern", "threat" or "risk" and their counterparts in Italian (e.g. "minaccia", "pericoloso", "rischio" etc.) and German ("drohen", "bedrohen" etc.). Getting to know what the different parliaments regard as danger and how lexicon is used within a sentence is indeed revealing for critical (political) purposes. In this way, the study reveals that the German Parliament invokes danger and reaction to danger more often than the other two parliaments. Furthermore, in analysing "threat", it comes out that all three parliaments perceive "stability" and "prosperity" in danger. However, while in the House of Commons the source of this threat is the process of European Unification, in the Italian and German Parliaments, the source of danger is precisely the opposite; that is, the potential failure to achieve European integration.

Hence, as is well known by now, CTS manages to provide abundant data to reinforce the link between contextual and textural levels of language. However, what impact does this have upon the translation class?

ECPC and DDL: the pedagogical piece

In recent years a lot of investigation has been devoted to how corpus-based studies can facilitate learning. One of the most popular ways in which this is made possible is through the use of corpora as "pedagogue", or what is currently known as Data Driven Learning ─DDL─, which basically attempts to "develop the ability to see patterning in the target language and to form generalisations to account for that patterning" (Johns 1991: 2-3). In this sense, the primary aim of DDL is to raise learners' awareness by "favour[ing] learning by discovery ─the study of grammar (or vocabulary, or discourse, or style)─, [which] takes on the character of research, rather than spoonfeeding or rote learning" (Tribble and Jones 1990: 12). Learners, thus, become real researchers into language performance and their working methodology is basically based on the rationale that "what the concordancer does is make the invisible visible" (Tribble and Jones 1990:11). In order to do this, students subjected to DDL learning require “a great deal of exposure to language data” (McEnery, Xiao and Tono: 99). Following Johns (1991, quoted in McEnery, Xiao and Tono, 2006), in DDL, trainees go through three stages of learning:

- observation (of corpus-based evidence),
- classification (of salient features), and
- generalization (of rules).

These three stages may interact with the most traditional approach to corpus-based research consisting in the scrutiny of:

- Statistical data
- Wordlists
- Keywords
- Concordances in the form of: 1) collocations, 2) colligations, 3) semantic prosody and 4) semantic preference.

The final sections of this paper, hence, are devoted to exemplifying how students can: 1) observe statistical data; 2) classify keywords; and 3) generalize from concordances (with a special focus on semantic prosody). All these learning tasks show that corpus-based DDL has a great potential to raise students’ awareness not only to translation procedures but also and mainly to CDA critical goals and their impact upon the translation task.

Observing statistical data in the translation classroom

As is well-known, corpus-based work is performed with the aid of concordancers, the most popular of which is, by far, WordSmith Tools (WST, currently under its 4.0 version). Once you upload your data in WST and you place a wordlist query, you immediately obtain overall statistical data from the corpus under scrutiny. This task is easy enough to get students started with corpus-based methodology. Hence, it becomes the inspiration for a set of sub-tasks related to the ECPC Archive to be performed in the translation class.

Task 1: The ECPC Archive contains the ECPC_EN and ECPC_ES corpora. The former comprises all of the speeches in English that were uttered in the EP in 2005; the latter comprises all of the speeches in Spanish that were uttered in the EP in 2005. The speeches contained in these two corpora are XML-tagged, which allows WST to discriminate between original and translated speeches. Analyse the statistical table below:

	Ecpc: original Spanish	Ecpc: Spanish translation from English	Ecpc: Spanish translation from any language	Ecpc: original ENGLISH	Ecpc: English translation from Spanish	Ecpc: English translation from any language
Tokens	223,455	959,443	3,679,473	853,113	194,923	2,723,271
Types	14,935	28,818	50,192	20,901	9,711	29,616
s. TTR	42.25	42.60	42.70	41.32	40.32	41.39
Sd	57.52	57.09	57.25	59.21	59.31	58.69
Word length	3.20	3.19	3.21	2.92	2.93	2.88

Results aimed at:

- The standardised TTR of Spanish originals (42.25) shows that they are more varied, as far as lexis is concerned, than its "equivalent" English translations, whose standardised TTR (40.32) is lower.
- The standardised TTR of Spanish translations (42.60) shows that they are more varied, as far as lexis is concerned, than its "equivalent" English originals, whose standardised TTR (41.32) is lower

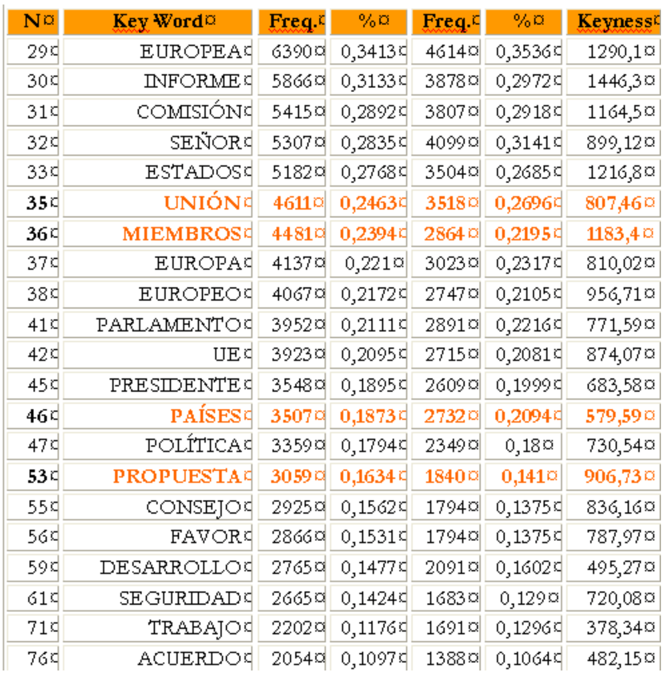
Significance of results:

These results seem to contradict the translational norm of simplification (upheld ─up until now─ by well-known descriptive translation scholars such as Gideon Toury or Mona Baker), which claims that translations are simplified versions of originals; or, in order words, that original texts are more varied than their translated counterparts. These results throw evidence to the contrary and may unleash a revision of the results advocated by the influential descriptive school, with the ideological consequences this may have for translation studies themselves.

Classifying (and comparing) keywords in the translation classroom

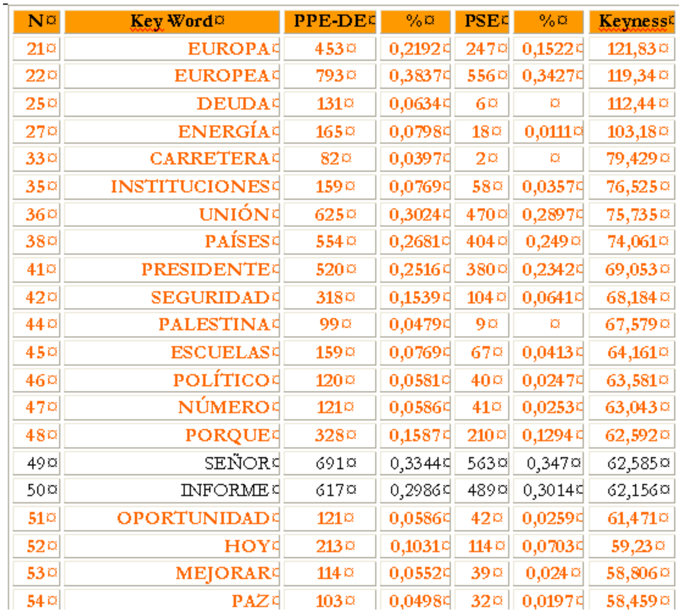
WordSmith Tools 4.0 also allows to produce frequency lists and keyword lists that give a fairly clear picture of the topical networks or nodal points of the corpora analysed. Classifying frequent words according to their degree of significance and comparing them against other frequent words can also be a source of (critical) data for would-be translators.

Task 2: There is an influential theory known as the "grand coalition" (see Kreppel 2000), which claims that ideologies are of "little" importance within the EP. Propounders of the grand coalition argue that regardless of the political parties MEPs represent, they basically speak about the same things and in a pretty similar fashion. Table X below shows keyword lists from Conservative (PPE-DE) and Socialist (PSE) MEPs speaking in Spanish during 2005 (regardless of whether their speeches are original or translated texts). Table Y, for its part, reproduces keywords from the original speeches uttered in Spanish during 2005 by Conservative and Socialist MEPs. Examine both tables and explain the conclusions you may draw from them. Notice that significant differences of lexical use between Conservative and Socialist behaviour are marked in red. These are keywords where there is a difference of use between both groups of 0.02% or above.



N	Key Word	Freq.	%	Freq.	%	Keyness
29	EUROPEA	6390	0,3413	4614	0,3536	1290,1
30	INFORME	5866	0,3133	3878	0,2972	1446,3
31	COMISIÓN	5415	0,2892	3807	0,2918	1164,5
32	SEÑOR	5307	0,2835	4099	0,3141	899,12
33	ESTADOS	5182	0,2768	3504	0,2685	1216,8
35	UNIÓN	4611	0,2463	3518	0,2696	807,46
36	MIEMBROS	4481	0,2394	2864	0,2195	1183,4
37	EUROPA	4137	0,221	3023	0,2317	810,02
38	EUROPEO	4067	0,2172	2747	0,2105	956,71
41	PARLAMENTO	3952	0,2111	2891	0,2216	771,59
42	UE	3923	0,2095	2715	0,2081	874,07
45	PRESIDENTE	3548	0,1895	2609	0,1999	683,58
46	PAÍSES	3507	0,1873	2732	0,2094	579,59
47	POLÍTICA	3359	0,1794	2349	0,18	730,54
53	PROPUESTA	3059	0,1634	1840	0,141	906,73
55	CONSEJO	2925	0,1562	1794	0,1375	836,16
56	FAVOR	2866	0,1531	1794	0,1375	787,97
59	DESARROLLO	2765	0,1477	2091	0,1602	495,27
61	SEGURIDAD	2665	0,1424	1683	0,129	720,08
71	TRABAJO	2202	0,1176	1691	0,1296	378,34
76	ACUERDO	2054	0,1097	1388	0,1064	482,15

Table X



N	Key Word	PPE-DE	%	PSE	%	Keyness
21	EUROPA	453	0,2192	247	0,1522	121,83
22	EUROPEA	793	0,3837	556	0,3427	119,34
25	DEUDA	131	0,0634	6		112,44
27	ENERGÍA	165	0,0798	18	0,0111	103,18
33	CARRETERA	82	0,0397	2		79,429
35	INSTITUCIONES	159	0,0769	58	0,0357	76,525
36	UNIÓN	625	0,3024	470	0,2897	75,735
38	PAÍSES	554	0,2681	404	0,249	74,061
41	PRESIDENTE	520	0,2516	380	0,2342	69,053
42	SEGURIDAD	318	0,1539	104	0,0641	68,184
44	PALESTINA	99	0,0479	9		67,579
45	ESCUELAS	159	0,0769	67	0,0413	64,161
46	POLÍTICO	120	0,0581	40	0,0247	63,581
47	NÚMERO	121	0,0586	41	0,0253	63,043
48	PORQUE	328	0,1587	210	0,1294	62,592
49	SEÑOR	691	0,3344	563	0,347	62,585
50	INFORME	617	0,2986	489	0,3014	62,156
51	OPORTUNIDAD	121	0,0586	42	0,0259	61,471
52	HOY	213	0,1031	114	0,0703	59,23
53	MEJORAR	114	0,0552	39	0,024	58,806
54	PAZ	103	0,0498	32	0,0197	58,459

Table Y

Results aimed at:

From a statistical viewpoint, table X shows that, when there is no distinction between original and translated speeches, there are virtually no differences between Conservative and Socialist nodal points.

From a statistical viewpoint, table Y shows that, when researchers focus on original speeches, differences between Conservative and Socialist nodal points largely increase.

Significance of results:

The grand coalition seems to be applicable when no difference is established between original and translated speeches. However, when original speeches only are analysed, results seem to put into question the grand coalition. This means that translational behaviour may have something to do with the converging / homogenising criteria applied within the EP. More research is needed in this respect. But these results underline the importance of translation to gather discoursal communities together or to set them apart.

Generalizing from concordances in the translation classroom

Indeed one of the most widely-used facilities offered by WST 4.0 is that of generating concordances of desired words or lexical items. By looking at concordances and by generalizing "partial theories" from their behaviour, a lot of information may be gathered since, as Tribble and Jones (1990:11) remark, "what the concordancer does is make the invisible visible". Concordance lines offer information about lexical collocations, linguistic colligations, semantic preference and semantic prosody. The tasks designed under this heading are related to semantic prosody, which has been defined by Hunston and Thompson (2000: 5) as "the speaker's or writer's attitude or stance towards, viewpoint or feeling about entities and propositions that he or she is talking about". Semantic prosody often unveils speakers' ideological tendencies.

Task 3: Look up the word "Perpetuate" in your monolingual and bilingual dictionaries.

Results aimed at:

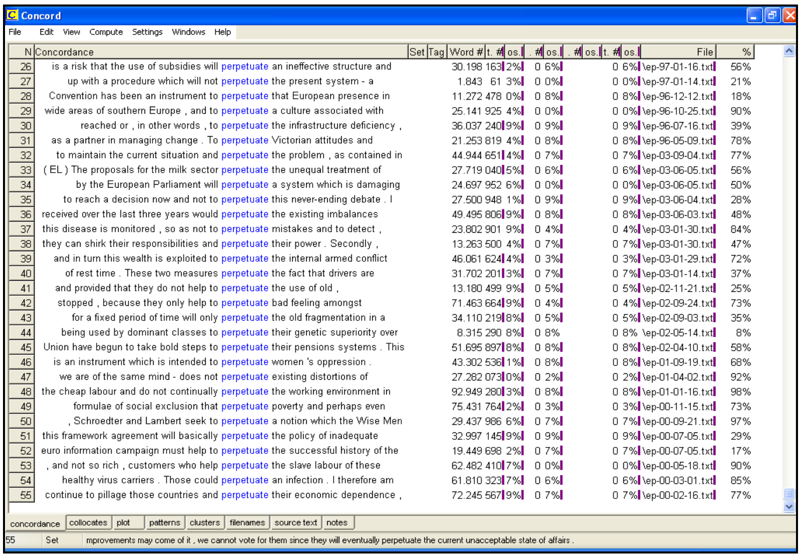
Cambridge Advanced Learner's Dictionary: continuing forever in the same way // frequently repeated.

Oxford Bilingual Dictionary (English-Spanish): Perpetuar.

Task 4: Now query ECPC_EN (Original) and produce a table of concordances for the nodal verb "perpetuate":

Results aimed at:

Concordances for "perpetuate":

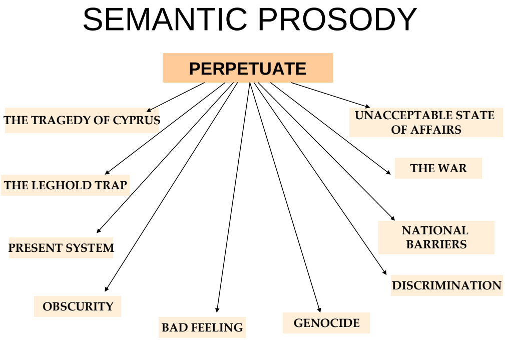


N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
26	is a risk that the use of subsidies will perpetuate an ineffective structure and			30.198	163	2%	0	6%			0	6%	\ep-97-01-16.txt	56%
27	up with a procedure which will not perpetuate the present system - a			1.843	61	3%	0	0%			0	0%	\ep-97-01-14.txt	21%
28	Convention has been an instrument to perpetuate that European presence in			11.272	478	0%	0	8%			0	8%	\ep-96-12-12.txt	18%
29	wide areas of southern Europe , and to perpetuate a culture associated with			25.141	925	4%	0	0%			0	0%	\ep-96-10-25.txt	90%
30	reached or , in other words , to perpetuate the infrastructure deficiency ,			36.037	240	9%	0	9%			0	9%	\ep-96-07-16.txt	39%
31	as a partner in managing change . To perpetuate Victorian attitudes and			21.253	819	4%	0	8%			0	8%	\ep-96-05-09.txt	78%
32	to maintain the current situation and perpetuate the problem , as contained in			44.944	651	4%	0	7%			0	7%	\ep-03-09-04.txt	77%
33	(EL) The proposals for the milk sector perpetuate the unequal treatment of			27.719	040	5%	0	6%			0	6%	\ep-03-06-05.txt	56%
34	by the European Parliament will perpetuate a system which is damaging			24.697	952	6%	0	0%			0	0%	\ep-03-06-05.txt	50%
35	to reach a decision now and not to perpetuate this never-ending debate . I			27.500	948	1%	0	9%			0	9%	\ep-03-06-04.txt	28%
36	received over the last three years would perpetuate the existing imbalances			49.495	806	9%	0	8%			0	8%	\ep-03-06-03.txt	48%
37	this disease is monitored , so as not to perpetuate mistakes and to detect ,			23.802	901	9%	0	4%			0	4%	\ep-03-01-30.txt	84%
38	they can shirk their responsibilities and perpetuate their power . Secondly ,			13.263	500	4%	0	7%			0	7%	\ep-03-01-30.txt	47%
39	and in turn this wealth is exploited to perpetuate the internal armed conflict			46.061	624	4%	0	3%			0	3%	\ep-03-01-29.txt	72%
40	of rest time . These two measures perpetuate the fact that drivers are			31.702	201	3%	0	7%			0	7%	\ep-03-01-14.txt	37%
41	and provided that they do not help to perpetuate the use of old ,			13.180	499	9%	0	5%			0	5%	\ep-02-11-21.txt	25%
42	stopped , because they only help to perpetuate bad feeling amongst			71.463	664	9%	0	4%			0	4%	\ep-02-09-24.txt	73%
43	for a fixed period of time will only perpetuate the old fragmentation in a			34.110	219	8%	0	5%			0	5%	\ep-02-09-03.txt	35%
44	being used by dominant classes to perpetuate their genetic superiority over			8.315	290	8%	0	8%			0	8%	\ep-02-05-14.txt	8%
45	Union have begun to take bold steps to perpetuate their pensions systems . This			51.695	897	8%	0	8%			0	8%	\ep-02-04-10.txt	58%
46	is an instrument which is intended to perpetuate women 's oppression .			43.302	536	1%	0	8%			0	8%	\ep-01-09-19.txt	68%
47	we are of the same mind - does not perpetuate existing distortions of			27.282	073	0%	0	2%			0	2%	\ep-01-04-02.txt	92%
48	the cheap labour and do not continually perpetuate the working environment in			92.949	280	3%	0	8%			0	8%	\ep-01-01-16.txt	98%
49	formulae of social exclusion that perpetuate poverty and perhaps even			75.431	764	2%	0	3%			0	3%	\ep-00-11-15.txt	73%
50	, Schroedter and Lambert seek to perpetuate a notion which the Wise Men			29.437	986	6%	0	7%			0	7%	\ep-00-09-21.txt	97%
51	this framework agreement will basically perpetuate the policy of inadequate			32.997	145	9%	0	9%			0	9%	\ep-00-07-05.txt	29%
52	euro information campaign must help to perpetuate the successful history of the			19.449	698	2%	0	7%			0	7%	\ep-00-07-05.txt	17%
53	, and not so rich , customers who help perpetuate the slave labour of these			62.482	410	7%	0	0%			0	0%	\ep-00-05-18.txt	90%
54	healthy virus carriers . Those could perpetuate an infection . I therefore am			61.810	323	7%	0	6%			0	6%	\ep-00-03-01.txt	85%
55	continue to pillage those countries and perpetuate their economic dependence ,			72.245	567	9%	0	7%			0	7%	\ep-00-02-16.txt	77%

55 Set mprovements may come of it , we cannot vote for them since they will eventually perpetuate the current unacceptable state of affairs .

Task 5: Classify examples and try to a create cluster chart that portrays the semantic prosody of "Perpetuate" in our corpus:

Results aimed at:


SEMANTIC PROSODY

PERPETUATE

THE TRAGEDY OF CYPRUS

UNACCEPTABLE STATE OF AFFAIRS

THE LEGHOLD TRAP

THE WAR

PRESENT SYSTEM

NATIONAL BARRIERS

DISCRIMINATION

OBSCURITY

BAD FEELING

GENOCIDE

Significance of results:

Two are the main conclusions students may draw after generalizing from the concordances for the word "perpetuate":

1) While in Spanish perpetuate has a neutral (neither positive nor negative) prosody (see concordances for the word at www.rae.es), the way the word is used in our corpus acquires a predominantly negative nuance. This is key translational information.

2) Students may develop an interest for the kind of words which are associated with the verb "perpetuate" among speakers from (for example) certain political groups. By doing so, they may be able to establish the political friends and foes of the various parties and MEPs speaking at the EP during 2005. This may have a decisive influence when translating speeches from English into Spanish since would-be translators may try to compensate for the potential loss of semantic prosody by scattering relevant nuances throughout the text.

Conclusion

Although the ECPC archive so far is largely limited to the linguistic behaviour of MEPs (and MPs at other national parliaments) for the year 2005, the data discussed here shows that the archive has potential for the critical analysis of the parliamentary genre and for its exploitation within the translation class. So far, of course, the results presented here are only tentative but once the Archive is enriched with further data, it will be a valuable source for CDA qualitative, CTS quantitative and DDL studies.

References

Bärenreuter, C. 2005. "'It is not sufficient to have a moral basis, it has to be democratic too'. Constructing 'Europe' in Swedish media reports on the Austrian political situation in 2000". In *A New Agenda in (Critical) Discourse Ansalysis. Theory Methodology and Interdisciplinarity*, Wodak, R. and P. Chilton (ed.). Ámsterdam / Filadelfia: John Benjamins.

Bärenreuter, C. 2005. " Researching the European public sphere and its political functions. A proposal". *Freedom, Justice, and Identity* Vol. 18/Online.

Bayley, P., Bevitori, C. and Zoni, E. 2004. "Threat and fear in parliamentary debates in Britain, Germany and Italy". In *Cross-Cultural Perspectives on Pàrliamentary Discourse*, Bayley, P. (ed.). Ámsterdam / Filadelfia: John Benjamins.

Bloor, M. and Bloor, T. 2007. "The practice of critical discourse analysis. An introduction".

Caldas-Coulthard, C. and Coulthard, **M.** 1995. "Texts and practices: Readings in critical discourse analysis".

Calzada Pérez, M. 2007. *Transitivity in Translating: The Interdependence of Texture and Context. Preface by Ian Mason. Bern: Peter Lang, 2007*. Berna: Peter Lang.

Calzada Pérez, M. and Luz, S. 2006. "ECPC - Technology as a Tool to Study the

(Linguistic) Functioning of National and Trans-National European Parliaments". *International Journal of Technology Knowledge and Society 2/5* http:// www.Technology-Journal.com. html [Access date 10/05/2008].

Carbó, T. 2004. "Parliamentary discourse when things go wrong: Mapping histories, contexts and conflicts". In *Cross-Cultural Perspectives on Parliamentary Discourse*, Bayley, P. (ed.). Ámsterdam / Filadelfia: John Benjamins.

Chilton, P. and Ilyn, M. 1993. "Metaphor in political discourse: the case of the 'common European house'". *Discourse & Society* 4/1: 7-31.

De Goede, M. 1996. "Ideology in the US welfare debate: neo-liberal representations of poverty". *Discourse & Society* 7/3: 317-357.

Elpass, S. 2002. "Phraseological units in parliamentary discourse". In *Politics as Text and Talk*, Chilton, P. and C. Schäffner (ed.). Ámsterdam / Filadelfia: John Benjamins.

Fairclough, N. 1985. "Critical and Descriptive Goals in Discourse Analysis". *Journal of Pragmatics* 9/739-763.

Fairclough, N. 1995. *Critical discourse analysis. The critical study of language.* London: Longman.

Garzone, G. and Santulli, F. 2004. "What can corpus linguistics do for Critical Discourse Analysis?" In *Corpora and Discourse*, Partington, A., J. Morley and L. Haarman (ed.). Bern: Peter Lang.

Hunston, S. and Thompson, G. 2000. *Evaluation in Text: Authorial stance and the construction of Discourse.* Oxford: Oxford University Press.

Johns, T. 1991. "Should you be persuaded: Two samples of data-driven learning material ". *Classroom Concordancing. Special issue of ELR Journal* 4/1-13.

Kreppel, A. 2000. "Rules, ideology and coalition formation in the European Parliament. Past, present and future". *European Union Politics* 1/3: 340-362.

Kress, G. 1990. "Critical Discourse Analysis". *Annual Review of Applied Linguistics* II/

Martín Rojo, L. and T. A. Van Dijk. 1997. "'There was a problem, and it was solved!': legitimating the expulsion of 'illegal' migrants in Spanish parliamentary discourse". *Discourse & Society* 8/4: 523-566.

McEnery, A. M., Xiao, R. and Tono, Y. 2006. *Corpus-Based Language Studies. An Advanced Resource Book*. London and New York: Routledge.

Mehan, H. 1997. "The discourse of the illegal immigration debate: a case study in the politics of representation". *Discourse & Society* 8/2: 249-270.

Muntigl, P. 2002. "Politicization and depoliticization. Employment policy in the European Union". In *Politics as Text and Talk*, Chilton, P. and C. Schäffner (ed.). Ámsterdam / Filadelfia: John benjamins.

Partington, A. 2003. *The Linguistics of Political Argument. The Spin-Doctor and the Wolf-Pack at the White House*. Londres

Nueva York: Routledge.

Quintrileo, C. 2005. "El debate parlamentario como género discursivo. Una primera aproximación". *VI Congreso Latinoamericano de Estudios del Discursos. América Latina en su Discurso* 1-14.

Risse, T. 1999. "To euro or not to euro?: The EMU and identity politics in the European Union". *European Journal of International Relations* 15/2: 147-187.

Tribble, C. and Jones, G. 1990. *Concordances in the Classroom*. London: Longman.

Van Dijk, T. A. 2000. "On the analysis of parliamentary debates on immigration". In *The Semiotics of Racism.. Approaches to Critical Discourse Analsysis*, Reisigl, M. and R. Wodak (ed.). Viena: Passagen Verlan.

Van Dijk, T. A. 2002. "Political discourse and political cognition". In *Politics as Text and Talk*, Chilton, P. and C. Schäffner (ed.). Ámsterdam /Filadelfia: John Benjamins.

Wodak, R. 2002. "Fragmented identities. Redefining and recontextualizing national identities". In *Politics as Text and Talk. Analytic Approaches to Political Discourse*, Chilton, P. and C. Schäffner (ed.). Ámsterdam / Filadelfia: John Benjamins.

Wodak, R. 1989. "Language Power and Ideology: Studies in Political Discourse".

Wodak, R. and P. Chilton. 2005. "A New Agenda in (Critical) Discourse Analysis. Theory, Methodology and Interdisciplinarity".

Wodak, R. and Weiss, G. 2004. "Visions, ideologies, and utopias in the discursive construction of European identities: Organizing, representing and legitimizing Europe". In *Communicating Ideologies: Multidisciplinary Perspectives on Language, Discourse, and Social Practice*, Pütz, M., J. N.-v. Aesterkaer and T. A. Van Dijk (ed.). Berna: Peter Lang.

USING A CORPUS TO TEACH RHETORICAL FUNCTIONS: STUDENTS' EVALUATION OF A HANDS-ON CONCORDANCING APPROACH

Maggie Charles[14]

Abstract

Much corpus work on academic writing examines lexico-grammatical patterning at a local level, but has had little impact on syllabuses or materials because it fails to address pedagogical issues adequately. One problem is that the use of corpus consultation to explore a series of lexico-grammatical patterns does not in itself constitute a coherent set of teaching materials. This paper suggests that if corpus work is to be incorporated routinely into academic writing classes, it must also help students to tackle higher level discourse concerns. The study proposes a pedagogic approach which combines discourse analysis with corpus investigation and reports on students' evaluation of the materials derived. The six-week, twelve-hour course is primarily designed to develop students' recognition and understanding of key rhetorical functions and an example of the material on 'Criticising the Work of Other Researchers' is discussed. Each unit begins with discourse-based tasks to raise students' awareness of a given function and continues with hands-on concordancing to focus on specific lexico-grammatical options for performing that function. The corpora consist of native-speaker theses (approximately 500,000 words) and were examined using WordSmith Tools (Scott 2005). The participants were forty-nine international graduates, who evaluated sixteen statements about corpus work on a five-point scale from 'strongly disagree' to 'strongly agree'. The results show that attitudes towards corpus work were generally very favourable: percentages of those agreeing with positive statements about the tasks ranged from 61% to 96%. This study argues that corpus work provides access to extended context, which facilitates the study of discourse and helps students bridge the gap between their rhetorical concerns and lexico-grammatical knowledge. It concludes by stressing that large mixed-discipline academic writing classes need an approach to corpus work that is both systematic and pedagogically valid.

Keywords: hands-on concordancing, corpus materials, academic writing, student attitudes, corpus pedagogy

Introduction

Many corpus studies have investigated aspects of written academic discourse, but concern has been expressed that much of this work concentrates on examining lexico-grammatical patterning at a local level, with little attention given to macro-textual features (Swales 2002). Following on from this, it has also been argued that such research has had little impact on syllabuses or materials because it fails to address pedagogical issues adequately (Flowerdew 1998, 2002). One of the problems is that the use of a corpus to investigate a series of individual lexico-grammatical patterns does not constitute a systematic and pedagogically valid set of course materials. While it is certainly true that advanced EAP writing students need lexico-grammatical 'fine-tuning' (Lee & Swales 2006), they also have higher level discourse concerns. This paper argues that if corpus work is to be incorporated routinely into the class teaching of academic writing, it must also help students to tackle such issues. The aim of the study is to propose an approach which combines discourse analysis with corpus investigation and to report on students' evaluation of the materials derived.

Background to the study

The Course

The course described here was offered as part of Oxford University Language Centre's programme of open-access academic writing classes for graduates. It expands and takes forward the pilot version presented in Charles (2007).

[14] Maggie Charles is a tutor in EAP at Oxford University Language Centre, where she teaches academic writing to graduates. Her research interests include the pedagogical applications of corpus linguistics, the study of stance and discipline-specific discourse. She has published in Applied Linguistics, Journal of English for Academic Purposes and English for Specific Purposes and is currently co-editing a volume entitled Academic Writing: At the Interface of Corpus and Discourse for Continuum.

Five parallel groups, each with around 16 participants, attended one weekly two-hour session on ‘Investigating Rhetorical Functions’ for six weeks. This course differs from many of those reported elsewhere in that its primary purpose was to develop students’ recognition and understanding of key rhetorical functions. Corpus work was introduced as a means of achieving this goal, rather than constituting the main focus of the course itself. The materials begin with discourse-based tasks to raise students’ awareness of a given function and continue with hands-on concordancing, which uses the corpus to focus on specific lexico-grammatical options for performing that function. The rhetorical functions studied are: Situating your research in the field; Defending your research against criticism; Portraying your professional competence; Making and modifying claims; Criticising others’ research; Making and countering arguments. The corpora consist of theses written by native-speakers: approximately 190,000 words in politics and 300,000 words in materials science and were examined using the concordancer of WordSmith Tools (Scott 2005).

The Participants

The participants were international graduates or researchers. Forty-nine students completed both the initial questionnaire, adapted from Yoon and Hirvela (2004), which provides information on backgrounds and attitudes towards computer use and the final evaluation, which asks participants to rate sixteen statements about the corpus work on a five-point scale from *strongly disagree* to *strongly agree*. Roughly 55% of the participants were doing doctorates, 41% were Master’s students and two were post-doctoral researchers. They worked in 33 different research fields: 33% in natural sciences, 45% in social sciences and 22% in humanities. Participants spoke 21 native languages. All but one liked using computers and most did so several times per day. 47% of them had used a corpus before, which reflects the fact that the British National Corpus had been introduced in earlier courses. Attendance was not necessarily regular, but 47% of the students attended all six corpus sessions.

It is clear that these classes were very heterogeneous, particularly in discipline, native language and experience in corpus use and indeed such diversity of background may be characteristic of academic writing classes more generally. By contrast, published studies on corpus-based courses tend to concern groups which were relatively small (Lee & Swales 2006; Yoon & Hirvela 2004) and/or homogeneous (Bondi 2001; Gavioli 2005; Hafner & Candlin 2007; Weber 2001). Thus the Oxford classes provide a useful test case for two research questions. 1. Can concordancing help students learn features of discourse? 2. What type of corpus work can be used to benefit relatively large mixed-discipline classes?

An example of classroom procedure: Criticising the work of other researchers

One difficulty often mentioned by students is that of making an acceptable criticism of other researchers’ work. They are frequently urged to ‘be critical’, but given little specific guidance on how to achieve this. It seemed likely, therefore, that a combination of discourse and corpus work would enable students to discover some of the rhetorical and lexico-grammatical patterns associated with this function.

The discourse session begins by focusing on two extracts from successful theses (given below as Extracts A and B), both of which criticise others’ research. Students are asked to underline the linguistic features that construct the criticism and, in discussion with a partner, to compare and contrast the extracts, commenting on their acceptability and persuasive effect. The first is extremely forthright in its criticism, making an overt attack (*gross oversimplification, disregards all the known data*), rather than a reasoned argument, while the second is more nuanced and includes a concession to the other researcher’s view (*It is true that…*), before giving the writer’s criticism, signalled by *but*. The purpose of the task is not just to illustrate how criticisms can be made, but to help students reflect on their own attitudes to criticising others’ work and the norms of their discipline in this regard.

Extract A
Statements such as a bone collagen δ^{13}C value of -18.07‰ implies a dietary content of 16% C_4 plants (White & Schwarcz 1994) have been made *and accepted* as possible conclusions. The assumptions on which this statement is based is that all C_3 and C_4 plants can be assigned fixed single δ^{13}C values... This is a **gross oversimplification and disregards all the known data** on the wide variation in plant isotopic values…

Extract B
There exists an important strand of thought concerning state socialization that limits the scope of investigation to processes affecting individuals *within* states (see Luard: 59-60; Ikenberry and Kupchan: 289-290). **But** states are **not merely** aggregates of individuals. States are corporate actors, possessed of centralized administrative organs. **It is true that** state policy is made by individuals, **but** these individuals are subject to pressure from domestic interests…

In the concordancing part of the class, the students focus on two-part rhetorical patterns similar to that identified in Extract B, in which the writer first acknowledges the work of another researcher before criticising it in some way. They are asked to perform corpus searches on *but, however* and *though* with _(19*_ in the context. As the corpora were compiled in the 1990s, this retrieves citations which occur within the context of a contrasting statement signalled by the conjunction. In each concordance, this provides students with several examples of the rhetorical pattern studied. Thus in the materials science corpus, the concordance for *but* in the context of _(19*_ retrieves 19 lines, of which about half show evidence of this pattern, for example:

> 1 This type of difficulty in imaging n+p-junctions has also been reported by Venables and Maher (1996). Their work successfully imaged ion-implanted p+n-junctions, **but** did not report the observation of any contrast in similar n+p-junctions.
> 2 Humphreys (1993) and Gater (1994) studied the grain boundaries of MA6000 **but** failed find any evidence of carbides.
> 3 The early parameterisation of Chadi(1979c) described the band structure rather well, and had the correct separation of Es and Ep, **but** gave poor elastic constants,
> 4 A similar insensitivity of SE-imaging was reported, **but** not explained, by Venables and Maher (1996)
> 5 Norenberg, Bowler and Briggs (1997) also found an increase in the amount of c(4x4) with Ga exposure, **but** were unable to measure the Ga coverage in Auger.

Part of concordance on *but* in the context of _(19*_ from the materials science corpus

Students are asked first to identify which lines construct criticisms and how they are signalled (by *but,* often with negatives e.g. *unable, not explained* or with negative evaluation e.g. *poor*). They are also asked to note any features which soften the effect of the criticisms (e.g. factive verb *found*, positive evaluation e.g. *rather well, successfully*). The aim of the task is to draw students' attention to a way of mitigating the effect of criticism by first indicating what the other researcher **has** achieved before drawing attention to limitations or flaws. In pairs students discuss their findings and the class ends with a report back session for the whole group. Paper versions of the concordances are handed out at the end of the session so that students have a record of the data they worked with. Homework is to write a short paragraph criticising the research of others in their own field.

Results of the students' evaluation

The results of the final questionnaire show that attitudes towards the corpus work were generally very favourable. Thus the average percentage of students who *somewhat agree* or *strongly agree* with positive statements about corpus work was 82%. In providing the following brief overview, I combine the categories '*somewhat*' and '*strongly*' and refer here to 'agree' or 'disagree'. Thus 90% or over agreed that concordances helped them learn about rhetorical functions and how they are expressed and would recommend other international students to use a corpus for help with their English. Between 80% and 90% agreed that working with the concordances was a useful supplement to working with the texts and intend to use a corpus in the future; it also helped them to understand how rhetorical functions are used, to learn the grammar and vocabulary associated with them and to become more aware of them in their reading. Between 70% and 80% agreed that working with concordances was interesting, that it helped them to use rhetorical functions in their own writing and that they would like to use corpora in a future English class. Lower percentages were recorded for agreeing that working with concordances helped students learn other aspects of academic writing (63%) and for disagreeing with the negative statement that analysing the concordance lines was difficult because of the language (63%). However, the results were generally very positive, showing that working with concordances is both possible and fruitful for teaching discourse features to mixed-discipline classes.

I now discuss in more detail the students' evaluation of two statements which concerned the practicalities of corpus work. These raise wider issues relevant to incorporating concordancing into an academic writing course, particularly with regard to the two research questions posed earlier. The full results appear in Table 1.

Statement	strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
Analysing the concordance lines took too much time because there was a lot of data.	20%	22%	22%	24%	10%
It was easy to perform the searches using the WordSmith Concordance program.	0%	18%	20%	35%	27%

Table 1 Student evaluation of two statements concerning the practicalities of corpus work

Discussion

'Analysing the concordance lines took too much time because there was a lot of data.'

The first issue concerns the amount of data consulted, which caused difficulty for some students: 35% of participants agreed with the statement above, while only 43% disagreed. This is a problem that has been noted before (e.g. Chambers 2005; Kennedy & Miceli 2001), although usually because of the number of individual concordance lines. Here, the numbers of lines were limited (typically around 10-20), but a greater amount of context was examined. Nonetheless, students tended to read much more context than necessary and had to be encouraged to grow or shrink lines depending on the information needed. This issue can be addressed through demonstrating how different amounts of context reveal different aspects of the discourse feature studied. It is important that students examine enough context to ensure that discourse patterns can be observed, without being overwhelmed by the quantity of data.

Nonetheless, I suggest that concordances do have an important role to play in investigating context. In particular I would argue that, far from being instances of 'decontextualised' language (Widdowson 2000), concordance lines enable students to understand the importance of context in certain respects better than paper-based materials. This is because academic texts are so long that students in academic writing classes do not normally work with whole texts, but rather with extracts, which are limited in the amount of context they provide. Moreover, students often read such extracts as if they **were** complete: they come to conclusions and make judgments without taking into account the fact that they only have access to part of the writer's text. This tendency is actually fostered by the appearance of extracts on the written page: they begin at the beginning and end at the end of sentences and are surrounded by white space. By contrast, when a student reads a concordance line, however much context is displayed, it is always clear that they are looking at only a fragment of text. This acts as a stimulus for students to expand lines and/or access the original file, two ways of examining context which are simply unavailable on paper. For these reasons, then, concordances are particularly valuable not just for studying relatively short lexico-grammatical patterns, but also for working with extended contexts and investigating the discourse level of features such as rhetorical functions.

'It was easy to perform the searches using the WordSmith Concordance program.'

Only 62% agreed with the statement above, while 18% 'somewhat' disagreed. Although this may not seem a particularly negative result, the disagreement percentage was one of the highest and the agreement level one of the lowest in the evaluation. I would suggest that there are two main reasons for this. First, the necessity of ensuring a limited number of concordance lines leads to more complicated searches, involving, for example, context items or wild cards. For a beginner in corpus work, this undoubtedly presents a challenge. It is more likely, for example, that typing errors will occur, which may cause no concordance entries to be found and result in frustration. Secondly, if the main focus of the class is on learning the rhetorical function, it is more difficult to present the corpus tasks in a systematic and graded sequence, with the result that students may be overloaded with new technical information at the beginning of the course and there may be little sense of progression in concordancing skills. Undoubtedly, these disadvantages result from the decision to subordinate the learning of corpus techniques to the study of discourse functions. While some further adjustments in the difficulty and sequencing of the corpus tasks can certainly be made, I would argue that the slightly more negative responses seen here are at an acceptable level in relation to the overall aims of the course.

However, consideration of this aspect of the students' evaluation raises the wider issue of the place of corpus work in EAP writing courses and in particular what sort of approach is feasible and useful with relatively large mainstream classes. Of the student groups described in the literature, the two which most resemble the Oxford classes are the one discussed by Lee and Swales (2006) and the advanced class in the Yoon and Hirvela study (2004). Both of these studies view corpus work as a response to individual writing problems, but they propose very different roles for corpora in the writing class. After initial training, the Yoon and Hirvela group was left to consult the corpus outside the class, while Lee and Swales made corpus work the focus of their course, which culminated in individual corpus-based student projects. Responses from the Yoon and Hirvela study indicated a less favourable attitude to corpus work among the advanced group, who had worked on corpora independently than among the intermediate group, who used corpora in class. The authors suggest that the advanced students' less positive response may have been due to the lack of teacher support and class-based practice. By contrast, the Lee and Swales group became skilled and enthusiastic corpus users, but, as the authors note, the success of this approach may depend on working with limited numbers of highly-motivated students.

Thus if corpus work is not to be relegated to the fringes of EAP writing classes, there is still a need to re-think its purpose and role. The results of the above studies show that even advanced students need considerable in-class support. However this is difficult to achieve in larger groups if each student is working on his/her own individual problems. Hence I propose a different role for corpus work in class. In this approach, corpus consultation is used not as a support for individual writing tasks, but as a way of studying rhetoric. Lexico-grammatical patterns are still examined and fine-tuning occurs, but in a more systematic way, since the patterns are investigated as realisations of the discourse function studied. Thus, corpus work is a key element of class activity, in which each individual task contributes to the overall aim of increasing rhetorical awareness and competence.

Conclusions

In this paper I have suggested that advanced students of academic writing need to learn not only local lexico-grammatical patterns, but also higher level discourse features and I have argued that corpus work provides access to extended context, which facilitates the study of discourse. I have also suggested that teaching large mixed-discipline EAP writing classes raises distinct issues with regard to the type of corpus work that is appropriate. The approach I propose brings corpus consultation into mainstream EAP classes for the purpose of studying rhetoric. To illustrate this, I have described a course which integrates hands-on concordancing very closely with work on rhetorical functions. As shown by the participants' positive evaluation, it is possible to learn discourse features through concordance work and students can benefit from engaging with corpora to bridge the gap between their rhetorical concerns and their lexico-grammatical knowledge. The use of corpus consultation as a support for discourse work can thus provide a systematic approach that is both feasible and pedagogically valid.

References

Bondi, M. 2001. "Small corpora and language variation." In *Small Corpus Studies and ELT*, M. Ghadessy, A. Henry and R. L. Roseberry (eds.). Amsterdam: Benjamins, 135-174.

Chambers, A. 2005. "Integrating corpus consultation in language studies." *Language Learning and Technology* 9/2: 111-125.

Charles, M. 2007. "Reconciling top-down and bottom-up approaches to graduate writing: using a corpus to teach rhetorical functions." *Journal of English for Academic Purposes* 6/4: 289-302.

Flowerdew, L. 1998. "Corpus linguistic techniques applied to textlinguistics." *System* 26/4: 541-552.

Flowerdew, L. 2002. "Corpus-based analyses in EAP." In *Academic Discourse*, J. Flowerdew (ed.). London: Longman, 95-114.

Gavioli, L. 2005. *Exploring Corpora for ESP Learning*. Amsterdam: Benjamins.

Hafner, C. A., and **Candlin, C. N.** 2007. "Corpus tools as an affordance to learning in professional legal education." *Journal of English for Academic Purposes* 6/4: 303-318.

Kennedy, C., and **Miceli, T.** 2001. "An evaluation of intermediate students' approaches to corpus investigation." *Language Learning and Technology* 5/3: 77-90.

Lee, D., and **Swales, J.** 2006. "A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora." *English for Specific Purposes* 25/1: 56-75.

Scott, M. 2005. *WordSmith Tools* (Version 4). Oxford: Oxford University Press.

Swales, J. M. 2002. "Integrated and fragmented worlds: EAP materials and corpus linguistics." In *Academic Discourse*, J. Flowerdew (ed.). London: Longman, 150-164.

Weber, J.-J. 2001. "A concordance-and genre-informed approach to ESP essay writing." *ELT Journal* 55/1: 14-20.

Widdowson, H. G. 2000. "On the limitations of linguistics applied." *Applied Linguistics* 21/1: 3-25.

Yoon, H., and **Hirvela, A.** 2004. "ESL student attitudes towards corpus use in L2 writing." *Journal of Second Language Writing* 13/4: 257-283.

3A DDL APPROACH TO LEARNING NOUN AND VERB PHRASES IN THE BEGINNER LEVEL EFL CLASSROOM

Kiyomi Chujo[15]

Kathryn Oghigian[16]

Abstract

In this study, we combine learning grammar through "observation, hypothesis formation, use" in a series of hands-on DDL (data-driven learning) activities with a CALL (computer-assisted language learning) vocabulary learning program for teaching TOEIC-specific grammar and vocabulary to beginner level Japanese university students. First, students follow carefully crafted guidelines to explore the structures of noun or verb phrases using parallel Japanese-English concordancing lines from a specially developed Japanese-English newspaper corpus (Utiyama & Isahara 2003) and a bilingual concordancer (Barlow 2004). Next, the vocabulary is consolidated with a 30-minute CALL program activity. Finally, they complete follow-up activities using targeted vocabulary to reinforce the grammar patterns discovered through the DDL activities. The evaluation of learning outcomes showed this course design was useful for learning grammar basics such as recognizing parts of speech and derivations, and understanding the basic structures of noun phrases and verb phrases, in addition to the more complex grammar questions such as those found on TOEIC. In response to questionnaires, students reported that they enjoyed this course and believed it was useful for both grammar and vocabulary learning. As a modification of a previous study in an on-going project, this study included new follow-up exercises which students also found useful.

Keywords: DDL, Japanese-English parallel corpus, beginner level EFL, CALL, grammar

Introduction

When Japanese high school graduates enter university, a dismaying number cannot conjugate verbs or recognize parts of speech (Chujo, Nishigaki & Uchibori 2006); they have difficulty identifying and understanding the structure of noun and verb phrases (Uchibori & Chujo 2005) and their overall understanding and knowledge of English grammar has decreased drastically in the last few years (Green 2006). At the same time, these students are faced with taking proficiency tests such as the Test of English for International Communication (TOEIC) and the Test of English as a Foreign Language (TOEFL), which are frequently used as both placement and achievement indicators for university, and as hiring and promotion indicators for companies. In a study comparing the grammar taught in the most widely used Japanese high school textbooks with the grammar found on TOEIC, researchers (Uchibori, Chujo & Hasegawa 2006) found, not surprisingly, that there was not much overlap. Most notably, the most frequent grammatical forms found in TOEIC were the structures of noun and verb phrases, and these are generally not taught in high school. Whether or not one assumes proficiency tests such as the TOEIC really do provide an effective measure of communicative ability, the reality is that many Japanese students will have to take these tests, and that they are ill prepared to do so. Similarly, Chujo (2003) compared the vocabulary taught in the best-selling Japanese junior and senior high school textbooks with the vocabulary found in TOEIC and found, again, that there is very little overlap. From the Uchibori et al (2006) study, we know what grammatical features are missing: the structures of noun and verb phrases, subject-predicate relations, adverbs, conjunctions, prepositions, and passive voice, among others. From the Chujo (2003) study, we have TOEIC-specific vocabulary. How then do we teach this grammar and vocabulary to beginner level university students in a way that actively engages them?

In Japan, the traditional pedagogical approach had been teacher-centered presentation-practice-production (PPP) methodology and this textbook-oriented approach is often used in EFL classrooms in Japan. McCarthy and Carter (1995) have argued that learners may need to supplement an inductive approach in addition to a deductive approach offered by PPP, and propose their three Is: illustration, interaction and induction. They also make the point that if teaching conversation or spoken English, it is important to teach spoken forms of language, not written forms as is so often seen in textbooks. Biber et al (2004:1) point out that the sequences of grammatical structures taught in most textbooks reflect intuitions about language rather than the language "actually used by

[15] Kiyomi Chujo teaches at the College of Industrial Technology, Nihon University, Japan. Her current research interests are vocabulary selection, e-learning, and the pedagogical applications of corpus linguistics.

[16] Kathryn Oghigian teaches at the Center for English Language Education for Science and Engineering, Waseda University.

speakers and writers in natural situations" and advocates basing learning material on real life language found in corpora. Other researchers (Aston 1995; Boulton 2007; Johns 1991) advocate allowing learners to find patterns in language through the exploration of corpus data, and propose a corpus-based "observation, hypothesis formation, use" methodology (Boulton 2007:3). On the other hand, it has been noted that learners may be more comfortable with grammatical rules and vocabulary lists rather than "lexical chunks" found in concordancing lines (Widdowson 2000 and Cook 1998, as discussed in Hunston 2002:193). Finally, Muranoi (2006) advocates a PCPP approach---presentation, comprehension, practice and production----incorporating cognitive processes such as noticing, hypothesis formation, hypothesis testing, automatization, and storage. In this case study, we have blended aspects of both an inductive corpus-oriented "observation, hypothesis formation, use" paradigm with a deductive PCPP approach (Muranoi 2006) that involves noticing, hypothesis formation and testing, automatization and storage.

Methodology: Subjects and Materials

Two classes of beginner level freshmen engineering students met for 90 minutes a week, for twenty-two weeks in one academic year. On the first day, students were given a TOEIC Bridge test which is a simplified version of the TOEIC that targets beginner and intermediate learners. Scores for 21 students in Group 1 averaged 124.6 out of 180 and averaged 143.5 for 26 students in Group 2; these results confirm their "beginner level" status. In addition, all students were given a grammatical features pre-test. At the end of the year, all students took a grammatical features post-test and responded to a questionnaire on the usefulness of the course and its various aspects.

A new unit was covered in each class, for a total of twenty units. These are shown in **Table 1**. The grammar basics targeted were those identified by the Uchibori et al (2006) study, i.e. those grammatical features found in TOEIC (primarily in Parts VI and VII) but not generally taught in Japanese high school textbooks such as the structure of a noun phrase (NP), the structure of a verb phrase (VP), the structure of a prepositional phrase (PP), subject-predicate relations, and adverbs. The syllabus was designed to present basic information first, such as inflections and derivations, since these are central to understanding the nouns, verbs, adverbs, and adjectives found in noun and verb phrases. In addition, the particular types of noun or verb phrases taught are those identified as the most common phrases used in English, according to Biber et al's (1999) *Longman Grammar of Spoken and Written English (LGSWE)*. For example, the most common pre-modifiers in an NP are general adjectives (*Ibid*: 589) and the most common post-modifiers are PPs (*Ibid*: 606); therefore, in teaching a noun phrase, we have emphasized [article + adj + noun] such as *a new location* and also [article + noun + PP] such as *the quality of education*. Our goal has been to use real language identified in the *LGSWE*, that will be encountered in grammatical structures found on TOEIC, as identified by Uchibori et al (2006), using TOEIC vocabulary identified by Chujo (2003) and to present them in a way that beginner level students can understand and use.

Spring semester		*Fall semester*	
1	Word classes, verb forms, derivations	1	Word classses
2	Verb forms	2	Structure of VP (1) (present/past progressive)
3	Structure of NP (1)	3	Structure of VP (2) (present perfect/progressive)
4	Structure of NP (2)	4	Structure of VP (3) (passive sentences)
5	Structure of NP (3), Adverbs (1)	5	Structure of VP (4) (transitive/intransitive verbs)
6	Structure of NP (4), Adverbs (2)	6	Structure of VP (5) (agreement)
7	Structure of NP (5)	7	Structure of VP (6) (to-infinitives，that-clauses)
8	Countable and uncountable nouns	8	Structure of VP (7) (to-infinitives，gerunds)
9	Structure of NP (6)	9	Structure of VP (8) (position of adverbs)
10	Structure of NP (7)	10	Structure of VP (9) (wh-clauses)

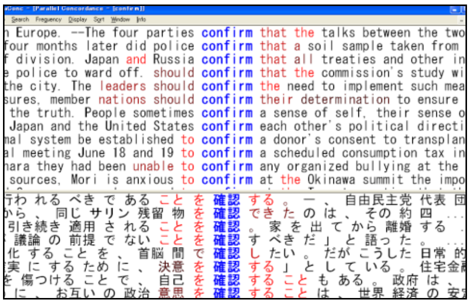
Table 1 2007 DDL Grammar Syllabus

The DDL Component

Observation

In the first step, the teacher introduced the topic and gave students clear instructions to follow guidelines on a carefully crafted DDL (data-driven learning) worksheet to find seven targeted words and their patterns of usage for the targeted grammar in the concordance lines on a computer by using a bilingual newspaper corpus (Utiyama

& Isahara 2003) and a parallel concordancer (Barlow 2004). Noticing is the first step in language building (Schmidt 2001:31), and as an effective feature of DDL, Key Word in Context (KWIC) concordance lines are considered effective in promoting this cognitive process. By having English concordancing lines together with their Japanese translations (see **Fig. 1**), beginner level students can understand what they are seeing. Thus, “learners can focus attention on a limited and hence controlled amount of data and the input becomes more manageable” (Gass 1997:8).



```
n Europe. --The four parties confirm that the talks between the two
four months later did police confirm that a soil sample taken from
f division. Japan and Russia confirm that all treaties and other in
e police to ward off. should confirm that the commission's study wi
the city. The leaders should confirm the need to implement such mea
sures, member nations should confirm their determination to ensure
 the truth. People sometimes confirm a sense of self, their sense o
 Japan and the United States confirm each other's political directi
mal system be established to confirm a donor's consent to transplan
al meeting June 18 and 19 to confirm a scheduled consumption tax in
hara they had been unable to confirm any organized bullying at the
 sources, Mori is anxious to confirm at the Okinawa summit the impo
```

```
行わ れる べき で ある こと を 確認 する 。 一 、 自由民主党 代表 団
から 、 同じ サリン 残留 物 を 確認 でき た の は 、 その 約 四 ...
引き続き 適用 さ れる こと を 確認 。 家 を 出 て から 離婚 する .
 議論 の 前提 で ない こと を 確認 す べき だ 」 と 語っ た 。 ...
化 する こと を 、 首脳 間 で 確認 し たい 。 だが こうした 日常 的
実 に する ため に 、 決意 を 確認 する 」 と し て いる 。 住宅金
を 傷つける こと で 、 自己 を 確認 する こと も ある 。 政府 は 、
に 、 お互い の 政治 意思 を 確認 する こと は 、 世界 経済 の 安
```

Fig. 1 An English-Japanese Concordancing Line

The goal for this exercise was for students to observe these seven words in a real context which illustrates the target grammar feature. As an example of a task, students were asked to search a word such as *quality* and list what words frequently come before and after (i.e., article before and preposition after, such as *the quality of education*) in order to understand the basic structure of a NP. Additional examples of DDL tasks appear in **Table 2**.

DDL tasks	*Examples*
examining different verb forms and derivations	Search ***develop**** and list both different verb forms and derivations. (Answer: *develop*, *develops*, *developed*, *developing*, and *development*.)
examining countable and uncountable nouns	Search ***furniture**** and ***passenger**** to understand which words are uncountable. (Answer: uncountable nouns have only one form, e.g. *furniture* NOT *furnitures*.)
examining word classes	Search a certain collocation such as ***a * organization*** to find what class of word comes between *a* and *organization*. (Answer: adjective, such as *a new organization*.)
examining the basic structure of NPs	Search a word such as ***quality*** and list what words frequently come before and after. (Answer: article before and preposition after; such as *the quality of education*)
examining the basic structure of VPs (1)	Search ***enjoyed*** and find which verb form frequently follows it. (Answer: *to*-infinitives or gerunds.)
examining the basic structure of VPs (2)	Search ***discuss***. Find and write down *wh*-clauses. (Various answers such as *to discuss when*.)
examining the basic structure of VPs (3)	Search ***agree*** and count how many *agree that* phrases you can find.

Table 2 Examples of DDL Tasks

Presentation and Hypothesis Formation

While some educators propose that an inductive process in grammar learning is essential (Seliger 1975), others advocate for a deductive approach (Shaffer 1989). We believe that what Corder (1973) claimed more than thirty years ago might well be true: that it is most effective to use a combination of both inductive and deductive approaches, and we use both as a basis for our instruction. By noticing patterns in the DDL tasks, students formed ideas about general rules for the targeted grammar. To confirm or clarify, next an explicit grammar explanation was presented with a generalized schema of the target structure by using a visual illustration such as a diagram (see **Fig. 2**). This was done in Japanese and students were provided with both a written and verbal explanation. In this particular illustration, the parentheses are used to show those words which are optional. There are two purposes for using an illustration such as this: (1) it is important to show as clearly as possible that the whole of a phrase is a coherent unit involving both obligatory and optional members of the phrase; and (2) it is necessary to show that elements within a phrase are grammatically related to one another, particularly with the head noun. The head noun determines the major properties of the phrase. In this example, the head noun forms a noun phrase.

[$_{NP}$	**(Determiner)**	**(Premodifier)**	**Head Noun**	**(Postmodifier)**	**]**
	article	adjective		prepositional phrase	
	quantifier	-ing		-ing/-ed	
	numeral	-ed		relative clauses	

Fig. 2 Grammar Explanation Used for Introducing the General Structure of the NP

Practice and Hypothesis Testing

After completing the DDL tasks, forming hypotheses about a discovered pattern or a rule, and having that understanding verified or corrected, the next phase is an intake or internalization phase to promote the hypothesis testing process. Students were given a follow-up worksheet, which contained tasks using the targeted grammar in contextualized exercises aimed at consolidating comprehension by providing practice and encouraging production. The follow-up exercises help students to understand the targeted structure inductively. One example of an activity would be for students to underline all noun phrases, or to complete a sentence by choosing the correct derivation of a word.

Production: Automatization and Storage

Just having a corpus and corpus inspection software is not enough (Braun et al 2006). We also know that "extra form and meaning focused input, tasks and assignments have to be added in order to make instruction more effective" (Koenraad & Hajer 2006:6). The production phase corresponds to the automatization and storage phases in the final cognitive process. In the follow-up worksheet, students were asked to perform tasks such as describing pictures and creating dialogues using targeted vocabulary.

The CALL Component

In order to review and learn vocabulary, students completed a 30-minute CALL program. They are given 20 words and their meanings, including the seven words given in the DDL component, and they learn these words in a series of CALL activities.

Results and Discussion

For these two classes, students in both Group 1 and Group 2 made significant gains in all areas of the grammatical features test except for countable and uncountable nouns. Specifically, there were gains for word classes, derivations, structure of noun phrases, structure of verb phrases, and multi-aspect (TOEIC-formatted) questions, but not for countable and uncountable nouns (see **Table 3**). In the noun phrases section of the test, students were asked to underline noun phrases found in sentences. The post-test scores showed a significant gain and we can say the enhanced noun phrase DDL instruction conducted in this study was effective. On a specially constructed verb phrases section of the test, students were asked to answer fill-in-the-blank questions. The post-test showed significant gains for the use of transitive and intransitive verbs, adverb placement, gerunds, *wh*-clauses, to-infinitives, passive tense verbs, past tense verbs, and *that*-clauses.

Grammatical features	*Class 1*			*Class 2*		
	Pre-test (%)	*Post-test (%)*	*Gain*	*Pre-test (%)*	*Post-test (%)*	*Gain*
(1) Word classes	69	81	12**	82	90	8**
(2) Countable & uncountable nouns	66	74	8	73	75	2
(3) Verb forms	60	75	15**	75	80	5
(4) Derivations	33	53	20**	51	67	16**
(5) Structure of noun phrases	33	68	35**	56	76	20**
(6) Structure of verb phrases	55	73	18**	70	78	8**
(7) TOEIC-formatted test questions	39	51	12**	58	71	13**

Table 3 Comparisons of the Pre- and Post-test Scores

It is important to note that most of the exercises and the questions on the grammatical features tests target a single aspect of grammar, i.e. identifying a part of speech, providing the correct derivation of a word, or using the correct verb form. In contrast, the multi-aspect or more complex TOEIC-formatted test questions present the learner with more difficult questions. An example question is as follows, and the answer must be chosen from four choices given for each question:

We will (   ) have a trip to China next spring. {probable, prove, probably, probability}

In order to choose the correct answer, the respondent would have to understand not only that an adverb is required, but that an adverb generally ends in *–ly* and that the adverb could appear in the space indicated. Compared to the other six types of features shown in **Table 3**, which focus on a single grammar issue, this particular type of TOEIC-formatted test question is complex and requires learners to bring together knowledge of more than one aspect of grammar.

With regard to student questionnaires, on a five-point scale, students reported that the DDL concordancing activities, grammar presentation, and follow-up activities were useful and easy to understand (see **Table 4**) and that in general, this course was useful for learning both grammar and vocabulary. The students particularly appreciated the explicit explanation that was given after they had an opportunity to explore and discover the targeted grammar. This helped them to clarify their understanding in a way that was familiar to them, and providing it after rather than before seemed to reinforce their understanding..

Activities	*Evaluation*	*Group 1*	*Group 2*
DDL Worksheets	useful easy to understand useful for learning grammar	4.0 3.7 3.9	4.0 4.0 4.0
Grammar Presentation	useful easy to understand useful for learning grammar	3.8 3.6 3.9	4.0 3.9 4.0
Follow-Up Activities	useful easy to understand useful for learning grammar	4.0 3.7 3.9	4.2 4.0 4.1
Feedback from Teacher	useful	4.3	4.4
DDL/CALL Course	enjoyable useful for learning grammar useful for learning vocabulary	3.5 3.8 4.0	3.7 3.9 4.3

Table 4 Student Assessment of the DDL Course

In addition, students also provided written responses to open ended questions such as (1) “What are some of the differences between this DDL class and your high school English class?” (2) “What did you find using DDL?” and (3) “How do you want to use a parallel corpus?” An excerpt of sample responses is available in the **Appendix**.

Conclusion

There are those who might argue that this kind of course is teaching to the TOEIC exam, rather than teaching the communication skills that the TOEIC would measure. The course is designed for beginner level university freshmen only to close the gap in what high school students are not taught so that as they continue through their other English courses, they are equipped to recognize and produce these important language forms. In addition, the follow-up activities are geared to production, not to exam strategies.

In his interesting 2007 article, Boulton remarks that there is a lack of data in the literature on whether or not “big themes” can be learned through corpus studies such as concordancing. He cites researchers who believe big themes can only be understood at the discourse level (Boulton 2007:4) but concludes that there is a lack of studies on learning grammatical themes with DDL. He reports having some success with learners’ ability to detect patterns of use for *will* and *going to*, but acknowledges that this is only one small aspect grammar.

As an on-going research project, we make modest gains each year, and we believe we are able to teach “big themes” in grammar patterns. We have been successful in addressing important vocabulary and grammatical features that appear not only on TOEIC tests, but in the English language. The forms focused on are those cited in the *LGSWE* as those most frequent in English. The students continue to report that they find this combined method to be both enjoyable and effective. We hope that it will provide these students with the tools they need to be successful in language classes and in communicating in English.

Appendix An Excerpt of Sample Responses from Students

- DDL was different from traditional learning. Rather than using textbooks, we used computers and learned actively at our own pace.
- It was discovery learning, so we can understand grammar gradually and solidly. We think this is a suitable style for engineering students.
- We can learn grammar visually by observing a large number of authentic and practical examples.
- We learned different meanings from those in dictionaries.
- We learned that one word has various usages.
- We learned that some words are used commonly and others are used rarely.
- I'd like to use DDL when I'd like to know if the word is used in a real context or not.

References

Aston, G. 1995. “Corpora in language pedagogy: matching theory and practice.” In *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson.* G. Cook and B. Seidlhofer (eds.). Oxford: Oxford University Press, 257-270.

Barlow, M. 2004. ParaConc (computer software). http://www.athel.com/para.html.

Biber, D., Johansson, S., Leech, G., Conrad, S., and **Finegan, S.** 1999. *Longman Grammar of Spoken and Written English.* Edinburgh Gate: Pearson Education Limited.

Biber, D., Conrad, S., Reppen, R., Byrd, H.P., Helt, M., Clark, V., Cortes, V. Csomay, E., and **Urzua, A.** 2004. “Representing language use in the university: analysis of the TOEFL 2000 spoken and written academic language corpus.” *ETS TOEFL Monograph Series MS-25; RM-04-03.* Princeton, New Jersey: Educational Testing Service. http://www.ets.org/research/researcher/RM-04-03.html [Access date 2/10/08]

Boulton, A. 2007. “DDL is in the details….and the big themes.” Proceedings from the *2007 Corpus Linguistics Conference, University of Birmingham, UK.* July 27-30, 2007. http://www.corpus.bham.ac.uk/corplingproceedings07/ [Access date 2/10/08]

Braun, S., Kohn, K., and **Mukherjee, J.** 2006. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods.* Frankfurt am Main: Peter Lang.

Chujo, K. 2003. “Eigo shokyuusha-muke TOEIC-goi 1 to 2 no sentei to sono kouka (Selecting TOEIC vocabulary 1 & 2 for beginner level students and measuring its effect on a sample TOEIC test).” *Journal of the College of Industrial Technology, Nihon University* 36: 27-42. (In Japanese.)

Chujo, K. , Nishigaki, C. and **Uchibori, A.** 2007. “Parallel corpus wo riyoushita bunpou hakken-gakushuu no kokoromi (Discovering grammar basics with parallel concordancing).” *Journal of the College of Industrial Technology, Nihon University* 40: 33-46. (In Japanese.)

Cook, G. 1998. “The uses of reality: a reply to Ronald Carter.” *ELT Journal* 52 /1: 57-63.

Corder, S. P. 1973. “Pedagogic grammars.” In *Grammar and Second Language Teaching: A Book of Readings,* W. E. Rutherford and M. S. Smith (eds.) 1988. New York: Newbury House, 123-145.

Gass, S. 1997. *Input, Interaction, and the Second Language Learner*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Green, B. 2006. “A framework for teaching grammar to Japanese learners in an intensive English program.” *The Language Teacher,* 30 /2: 3-11.

Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Johns, T. 1991. “Should you be persuaded: two examples of data-driven learning.” In *Classroom Concordancing,*

T. Johns & P. King (eds.) *ELR Journal,* 4. Birmingham: Birmingham University Press, 1-16 (as cited in Partington A.,1998).

Koenraad, T. and **Hajer, S.** 2006. "Towards a linguistically scaffolded curriculum. How can technology help?" In *Technologies for a Content and Language Integrated Approach to Dropout Problems in Higher Education,* (eds.) T. Koenraad, M. Hajer, M. Simons, and R. van der Werf. Arno Academic Publications Online,1-16.

http://hbo-kennisbank.uvt.nl/cgi/hu/show.cgi?fid=11820 [Access date 4/16/08]

McCarthy, M. and **Carter, R.** 1995. "Spoken grammar: what is it and how can we teach it?" *ELT Journal,* 49/3: 207-218.

Muranoi, H. 2006. *Daini Gengo Shuutoku kara Mita Koukatekina Eigo Gakushuuhou Shidouhou [SLA Research and Second Language Learning and Teaching]*. Tokyo: Taishukanshoten. (In Japanese.)

Schmidt, R. 2001. "Attention." In: P. Robinson (ed.) *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, 3-32.

Seliger, H. 1975. "Inductive method and deductive method in language teaching: a re-examination." *International Review of Applied Linguistics*, 13: 1-18.

Shaffer, C. 1989. A comparison of inductive and deductive approaches to teaching foreign languages. *The Modern Language Journal*, 73/4: 395-402.

Uchibori, A.and **Chujo, K.** 2005. Daigaku shokyuu reberu gakushuusha no Eigo communication nouryoku koujou ni muketa CALL bunpou-ryoku yousei-you software no kaihatsu [The development of grammar CD-ROM material to improve communicative proficiency of beginner level college students], *Journal of the College of Industrial Technology*, *Nihon University*, 38: 39-49. (In Japanese.)

Uchibori A., Chujo, K. and **Hasegawa, S.** 2006. "Toward better grammar instruction: bridging the gap between high school textbooks and TOEIC." *The Asian EFL journal* 8/2: 228-253. http://www.asian-efl-journal.com/index.php

Utiyama, M. and **Isahara, H.** 2003. "Reliable measures for aligning Japanese-English news articles and sentences." *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (ACL-2003): 72-79.

Widdowson, H.G. 2000. "On the limitations of linguistics applied." *Applied Linguistics* 21: 3-25.

MULTIMODAL FUNCTIONAL-NOTIONAL CONCORDANCING

Francesca Coccetta[17]

Abstract

Spoken texts provide a large quantity of information which extends beyond language; they include semiotic resources such as gesture, posture, gaze and facial expressions which, like language, contribute to the overall meaning-making of the texts (Kress and van Leeuwen 2006: 41). However, until recently their investigation has completely relied on their 'basic' orthographic transcriptions (Leech 2000), partly due to the lack of adequate concordancing software tools. This has somewhat limited the potential spoken texts bring to language teaching and learning. Based on the theoretical and technical innovations which have taken place in the field of multimodal corpus linguistics (Baldry and Thibault 2001, 2006b, forthcoming), especially within the MCA project (Baldry in press, Baldry and Thibault in press), this study presents a pedagogical application of spoken corpora in the promotion of communicative language competence by language learners at various levels of proficiency. In particular, it illustrates how MCA, a multimodal concordance (Baldry 2005, Baldry and Beltrami 2005), can be used to create, annotate and concordance spoken corpora in terms of functions and notions (van Ek and Trim 1998, 2001). The study illustrates the kind of information the concordance lines and their associated film clips provide in terms of: a) the linguistic forms realizing a specific language function and b) the ways in which language interacts with its multimodal co-text (Baldry in press). In so doing, the paper introduces a new concordancing technique, namely multimodal functional-notional concordancing (Coccetta in press b), and presents two multimodal data-driven-learning (DDL) activities which show how this new approach to the analysis of spoken texts can enhance language learning.

Keywords: Spoken corpus, multimodal functional-notional concordancing, communicative competence, multimodality, teaching materials

Introduction

Spoken corpora are particularly useful for teaching the spoken language because they "can achieve high authenticity, serve as communication aids, and provide irreplaceable models of the target language" (Mauranen 2004a: 208). However, the common practice of adopting the approach used to investigate corpora of written texts has greatly limited their potential for doing so. In the light of the recent theoretical and technological developments regarding the analysis of multimodal corpora (Baldry and Thibault 2001, 2006a, forthcoming), this study investigates *multimodal functional-notional* concordancing, a new approach to the analysis of spoken corpora, which focuses on language and on the ways speakers express things, but also considers how language combines with other semiotic resources such as gesture, facial expressions and gaze to create meaning. This approach is facilitated by the use of the online multimodal concordancer *MCA* (*Multimodal Corpus Authoring System*) (Baldry 2005, Baldry and Beltrami 2005). This study analyses some texts in the Padova Multimedia English Corpus (Padova MEC) (Ackerley and Coccetta 2007) and in so doing demonstrates how *MCA* can be used to investigate spoken corpora for functions and notions. It also describes the ties existing between verbal and non-verbal information that the individual concordances and their related video clips reveal. In passing, we may recall that functions are defined as "the kind of things people may *do* by means of language" (van Ek and Trim 1998: 28). Notions, on the other hand, are "the concepts we may refer to while fulfilling language functions" (van Ek and Trim 1998: 28). Van Ek and Trim distinguish between general and specific notions. The former can be expressed in any situation and include concepts such as *space*, *quantity* and *time*. The latter are topic-related and can be expressed in particular situations only. They include thematic categories such as *education*, *travel* and *personal identification*. The last part of the study gives examples of teaching materials developed for language learners at various levels of proficiency that are designed to promote their communicative language competence and raise their awareness of the multimodal nature of spoken texts. The teaching materials adopt the data-driven (DDL) approach "which gives the learner access to the facts of linguistic "performance"" (Johns 1991: 2) and

[17] *Francesca Coccetta is a PhD student in English Linguistics at the University of Padua with a research grant awarded by the inter-University research project eColingua. Her fields of interest include multimodal corpus linguistics, the use of ICT in language teaching and learning, and in e-learning in particular. She has worked as a language advisor at the University of Padua's Language Centre for about 3 years and as an online tutor for the University of Venice's School of Educati on teacher training programme.*

which, through deductive and inductive reasoning, encourages the learner to discover patterns in the target language.

Multimodal concordancing

The research carried out in the new field of multimodal corpus linguistics (Baldry and Thibault 2001, 2006b, forthcoming) in the last ten years offers both the conceptual and software tools to examine a variety of multimodal texts, such as advertisements (Baldry and Thibault 2006a), conversational texts (Coccetta in press a/b, Dalziel and Metelli in press) and websites (Baldry and O'Halloran forthcoming). As regards the study of spoken texts, the development of the online multimodal concordancer *MCA*, capable of creating, annotating and concordancing multimodal texts, has overcome some of the limitations imposed by concordancers such as *WordSmith Tools* (Scott 2008) and *AntConc* (Anthony 2005) which arise as a result of the process of transposing a speech event to a written medium (Coccetta in press a/b). Unlike *WordSmith Tools* and *AntConc*, in *MCA* multimodal texts (film clips) are preserved in, more or less, their original format; for each concordance line the system provides contextualised access to specific parts of these texts, thus allowing the investigation of what Baldry (in press) defines as a concordance's *multimodal co-text*, namely a co-text (Sinclair 1991) which extends beyond the concordance itself and includes, besides language, other semiotic resources featured in the text. This concordancing practice is intersemiotic in its orientation as it considers language as being integrated rather than isolated from other semiotic modalities; in other words, language constantly interacts with other resources in the meaning-making process within the scalar organisation of texts (Baldry and Thibault 2006a: Chap. 4, 2006b, in press, Baldry in press).

Extensive research into multimodal concordancing within the *MCA* project has led to the development of a *Concordance Matrix* which combines four concordance types with four concordancing procedures within this scalar approach to concordancing based on the hypothesis that multimodal concordances can explore interactions across different textual levels in multimodal texts (Baldry in press, Baldry and Thibault, 2006b, in press, see below). Table 1 outlines this *Matrix* and indicates how from a theoretical standpoint any concordance type can be combined with any procedure.

Concordance types	Concordancing procedures
1. monomodal form-oriented concordances;	a. default type (KWIC concordancing);

2. monomodal meaning-oriented concordances;	b. media-indexed type;
	c. tabulated type;

3. multisemiotic form-oriented concordances;	d. overlay/captioned type.
4. multisemiotic meaning-oriented concordances.	

Table 1: A Concordance Matrix presenting a series of options to investigate multimodal corpora within the MCA project (Baldry in press, Baldry and Thibault in press);

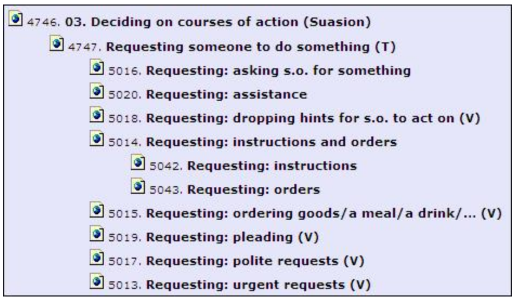
The *multimodal functional-notional concordancing* approach presented in this study is an example of Option 2b in the *Matrix*, namely the *monomodal meaning-oriented, media-indexed* concordance. In other words, this concordancing technique is *monomodal* because the concordance itself only provides information on one semiotic resource, namely language, and excludes an analysis of, say, gaze or gesture; it is *meaning-oriented* because it focuses on a specific language function and investigates the different language forms which enact this function; finally, it is *media-indexed* because it gives access to the film sequence indexed in each concordance line, i.e. its *multimodal co-text*. The word 'multimodal' included in the expression 'multimodal functional-notional concordancing' thus refers to the *multimodal co-text* rather than the concordance itself.

Sections 3 and 4 illustrate how *MCA* can be used to construct, annotate and concordance spoken texts.

Annotating a spoken corpus for functions and notions with MCA

The *MCA* system adopts a manual approach to tagging which allows users to: *a)* segment texts into functional units; *b)* create *mini-grammars*, namely the sets of descriptors which specify the features to be investigated in texts; and *c)* annotate texts.

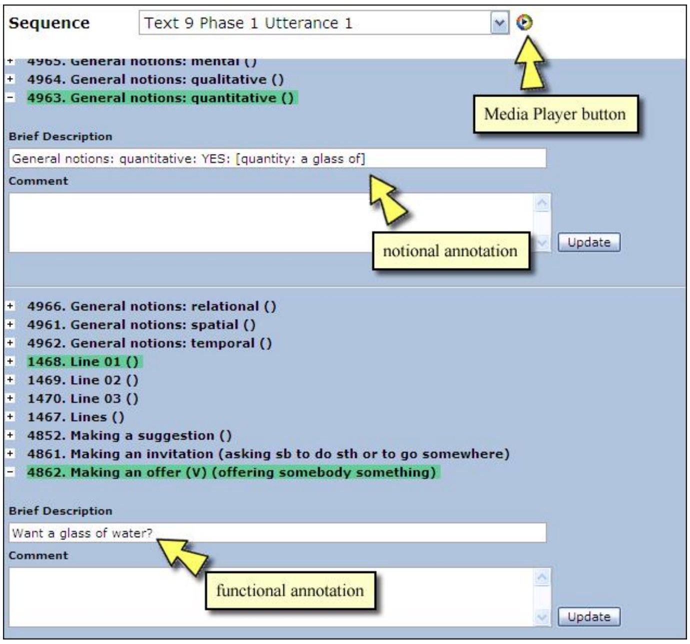
To be searchable, an MCA corpus needs to have at least one mini-grammar. To annotate the linguistic features of the texts in the Padova MEC, three mini-grammars have been developed: one for language functions, one for general notions and one for specific notions. The mini-grammars are based on the specifications given by van Ek and Trim in *Threshold 1990* (1998) and *Vantage* (2001). An excerpt from the mini-grammar for the function ‘requesting someone to do something’ is given in Figure 1.



- 4746. 03. Deciding on courses of action (Suasion)
  - 4747. Requesting someone to do something (T)
    - 5016. Requesting: asking s.o. for something
    - 5020. Requesting: assistance
    - 5018. Requesting: dropping hints for s.o. to act on (V)
    - 5014. Requesting: instructions and orders
      - 5042. Requesting: instructions
      - 5043. Requesting: orders
    - 5015. Requesting: ordering goods/a meal/a drink/... (V)
    - 5019. Requesting: pleading (V)
    - 5017. Requesting: polite requests (V)
    - 5013. Requesting: urgent requests (V)

Figure 1: An excerpt from the mini-grammar for the function ‘requesting someone to do something’;

As Figure 1 exemplifies, *MCA* mini-grammars are hierarchically organized on the basis that the deeper you go into the hierarchy the more specific the descriptors are. For example, the function in question belongs to the category of functional use ‘Suasion’ and comprises functions such as ‘polite requests’, ‘pleading’ and ‘instructions and orders’ which are more specific than the function ‘requesting someone to do something’. Similarly, the function ‘requesting: instructions and order’ comprises the more specific functions ‘requesting: instructions’ and ‘requesting: orders’. The hierarchical organization of the mini-grammars allows users to decide how deeply in the analysis of the texts they want to go. Once the tagging system has been created, the corpus is divided into functional units of any length within a prototypically scalar approach to concordancing that divides texts into macro and micro structures (Baldry, Thibualt 2001, 2006a, 2006b, in press, forthcoming). Thus, in the Padova MEC, texts are divided into phases (Thibault 2000: 320, Baldry and Thibault 2006a: Chap. 4) and utterances. Finally, each sequence is associated to the descriptors that characterize it. Figure 2 shows the *Sequence Analysis* tool used to annotate texts and which also contains the tagging systems created.



Sequence: Text 9 Phase 1 Utterance 1

+ 4965. General notions: mental ()
+ 4964. General notions: qualitative ()
− 4963. General notions: quantitative ()

Brief Description: General notions: quantitative: YES: [quantity: a glass of]

Comment:

+ 4966. General notions: relational ()
+ 4961. General notions: spatial ()
+ 4962. General notions: temporal ()
+ 1468. Line 01 ()
+ 1469. Line 02 ()
+ 1470. Line 03 ()
+ 1467. Lines ()
+ 4852. Making a suggestion ()
+ 4861. Making an invitation (asking sb to do sth or to go somewhere)
− 4862. Making an offer (V) (offering somebody something)

Brief Description: Want a glass of water?

Comment:

Figure 2: An example of functional-notional annotation;

Figure 2 gives an example of how an utterance is annotated for functions and notions. As for the utterance ‘Want a glass of water?’, the function ‘making an offer’ is selected and the utterance in question is written in the empty box below it. Similarly, to tag the general notion of quantity “a glass of”, the descriptor ‘general notions: quantitative’ is selected and “a glass of” is written in the empty box below it. The *Media Player* button at the top of the page facilitates the annotation of the sequence, as it allows the user to watch it.

Multimodal concordancing for functions and notions

The use of a spoken corpus to learn about typical ways of expressing something has been called for by Mauranen (2004b: 95). With reference to the functional tagging system described above and the use of the search engine incorporated in *MCA*, this section will exemplify how the Padova MEC can be investigated for functions and, in particular, how a function and a notion can be combined.

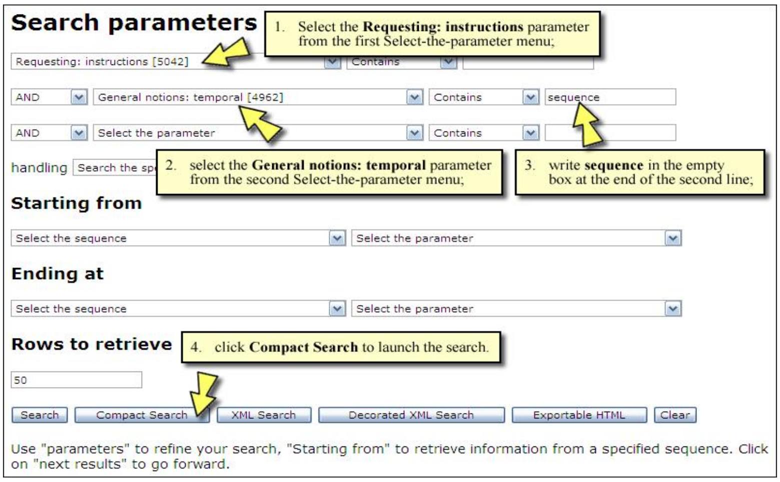
MCA incorporates a search engine (see Figure 3) which allows users to find and isolate sequences in a corpus sharing the same characteristics. For example, to see if the function ‘declining an offer’ is expressed in a Padova MEC subcorpus relating to *requests*, *invitations* and *offers* and to find what linguistic forms realize this function, the respective parameter is selected from one of the three drop-down menus included in *MCA*’s search engine. Table 2 presents the concordance lines retrieved.

1. No thanks
2. no thanks.
3. No thanks. I mean, that - that water's been there for ages.
4. No.
5. No thanks. I'm not - I'm not hungry.
6. Uh, no thanks
7. I've already had one, thanks.

Table 2: A set of results for the function ‘declining an offer’ retrieved with MCA;

Access to the film clip for each concordance line gives information about the multimodal co-text which might turn out to be useful for language learners. Research has shown the relevance of texts that use many semiotic resources in helping language learners, in particular lower level ones (Mueller 1980, Hoven 1999), to understand spoken texts (Kellerman 1992, Sueyoshi and Hardison 2005). For example, in the majority of the concordance lines presented in Table 2 the speaker shakes her/his head when s/he says “No”, thus showing the synchronisation that exists between the gesture and speech in the meaning making process, and giving learners clues as to how these resources are typically codeployed and used concurrently.

We may illustrate the capacity for integration of textual levels and resources in this approach to language learning and teaching with a further example. As *MCA* allows users to combine up to three descriptive parameters, it is easily possible to analyse adjacency pairs (Schegloff and Sacks 1973) (see Coccetta in press a/b). Similarly, users can combine a language function with one or two general or specific notions. With reference to a cooking demonstration taken from the Padova MEC, Figure 3 shows how *MCA*’s search engine can be set in such a way as to find the utterances expressing the function ‘requesting: instructions’ which contain a temporal notion of the sequence type (e.g. *then*, *first* and *afterwards*).



Search parameters

1. Select the **Requesting: instructions** parameter from the first Select-the-parameter menu;

Requesting: instructions [5042] | Contains

AND | General notions: temporal [4962] | Contains | sequence

AND | Select the parameter | Contains

2. select the **General notions: temporal** parameter from the second Select-the-parameter menu;

3. write **sequence** in the empty box at the end of the second line;

Starting from

Select the sequence | Select the parameter

Ending at

Select the sequence | Select the parameter

Rows to retrieve

4. click **Compact Search** to launch the search.

50

Use "parameters" to refine your search, "Starting from" to retrieve information from a specified sequence. Click on "next results" to go forward.

Figure 3: MCA’s Search Inquiry tool with instructions to find utterances expressing the function ‘requesting: instructions’ which contain a temporal notion of the sequence type;

Table 3 shows the results that this search produces for this text:

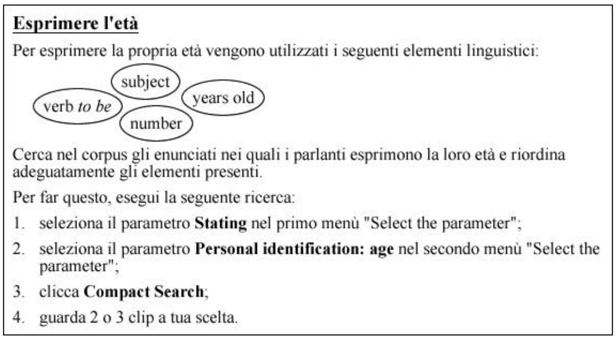
1. You take yogurt and then you mix it with tandoori special blend.
   General notions: temporal: YES: [reference without time focus: simple present] [sequence: then]
2. we cut it first [...] Do that and you do some slabs on the chicken. [ ] And do the same on the other bit.
   General notions: temporal: YES: [reference without time focus: simple present] [sequence: first] [present reference: now]
3. after you've left this for 30 minutes which we haven't done, but it doesn't matter, um, you coat it with this marinade.
   General notions: temporal: YES: [sequence: after + sub-clause] [duration: for] [divisions of time: minutes] [past reference: present perfect]
4. then you put this on. [ ] You put this on and, um, you turn the oven on at 170 degrees [...] And then you put salt all over it.
   General notions: temporal: YES: [reference without time focus: simple present] [sequence: then]

Table 3: The set of results for the function 'requesting: instructions' and the temporal notion of the sequence type retrieved with MCA;

Section 5 will further exemplify this approach by giving some sample DDL activities which analyse language functions and notions.

Multimodal DDL activities for functions and notions

Giving personal details about name, age, profession, family and address is one of the first things language learners learn to do in a language (see the specifications for the A1 level in the global scale of the *Common European Framework of Reference*, Council of Europe 2001: 24). Exercise 1 illustrates how a Padova MEC subcorpus relating to short introductions can be used to help language learners with a low level of proficiency to express their age. Because of the learners' level, the exercise is given in the native language.



Esprimere l'età

Per esprimere la propria età vengono utilizzati i seguenti elementi linguistici:

subject
verb *to be*
years old
number

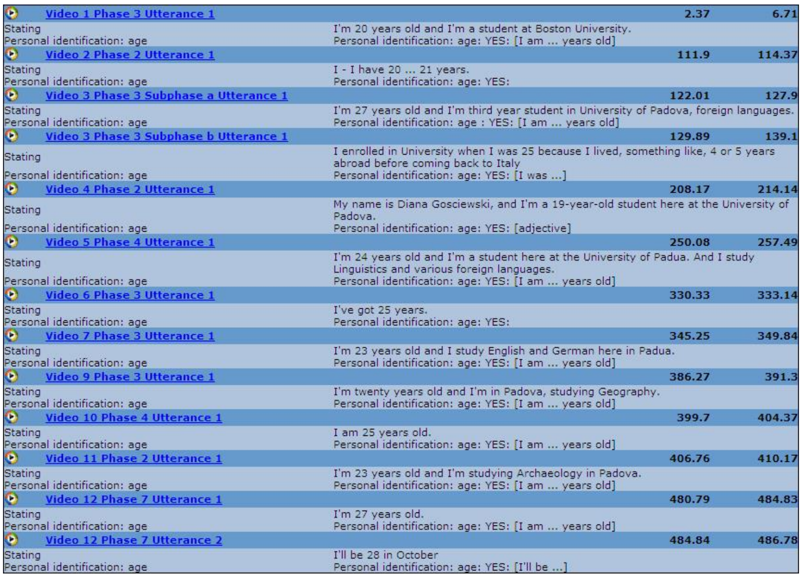
Cerca nel corpus gli enunciati nei quali i parlanti esprimono la loro età e riordina adeguatamente gli elementi presenti.

Per far questo, esegui la seguente ricerca:

1. seleziona il parametro **Stating** nel primo menù "Select the parameter";
2. seleziona il parametro **Personal identification: age** nel secondo menù "Select the parameter";
3. clicca **Compact Search**;
4. guarda 2 o 3 clip a tua scelta.

Exercise 1: Expressing age

In Exercise 1 learners are given the elements used to express age and are asked to put them in the correct order. To do so, they are required to search the corpus for the function 'stating' and the specific notion 'personal identification: age'. This produces the concordance lines presented in Figure 4.

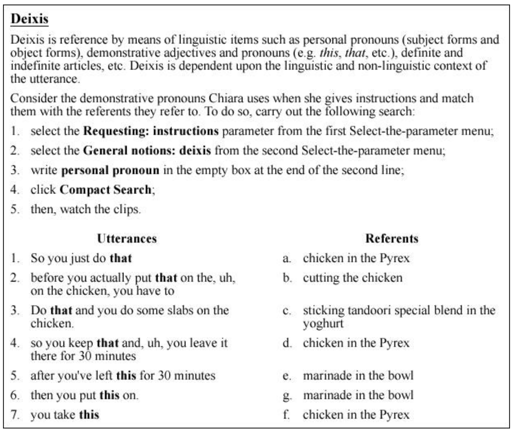


Video 1 Phase 3 Utterance 1		2.37	6.71
Stating Personal identification: age	I'm 20 years old and I'm a student at Boston University. Personal identification: age: YES: [I am ... years old]		
Video 2 Phase 2 Utterance 1		111.9	114.37
Stating Personal identification: age	I - I have 20 ... 21 years. Personal identification: age: YES:		
Video 3 Phase 3 Subphase a Utterance 1		122.01	127.9
Stating Personal identification: age	I'm 27 years old and I'm third year student in University of Padova, foreign languages. Personal identification: age : YES: [I am ... years old]		
Video 3 Phase 3 Subphase b Utterance 1		129.89	139.1
Stating Personal identification: age	I enrolled in University when I was 25 because I lived, something like, 4 or 5 years abroad before coming back to Italy Personal identification: age: YES: [I was ...]		
Video 4 Phase 2 Utterance 1		208.17	214.14
Stating Personal identification: age	My name is Diana Gosciewski, and I'm a 19-year-old student here at the University of Padova. Personal identification: age: YES: [adjective]		
Video 5 Phase 4 Utterance 1		250.08	257.49
Stating Personal identification: age	I'm 24 years old and I'm a student here at the University of Padua. And I study Linguistics and various foreign languages. Personal identification: age: YES: [I am ... years old]		
Video 6 Phase 3 Utterance 1		330.33	333.14
Stating Personal identification: age	I've got 25 years. Personal identification: age: YES:		
Video 7 Phase 3 Utterance 1		345.25	349.84
Stating Personal identification: age	I'm 23 years old and I study English and German here in Padua. Personal identification: age: YES: [I am ... years old]		
Video 9 Phase 3 Utterance 1		386.27	391.3
Stating Personal identification: age	I'm twenty years old and I'm in Padova, studying Geography. Personal identification: age: YES: [I am ... years old]		
Video 10 Phase 4 Utterance 1		399.7	404.37
Stating Personal identification: age	I am 25 years old. Personal identification: age: YES: [I am ... years old]		
Video 11 Phase 2 Utterance 1		406.76	410.17
Stating Personal identification: age	I'm 23 years old and I'm studying Archaeology in Padova. Personal identification: age: YES: [I am ... years old]		
Video 12 Phase 7 Utterance 1		480.79	484.83
Stating Personal identification: age	I'm 27 years old. Personal identification: age: YES: [I am ... years old]		
Video 12 Phase 7 Utterance 2		484.84	486.78
Stating Personal identification: age	I'll be 28 in October Personal identification: age: YES: [I'll be ...]		

Figure 4: The set of results for the function 'stating' and the specific notion 'personal identification: age' retrieved with MCA

By analysing the concordance lines learners are able to place the elements in the correct order. In addition, by accessing the original text they get used to hearing numbers in utterances. The results retrieved are typically used to develop some follow-up exercises. For example, the concordance line "I'm a 19-year-old student here at the University of Padova" illustrates the use of a compound adjective to express age. The concordance lines "I – I have 20 … 21 years" and "I've got 25 years", produced by non-native speakers of English, on the other hand, illustrate the incorrect use of the verb 'to have' to express age, a mistake which is very common in language learners, and Italians in particular. Drawing the learners' attention to the mistake helps them avoid it. However restricted the set of examples may be in this subcorpus, an illuminating set of options of linguistic forms for expressing age is, nevertheless, made available to learners. Similarly, learners can be asked to combine the function 'stating' and the specific notion 'personal identification: name' to find out how to express their names, or the specific notion 'personal identification: address' to learn how to say where they live.

Access to the original text allows learners to practice listening, and provides information which might be difficult to retrieve when only a linguistic co-text is available. Exercise 2 was created with reference to the cooking demonstration mentioned above, and focuses on the use of deictic elements, and demonstrative pronouns in particular, when giving instructions.



Deixis

Deixis is reference by means of linguistic items such as personal pronouns (subject forms and object forms), demonstrative adjectives and pronouns (e.g. *this*, *that*, etc.), definite and indefinite articles, etc. Deixis is dependent upon the linguistic and non-linguistic context of the utterance.

Consider the demonstrative pronouns Chiara uses when she gives instructions and match them with the referents they refer to. To do so, carry out the following search:

1. select the **Requesting: instructions** parameter from the first Select-the-parameter menu;
2. select the **General notions: deixis** from the second Select-the-parameter menu;
3. write **personal pronoun** in the empty box at the end of the second line;
4. click **Compact Search**;
5. then, watch the clips.

Utterances	Referents
1. So you just do **that**	a. chicken in the Pyrex
2. before you actually put **that** on the, uh, on the chicken, you have to	b. cutting the chicken
3. Do **that** and you do some slabs on the chicken.	c. sticking tandoori special blend in the yoghurt
4. so you keep **that** and, uh, you leave it there for 30 minutes	d. chicken in the Pyrex
5. after you've left **this** for 30 minutes	e. marinade in the bowl
6. then you put **this** on.	g. marinade in the bowl
7. you take **this**	f. chicken in the Pyrex

Exercise 2: Using deictic elements when giving instructions;

Because the participants in the cookery lesson, i.e. speaker and listener, share the same context, the speaker uses a large number of deictic elements whose referent can most often be identified only by watching the video. For example, by watching the sequence where the utterance "do that and you do some slabs on the chicken" is produced, learners come to realise that the personal pronoun 'that' refers to the action of cutting the chicken. In the same way, while watching the sequence where the utterance "so you keep that and, uh, you leave it there for 30 minutes" is produced, they realise that here the demonstrative pronoun 'that' refers to the chicken in the bowl. Thus, the video clips-cum-multimodal cotexts reveal the difference in use between 'this' and 'that' – the former typically refers to an object close to the speaker, while the latter refers to an object far from the speaker. Indeed, in the utterance "after you've left this for 30 minutes", 'this' refers to the chicken in the bowl the speaker is holding in her hands, while in "so you keep that and, uh, you leave it there for 30 minutes" 'that' refers to the chicken in the bowl placed far away from the speaker.

Conclusions

This study has briefly illustrated how spoken corpora can be annotated for functions and notions and how they can be investigated accordingly through the use of *MCA*. In so doing, it has introduced the concept of multimodal functional-notional concordancing and contextualised it in relation to the different concordancing techniques developed in the field of multimodal corpus linguistics. The study has also exemplified some DDL activities which focus on the ways speakers express things and consider the multimodal co-text in which utterances are produced. In so doing, it has illustrated some of the benefits that this approach to the study of spoken texts brings to language teaching and learning.

References

Ackerley, K. and **Coccetta, F.** 2007. "Enriching Language Learning through a Multimedia Corpus." *ReCALL* 19/3: 351-370.

Anthony, L. 2005. "AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom". In *Professional Communication Conference, 2005. IPCC. Proceedings.* International, 729-737.

Baldry, A.P. 2005. *A Multimodal Approach to Text Studies in English. The Role of MCA in Multimodal Concordancing and Multimodal Corpus Linguistics.* Campobasso: Palladino.

Baldry, A.P. in press. "Turning to Multimodal Corpus Research for Answers to a Language-course Management Crisis." In *Corpora for University Language Teachers*, C. Taylor Torsello, K. Ackerley and E. Castello (eds.). Bern: Peter Lang.

Baldry, A.P. and **Beltrami, M.** 2005. "The MCA Project: Concepts and Tools in Multimodal Corpus Linguistics." In *Multimodality: Text, Culture and Use. Proceedings of the Second International Conference on Multimodality*, M. Carlsson, A. Løvland and G. Malmgren, (eds.). Kristiansand: Agder University College and Norwegian Academic Press, 79-108.

Baldry, A.P. and **O'Halloran, K.** forthcoming. *Multimodal Corpus-Based Approaches to Website Analysis.* London: Equinox.

Baldry, A.P. and **Thibault, P.J.** 2001. "Towards multimodal corpora." In *Corpora in the Description and Teaching of English*, G. Aston and L. Burnard (eds.). Bologna: CLUEB, 87-102.

Baldry, A.P. and **Thibault, P.J.** 2006a. *Multimodal Transcription and Text Analysis. A Multimedia Toolkit and Coursebook.* London and New York: Equinox.

Baldry, A.P. and **Thibault, P.J.** 2006b. "Multimodal Corpus Linguistics." In *System and Corpus: Exploring Connections*, G. Thompson and S. Hunston (eds.). London and Oakville: Equinix164-183.

Baldry, A.P. and **Thibault, P.J.** forthcoming. *Multimodal Corpus Linguistics.* London: Routledge.

Baldry, A.P. and **Thibault, P.J.** in press. "Applications of Multimodal Concordances in the University Teaching and Testing Cycle." In *Hermes* 41.

Coccetta, F. in press a. "First Steps towards Multimodal Functional Concordancing." *Hermes* 41.

Coccetta, F. in press b. "Multimodal Corpora with MCA." In *Corpora for University Language Teachers*, C. Taylor Torsello, K. Ackerley and E. Castello (eds.). Bern: Peter Lang.

Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Dalziel, F. and **Metelli, M.** in press. “Learning about Genre with a Multimodal Concordancer.” In *Interdisciplinary Perspectives on Multimodality: Theory and Practice*, A. Baldry and E. Montagna (eds.). Campobasso: Palladino.

Hoven, D. 1999. “A Model for Listening and Viewing Comprehension in Multimedia Environments.” *Language Learning and Technology* 3/1: 88–103.

Johns, T. 1991. “Should You Be Persuaded: Two Examples of Data-driven Learning Materials.” *Classroom concordancing, ELR Journal* 4: 1–16.

Kellerman, S. 1992. “‘I See What You Mean’: The Role of Kinesic Behaviour in Listening and Implications for Foreign and Second Language Learning.” *Applied Linguistics* 13/3: 239-58.

Kress, G. and **van Leeuwen, T.** 2006. *Reading Images. The Grammar of Visual Design.* London and New York: Routledge.

Leech, G. 2000. “Grammars of Spoken English: New Outcomes of Corpus-Oriented Research.” *Language Learning* 50/4: 675-724.

Mauranen, A. 2004a. “Speech Corpora in the Classroom.” In *Corpora and Language Learners*, G. Aston, S. Bernardini and D. Stewart (eds.). Amsterdam and Philadelphia: John Benjamins, 195-211.

Mauranen, A. 2004b. “Spoken Corpus for an Ordinary Learner.” In *How to Use Corpora in Language Teaching*, J. Sinclair (ed.). Amsterdam and Philadelphia: John Benjamins, 89-105.

Mueller, G. 1980. “Visual Contextual Cues and Listening Comprehension: An Experiment.” *Modern Language Journal* 64: 335–340.

Schegloff, E. and **Sacks, H.** 1973. “Opening Up Closings.” *Semiotica* 6/4, 289-327.

Scott, M. 2008. *WordSmith Tools Version 5.* Liverpool: Lexical Analysis Software.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford and Singapore: Oxford University Press.

Sueyoshy, A. and **Hardison, D.** 2005. “The Role of Gestures and Facial Cues in Second Language Listening Comprehension.” *Language Learning* 55/4: 661-99.

Thibault, P.J. 2000. “The Multimodal Transcription of a Television Advertisement: Theory and Practice.” In A.P. Baldry (ed.). *Multimodality and Multimediality in the Distance Learning Age.* Campobasso: Palladino, 311-385.

van Ek, J.A. and **Trim, J.L.M.** 1998. *Threshold 1990.* Cambridge: Cambridge University Press.

van Ek, J.A. and **Trim, J.L.M.** 2001. *Vantage.* Cambridge: Cambridge University Press.

AntConc: http://www.antlab.sci.waseda.ac.jp/software.html/.

MCA: http://mca.unipv.it/.

WordSmith Tools: http://www.lexically.net/wordsmith/.

THE COLLOCATIONS DICTIONARY FOR LEARNERS OF ENGLISH:
CORPUS DATA AND USER NEEDS

Stephen Coffey[18]

Abstract

This study examines the Oxford Collocations Dictionary for students of English (OCD). It is not intended as a balanced review of the dictionary, but rather concentrates on specific areas of lexicographical presentation where there appears to be room for improvement. The study is based on the analysis of two four-page sections of the dictionary. These contained 80 headwords and over 1600 collocational phrases. Where doubts arose regarding the usefulness or accuracy of the dictionary data, searches were made in the British National Corpus (hereafter BNC), upon which OCD was largely based.

Analysis suggests that the following are areas where improvements might be made: the structure of the entries for prepositional collocates; the choice of which collocates to include for a given headword; the importance given to examples of usage; the treatment of extended collocational units; collocates of specific word forms; the precision of phraseological form in general. Suggestions are also made as regards certain additional features which it might be useful to include within some dictionary entries, specifically, semantic labelling, relationship between collocational phrase and text type, and frequency data.

Keywords: Collocations, corpora, dictionaries, English, pedagogy

Introduction

Ideally, a learners' collocations dictionary should be corpus-based, thus offering both a reliable inventory of typical collocational phrases and a useful set of authentic contextualized examples. In the present paper, I report on a study of *The Oxford Collocations Dictionary for students of English* (hereafter OCD), which is, to my knowledge, the only corpus-based, monolingual collocations dictionary compiled for learners' of English as a foreign language.[19] The object of the study has been to identify any shortcomings in OCD, especially from the point of view of its being an interface between the corpus and the language learner. The study is not to be taken as a balanced review of the dictionary, since I am largely ignoring its many good points.

The structure of OCD

Entries in OCD are presented alphabetically, and headwords are either nouns, verbs, or adjectives. Where there are different parts of speech for lexical units with the same spelling, each word becomes a headword in its own right. Within each entry, the organization is as follows (and see the Figure below): (1) where a headword is judged to have clearly distinguishable senses, the entry is split into sub-entries, each with an indication of meaning; (2) collocates within each entry, or sub-entry, are grouped in the first instance according to their part of speech; (3) following this, collocates with similar semantic functions are grouped together (e.g. *be, look, seem*), though without semantic labelling; (4) in some cases, the collocational phrases are provided with short contextualized examples.

[18] Originally from Devon, in south-west England, Stephen Coffey studied modern languages at London University, before embarking on a career in language education. After teaching for relatively brief periods in England and Spain, he settled in Italy, where he has lived since 1984. In recent years, he has devoted more time to research activities. The main focus of his research is the field of lexico-phraseology, which he has investigated from a purely descriptive point of view, including English-Italian contrastive description, from a lexicographical perspective, and from an applied, foreign language learning perspective. He currently works at the Department of English Studies in the University of Pisa.

[19] There are other pedagogical dictionaries devoted to collocation, notably Benson et al (1997) and Hill and Lewis, eds. (1998), but they are not corpus-based. There is also a corpus-based combinatorial dictionary (Kjellmer 1994), but it is not a pedagogical reference tool.

mad *ad).*

1 not sane; crazy/stupid

• VERBS **be**, **look**, **seem I go** *He went mad and spent the rest of his life locked up in a mental hospital.* ◊ *The world had gone completely mad.* **I drive sb** *His experiences in the First World War drove him mad.* ◊ *The children are driving me mad!* **I consider sb**, **think sb** *Her colleagues thought her quite mad.* I **pronounce sb**

• ADV. **absolutely**, **completely**, **quite**, **utterly I barking**, **(stark) raving** *What a barking mad idea!* ◊ *You must be stark raving mad to risk your money like that!* **I almost I a bit**, **half**, **a little**, **slightly I dangerously**

• PREP **with** *I went mad with joy and danced a little jig.*

2 angry

• VERBS **be**, **feel**, **look I get** *I get so mad when people don't take me seriously.* **I make sb** *It makes me really mad when people waste food.*

• ADV. **hopping**, **really I absolutely I pretty**

• PREP. **at/with** *My mum's absolutely mad with me!*

A sample entry from the Oxford Collocations Dictionary

Methodology

The methodology of the study was as follows. Sample entries from OCD were analysed in order to verify whether the collocational phrases present are pedagogically useful ones, and whether they are presented in a pedagogically useful fashion. The sample consisted of two sets of four consecutive pages (the first two pages of the letters G and M). A total of 80 headwords were involved, 62 of which were nouns, 11 verbs and 7 adjectives. Some of these entries comprise two or more semantically determined sub-entries, making a total of 111 entries or subentries. In the sample there were a total of 1635 collocational phrases.[20] Where doubts arose regarding the usefulness or accuracy of the dictionary data, searches were made in the British National Corpus (hereafter BNC), upon which OCD was largely based. With regard to 'usefulness', it is taken as axiomatic that the main purpose of the collocations dictionary is to offer the learner assistance while engaged in (usually written) production tasks, a purpose endorsed by the OCD in its Introduction.

Shortcomings

Structure of entries: prepositions

One way in which the structure of some entries might be improved relates to the presentation of collocations designated as 'prepositions'. In the case of NOUN + VERB collocational phrases, presentation distinguishes between two directions: HEADWORD (noun) + *verb* and *verb* + HEADWORD (noun), for example GAP + *widen* and *bridge* + GAP. A similar distinction, however, is not made for prepositions. Thus, at the entry for **gang 1** (= group of criminals), one finds the following grammatical sub-entry:

PREP. **in a/the ~** *We were in the same gang.* | **~ of** *a gang of skinheads.*

I think it fair to say that the class of words generally termed 'preposition' is a difficult one for learners, and as much help as possible needs to be given to raise learner awareness of the different phrasal types involved (the vast

[20] In arriving at this figure, I have counted certain sets of collocates as single items, notably semantically restricted sets (e.g. *autumn / winter / etc*) and closely related lexical alternatives (e.g. *round / around*).

majority of prepositional usage must be learnt in terms of phraseology). In the case in hand, the specificity of direction would make it more obvious to learners that, for example, in the phrases *a mania for*, *mad with* and *manage on*, the preposition is selected together with the preceding word, which might be a noun, an adjective or a verb.

Choice of collocates

There are three ways in which the choice of collocates might be improved. Firstly, there are a small number of collocates which seem out of place because they are too predictable. Examples are: **magazine** → *new magazine*, *old magazine*, *read a magazine*; **man** [male person] → *old man*, *poor man*, *rich man*. Some prepositional uses are also fairly predictable. For example, a standard use of the word *for* is the following: 'You use **for** when you state or explain the purpose of an object, action or activity.' (Cobuild. 5, 2006). It would seem superfluous, therefore, to include examples such as *gadget for*, *machine for* and *machinery for*.

One cannot generalise, however, when it comes to 'obvious' words. Where there are close synonyms, exclusion of a common item may lead the learner to suppose that the latter is not a normal collocate. At the headword **machine**, for example, the adjectives *huge* and *large* are listed, but not *big*. The latter, however, is the most frequent of the three in the BNC. This situation may be compared with the entry for **garden**, where *big* and *large* are both listed.

A second area for improvement relates to the fact that some collocational phrases would not normally be accessed through the headwords at which they are listed. There are a number of different situations involved here. The most frequent appears to be that of headword nouns which are made more specific by preceding modifiers. It is unlikely, for example, that *fax machine* will be accessed through *machine*, or *information gap* through *gap*. Some other situations in which the learner would not look for particular collocational phrases are the following: 1) the headword is not being used with its typical meaning within a particular collocational phrase, for example **gate** → *sluice gate*; 2) from the point of view of familiarity and salience, it seems more likely that the learner would look up the collocate than the base (e.g. **gauge** → *fuel gauge*); 3) a headword noun is very adjectival in nature and it is more probable that the phrase would be accessed via the second noun, for example **majority 1** (= most) → *majority opinion*, *shareholder*, *stake*; 4) the whole collocational phrase is better viewed as a collocate of something else, for example **maniac** → *like a maniac*. Here, the dictionary example is *He was driving like a maniac*, which is a typical usage of *like a maniac*. The latter phrase would be better treated as the collocate and included in the entry for *drive* (and a few other pertinent verbs).

The third area for improvement with regard to collocates is the fact that the meaning of some collocational phrases may not be clear to the learner (it is to be remembered that there are virtually no explanations of meaning in OCD). Examples of this are: **garage** [for keeping cars in] → *integral garage*; **garden** → *rock garden*; **gardener** → *jobbing gardener, market gardener*; **gathering** [collecting] → *intelligence gathering*; **magistrate** → *licensing magistrate*; **maid** → scullery maid; **mammal** → *higher mammals*; **manage** → *manage sustainably*.

Examples of usage

It would be useful to have more contextualized examples. This would sometimes help clarify meaning and also indicate typical patterns of usage. It is worth underlining in this respect that only 423 of the collocational phrases examined (25.87 %) are shown within longer contexts. At the headword **manifestation**, for example, of the five adjective collocates given (*concrete*, *physical*, *visible*, *obvious* and *public*), only *public manifestation* is exemplified. Similarly, at the entry for **mail** *noun* (sub-section **mail** + VERB*)*, the verbal collocate *go* has no accompanying example, and one is left wondering what this NOUN + VERB combination (*mail* + *go*) means. Some further examples of collocational phrases for which examples would be useful are: *away* + **game**, **gasp** + *of*, **gather** + *up*, *information* **gathering**, *address* a **gathering**, **gauge** *of*, **mail** (verb) + *direct*, **manage** + *somehow*, and the noun **gala**, which has 16 collocates but no examples at all.

Collocates of specific word forms

Some collocates relate to specific morphological forms of word, but this is not always shown. An example is the noun *gain*, for which a number of adjectival collocates are given, but only some of which are indifferent to number. Data from the BNC suggests that, for example, *considerable*, *significant*, *efficiency* and *ill-gotten* are used much more with *gains* than with *gain*. Conversely, *personal* and *territorial* are used much more with the singular form. Another example is at the entry for *magistrate,* where the collocating verbs *hear* and *remand* are used above all with the plural *magistrates*.

Extended phraseology

Another area which could be improved is that of 'extended phraseology'. For example, there is an indication at the word **malice** of the verbal collocate *bear*, but no explicit statement that it is often used in a negative context. Similarly, at the headword **majority** [= most], *great* is listed as a single-word collocate, but there is no indication of the phrase *the great majority of* .... (Of the 389 tokens of *great majority* in the BNC, 97.4% are preceded by the word *the*, and 79.4% occur in the phrase *the great majority of* + NOUN). Another example is the collocational phrase *at a gallop*, which sometimes has an intervening adjective, and for which a more precise description would be *at a (ADJ.) gallop.* Similarly, the phrase *for gain* is indicated, but again there is often an intervening adjective (e.g. *financial*, *commercial*, *personal*, *private*). An example with a verb involves the third sub-entry for **gallop** (= increase). Here, we see the structure VERB + *gather* realized as *begin to gather*, *start to gather*, followed by one example, *As the weeks passed, Charlotte began to gather strength*. The example is useful, but the reader should be told which other nouns typically follow this VERB + VERB pattern.

In a sense, many of the extended collocational units of the type exemplified are in the dictionary already, but they are not explicitly shown. The reader has either to put together different collocates within the same entry, or to discover the relevant phrases within examples. It would be useful to find at the end of some entries, and not associated with a particular collocate, contextualized examples containing extended collocational units, with the important words in bold type.

Miscellaneous phraseological imprecision

I will complete this overview of shortcomings with varied examples of imprecise phraseological presentation. I begin with two examples involving the use of prepositions. Firstly, at the entry for **gaze** verb, there are three prepositions indicated (*at*, *into* and *in*). Of these, *in* is misleadingly represented: corpus evidence shows that in the vast majority of cases, '*gaze* + *in*' is part of the template *gaze* + *in* [*wonder*, etc] (which also appears in the one OCD example). It would be more appropriate therefore to consider the collocate as *in wonder* (etc), an adverbial collocate to be placed in the same section as collocational phrases such as *gaze admiringly*. Secondly, at **majority 2** (= in an election) there is an indication of the prepositional collocate *in*, with the example *a majority in parliament*. It is true that *in* often occurs after *majority*, but I think that *in* should be associated with the following noun (in this case *parliament*), and the whole phrase *in parliament* be considered as the collocate.

A different example of imprecision is at the entry for **gait** *noun* (PREPOSITION section), where we find the phrase *with a* ~ (where '~' stands for *gait*). However, **gait** is a word of general meaning, which usually needs to be completed in some way (there are no corpus examples of *with a gait*). The entry would therefore be better as *with a* ADJ *gait*. A final example is at **gas** [heating / cooking], where the verbs *cook with*, *light* and *turn on* are listed, but there is no explicit indication that the first of the three is used without *the* and the other two with *the*.

Further suggestions for the print or electronic dictionary

The suggestions made in this section are not to be thought of as automatically applicable to the print dictionary. The OCD presents collocational information in a quite simple fashion, and the addition of further data types may result in the dictionary seeming more cluttered. The actual amount of space available may also be a problem, and this is why I have also made reference to the electronic medium.

Structure of entries: semantic labelling

It may be useful, especially in the case of longer, or more dense, entries, to label groups of collocates pertaining to the same semantic field. A case in hand is the entry for the headword *magazine*. Within the adjective subsection (including nouns functioning as adjectives), there are 41 collocates. Some possible semantic labels for the sub-groups are the following: PHYSICAL APPEARANCE (e.g. a colour magazine, a glossy magazine); FREQUENCY (e.g. a monthly / quarterly / weekly magazine); GENERAL DISTRIBUTION (e.g. a local / national magazine); GENERAL READERSHIP (a teenage magazine / a women's magazine); MORE SPECIFIC READERSHIP & DISTRIBUTION (e.g. a school / student / parish magazine); CONTENTS (e.g. a computer / fashion / gardening / satirical magazine).

Text type

Where there are very close associations between collocational patterning and text type, it may be useful to indicate these. An example from the sample pages of OCD is the adverb *unseeingly*, which is indicated as a collocate of the verb TO GAZE. A check in the corpus reveals that the 8 tokens of *unseeingly* which accompany the verb *to gaze* are all found in the domain of 'imaginative literature'. Furthermore, of the total of 41 tokens of the word *unseeingly*, (which mainly collocates with the verb *to stare*), only one does not appear in imaginative literature.

Relative frequency of collocates

It would sometimes be useful to give indications of relative frequency of alternative collocates. This would help the learner choose between otherwise similar lexical items, and also allow the compiler to include somewhat unusual, but perfectly acceptable, collocational phrases, which otherwise might be omitted. This would be of use especially to the more advanced learner. (Compare remarks by Heid 2004: 735).

Concluding remarks

To conclude, analysis suggests that, although OCD offers a great deal of reliable collocational information, there are a number of ways in which it could be further refined. The lexicographical-pedagogical work undertaken after initial corpus analysis is a very crucial phase in this respect. Very great attention must be paid to the accuracy of phraseological description and to the type of look-ups that will probably be made by learners. The number of contextualized examples should also be as great as space will allow.

References

Benson, M., Benson, E. and **Ilson, R**. 1997. *The BBI Dictionary of English Word Combinations*. Amsterdam / Philadelphia: John Benjamins. [revised edition of Benson, M., Benson, E. and Ilson, R., 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*.]

Cobuild 5 = *Collins COBUILD Advanced Learner's English Dictionary*, 5th ed. Glasgow: HarperCollins.

Heid U. 2004. "On the presentation of collocations in monolingual dictionaries". In *Proceedings of the Eleventh EURALEX International Congress*, G. Williams and S. Vessier (eds). Lorient: Université de Bretagne-sud, 729-738.

Hill, J. and **Lewis, M. (eds.)** 1998. *The LTP Dictionary of Selected Collocations*. Hove (GB): Language Teaching Publications. [originally published as two separate works: a) Dzierżanowska, H. and Kozłowska, C.D. 1982, *Selected English Collocations.* Warsaw: PWN, and b) Kozłowska C.D, 1991, *English Adverbial Collocations*. Warsaw: PWN.]

Kjellmer G. 1994. *A Dictionary of English Collocations: Based on the Brown Corpus*. Oxford: Clarendon Press.

Oxford Collocations Dictionary for students of English. 2002. Oxford: Oxford University Press.

CORPORA AND STYLE SHEETS FOR CONTEXT-BASED MT AND LSP

Alejandro Curado Fuentes

Héctor Sánchez Santamaría

Patricia Edwards Rokowski

Mercedes Rico García[21]

Abstract

Corpora have key functions in MT (Machine Translation) developments. In CBMT (Content-Based Machine Translation), corpus size and genre integration are significant aspects. In this paper, we focus on an important stage for CBMT performance: the role of written academic linguistic forms, particular linguistic-discursive features or high-frequency lexical / rhetorical items from specific texts in the corpus. These items would integrate reference information for CBMT in relation to LSP (Languages for Specific Purposes). A reference academic corpus of Spanish writing is compared for the purposes of enhancing possible MT procedures. MT is supported by a massive target language corpus in Spanish, a huge bilingual, fully inflected, dictionary (English / Spanish), and a smaller source language corpus of English. Segmenting tools are used in CBMT for synonym finding utilities (called near-synonym finder) in case of insufficient lexical information. We find that the CBMT approach may be enhanced or at least partially aided in its implementation by the integration of frequency-based lexical-grammatical information in the form of collocations, colligations, semantic associations, textual collocations, and textual colligations from the academic Spanish corpus. This type of aid is described as complementary corpus-based style sheets for CBMT. The integration of the electronic hand-outs is done as a means of supervised documentation for the corpus management by the tool. Although at a very initial prototype stage, the tests with such items tell us about their feasibility and use in CBMT.

Keywords: Spanish corpus, academic lexical items, LSP, CBMT, style sheets

Introduction

MT (Machine Translation) of LSP (Languages for Specific Purposes) from English to Spanish may be exploited more effectively by integrating specific features of discourse, in this case, academic writing. In agreement with Storch and Tapper (1997), there seems to be a prioritization of grammar and lexis in the case of non-native writers of scientific / technical English. Spanish LSP readers / writers such as university faculty and students tend to rely on certain aspects of discourse more than on others, e.g., indirectness tends to be quite important in Spanish writing, and discourse markers, hedges, and paraphrasing play an important part (e.g., Lahuerta Martínez 2004; Morales et al. 2007; Cademártori et al. 2007).

This paper focuses on written scientific-technical discourse in Spanish, particularly linguistic-discursive features that derive from the use of higher frequency rates for specific lexical / rhetorical items. In this line of work, a comparison with a reference academic corpus of writing is made at one of the stages to be developed for CBMT (Context-Based Machine Translation), a research project undertaken by our group this year. This type of approach to MT emphasises the need for the integration of context as key for disambiguation (cf. Abir et al. 2002; Carbonell et al. 2006). In addition to a massive target language corpus (i.e., Spanish in our project), a huge bilingual, fully inflected, dictionary (English / Spanish), and a smaller source language corpus (i.e., English),

[21] The four authors are members of the research group "GexCALL" (Group in Extremadura for research on Computer-Assisted Language Learning). Alejandro Curado Fuentes, PhD, Lecturer, teaches EFL / ESP at Polytecnic School, University of Extremadura, Cáceres, Spain. His research mainly focuses on academic English corpus linguistics and Spanish students' use of IT and ESP. Héctor Sánchez Santamaría, PhD, Permanent Assisting Teacher, teaches Programming at Mérida Polytechnic, University of Extremadura. His research includes free source software and Adaptive Hypermedia. Patricia Edwards, PhD, Senior Lecturer, teaches Tourism English and discourse analysis at School of Business, Cáceres. Her research focuses on language learning, IT and ESP. Mercedes Rico, PhD, Lecturer, teaches EFL / ESP at Mérida Polytechnic. Her research focuses on IT, Adaptive Hypermedia, and ESP.

segmenting tools are used in the CBMT project for synonym finding utilities (called near-synonym finder) in case that the process might still not find a given lexical item.

We find that the CBMT approach may also be enhanced or at least partially aided in its implementation by the integration of frequency-based lexical-grammatical information in the form of collocations, colligations, semantic associations, textual collocations and textual colligations from the Spanish corpus. Scientific-technical English writing is taken as source text, to be translated into target Spanish text. The tool can then be provided with stylistic sheets that may offer alternatives for the translated output in order to improve the translation into scientific Spanish. For instance, features of discourse in English such as the predominance of the passive voice, or the use of plural pronouns, have less weight in Spanish, where the active voice may be preferred.

Thus, our lexical investigation is projected at the improvement of CBMT approaches in the specific case of scientific-technical / academic discourse. Potential “customers” for this application, if implemented successfully in one year or so (evaluation phase in our project), can include university faculty aiming to translate English publications into correct (or at least decent) Spanish academic writing.

Methodology

The CBMT approach requires a massive target language corpus, in our case, Spanish. Such a large corpus (at least 70 GB) does not exist commercially, and thus, must be built and indexed by our research group. As we are only at a roughly 20 percent level in the compilation process, we are still ‘crawling’ the web and various electronic source sites in Spanish. If implemented successfully, the linguistic-discursive information from the corpus should enable the system to draw enough contrastive data for best lexical-grammatical alternative retrieval. Figure 1 shows the overall layout for the CBMT tools and resources, as provided by Carbonell et al. (2006).

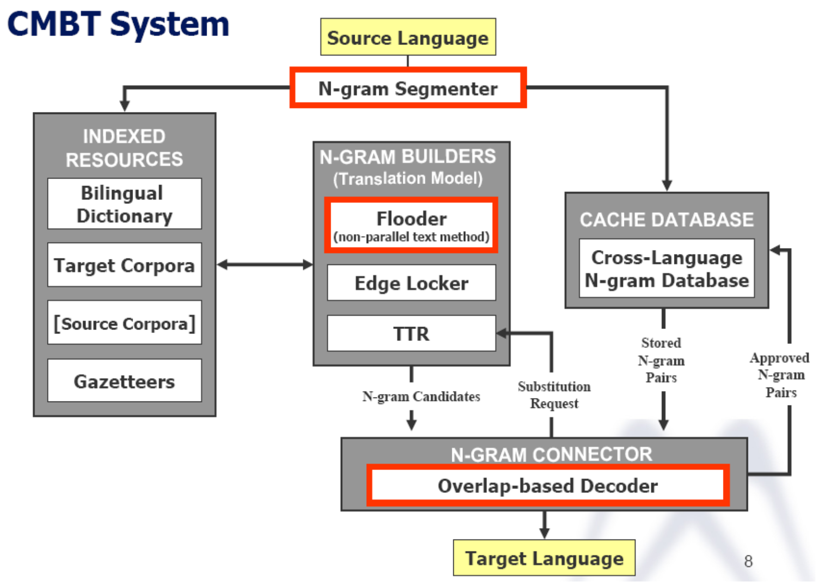


Figure 1: Distribution of resources and tools in the CBMT system

CBMT relies on a segmenting tool (n-gram builders) that takes n-grams and overlaps them in order to find suitable overlapping possibilities (e.g., *a soldier in the, soldier in the US, in the US army*, etc). A lot of text is thus crucial in the search for the candidates for translation, as the matching among the many possibilities is done according to word frequency and position in the n-grams. For this paper, we focus on the very early prototype texts that we have tested in order to draw the data for a very preliminary analysis.

The observation of function / grammatical words is key in the process. Lexical frequency is taken as a key measurement reference, and so, frequent grammatical words are selected. In specific academic texts and genres, the exploration of frequently co-occurring lexical items with these words leads to the observation of attributes of academic stance (cf. Biber et al. 2004). The academic register of scientific-technical discourse is checked by means of collocational strength degrees—i.e., in terms of content, grammatical-discursive position, semantic space, and text-item relationships (cf. Hoey 2005)—. Items are analysed in the form of collocations, colligations, semantic associations, textual collocations and textual colligations (cf. Hoey 2005).

Collocation refers to the statistically significant co-occurrence of two or more words in the texts. In Academic English, an example is the verb **appear** with **to be**, *appear** + *to be* (20.4 percent of co-occurrence probability according to measurements made in selected academic texts from the BNC Sampler—Burnard and McEnery 1999—, used as a source language sample). Colligation may involve frequency in the use of a given word class, such as nouns in academic discourse (e.g., *en el proceso de—in the process of—*followed by a noun that ends in the morpheme *ción—*see Table 1). Colligation may also include any given item where a grammatical aspect is related (e.g., *por qué no* (*why not*) is followed by the present simple tense in Academic Spanish, as derived from our analysis—Table 1—).

Semantic association refers to those collocations that have statistically significant semantic relationships. For instance, with the preposition **to**, the item *to be seeking* + 'work' is used characteristically in the BNC texts. Textual collocation refers to a pattern characterised by the positioning of a lexical item at a given point or place in the text. An example in the BNC Sampler is that *one of the most* + adjective appears at the beginning of sentences at a high proportion. Finally, textual colligations operate similarly to textual collocations, but in their case, a grammatical feature or pattern is involved. For example, in Academic Spanish, the pattern *teniendo en cuenta* (*bearing in mind*) is followed by a noun phrase without *that* at the beginning of sentences in 45 percent of the occurrences of *teniendo en cuenta* (see other examples in Table 1).[22]

Table 1 includes some examples withdrawn from the corpus and stored and managed in the MT system. These examples represent material that can be linked with one given word or item. For instance, the collocation *razón por la cual* (*reason for which*) can be provided in relation to the noun ***razón***.

Collocations	***Colligations***	***Semantic associations***	***Textual collocations***	***Textual colligations***
razón por la cual	¿por qué no + present simple	La mayor parte de + 'people'	. En la medida en que	, para lo que + reflexive verb
lo que se denomina	necesidad de + infinitive	'collected' + en la tabla #	. Por lo que respecta a	. Teniendo en cuenta + NP
hay que tener en cuenta que	en el proceso de + noun-ción	en el marco + 'institution'	. Contrariamente a lo que	. Desde el punto de vista + adjective

Table 1: Examples of lexical items collected from the corpus for the style sheet database

Results

The preliminary Academic Spanish corpus contains 154 Spanish texts that amount to a total of 688,481 running words. The STTR (Standardised Type-to-Token Ratio, i.e., lexical density) is 42.41 words, and the average number of words per sentence is 30.09. These two scores are similar to those calaculated for the academic English texts from the BNC Sampler: 41.54 and 26.23 respectively. Given these initial similarities, it seems that the

[22] The percentages for co-occurrence probability are calculated in relation to a content word (e.g., *appear to be* in relation to the inflected verb **appear***). These proportions can also be measured with reference to more than one single (non-content) word in order to specify discourse features (e.g., the colligation *I had* + participle is computed in relation to many instances of *I had* in the BNC Sampler).

academic register of the Spanish texts can parallel English discourse proportions of word use. It should be found, then, that the five types of lexical co-occurrence described above, abundant in specialised English discourse, are equally significant in Spanish academic texts. LSP discourse is, to a large extent, represented by such features, as being competent in a given academic discipline is closely related to having “mastery of collocations, colligations and semantic associations of the vocabulary (...) of the domain-specific and genre-specific primings” (Hoey 2005: 182).

Figure 2 displays the number of lexical features found with the first 25 grammatical words from the frequency word list. The results are classified according to either appearance of items across the whole corpus or distribution only within a certain genre.

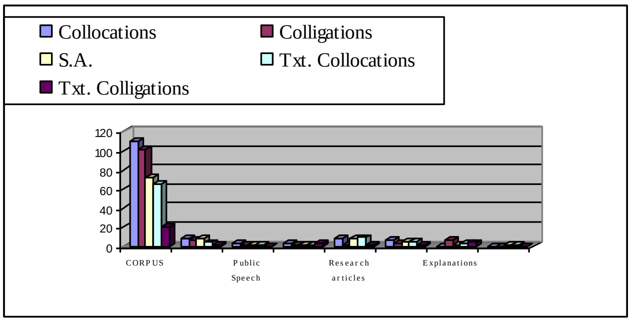


Figure 2: Number of items or features distributed according to whole corpus or genres

The items are selected and recorded on the basis of their stylistic features within the potential style sheets for the CBMT approach. As such, they should provide enriching alternatives in the English-to-Spanish translation. Figure 2 shows that there are more instances of collocations overall. These combinations are often easy to record and label in the style sheet for the provision of fitting translation options. For example, the construction *al mismo tiempo* (*at the same time*) is produced in 80 percent of the cases of *al mismo* (actually, a very close figure to *at the same time* in relation to *at* the same in the BNC sampler texts—79.1 percent—).

CBMT would work with the segmenting tool that takes n-grams and overlaps them in order to find suitable overlapping possibilities. As a result, collocations and other combinations can work as basic items in the style sheet on which this system may rely to find alternatives in the form of fabricated language. These items can complete given gaps in the CBMT segmentation analysis or when some lexical items may not be found in the bilingual dictionary. For instance, when the system encounters *al mismo*, the style sheet database should offer **tiempo** as highly likely to follow, given the statistical information stored. The collocation would then be selected, once the overlapping procedures should point to such appropriate alternatives.

Other statistically noteworthy examples from the corpus include several colligations where the adjacent item is grammatical: *al objeto de* + infinitive verb (86.6 percent) (*with the objective of*), *una vez* + past participle (73 percent) (*once*...), *si no se* + present simple indicative (86.6 percent) (*should he /it / she*...)—see others in Table 1 above—. In such cases, the information to be held should be labeled as grammatical features that must be invoked from the appropriate resources (e.g., if a specific verb tense is favoured, the system should check and select it from the lists of conjugations in the inflected dictionary).

In the case of the semantic associations, a similar process can take place with particular semantic sets previously arranged in the style sheet. For example, the form *se sitúa en* + ‘physical location’ appears in 42 percent of the instances of *situarse en* (*to be located in*). In this same scope, some items can be examined within particular genres, and this information can also be entered in the style sheet database. An example is *se encuentra en* + ‘virtual space’ (*it is found in*), only present in textbooks.

Finally, textual collocations and textual colligations operate a bit differently, as tagging their relevant information would involve potential changes in the arrangement of the translated sentences. For instance, the cluster *en este trabajo* (*in this paper*) should be re-positioned to the beginning of the sentence and followed by a comma. This rearrangement would be done after checking the information regarding this item in the style sheet, where it appears as a textual collocation with 72 percent of its occurrences at sentence initial position (i.e., . *En este trabajo*,). Similarly, the instances of textual colligation (see Table 1) can be stored as relevant linguistic-discursive information, where important lexico-grammatical aspects are intricately related, e.g., the form *por más que* (*no* *matter what*) should be placed at the beginning of sentences and followed by the present simple subjunctive (71 percent of the instances).

Conclusions

The integration of this material in CBMT may especially find an interested audience among Spanish readers of scientific written discourse who would like to access effective MT tools for English and Spanish. As teachers of English for Specific Purposes, we believe that rather than working with translation as an absolute concept, a more valuable information that can be given to faculty and students is in the form of MT improvements regarding thesaurus-like and / or stylistic / discursive handout-like resources for academic language effectiveness. The scope would entail a descriptive, and not prescriptive, approach to the material in a way that, for example, commonly used academic colligations may be offerred as better options to take in the processing of discourse (e.g., *hacer referencia a—make reference to—*, used in the present indicative in 72 percent of the instances of the verb, should make a better alternative in this verb tense in Academic Spanish).

Likewise, textual considerations may be made as regards genre and / or even subject traits, which serve to improve the academic register and tone of the translation. In English, for instance, sentences such as *There is no need for* or *There is no point in* may provide suitable alternatives in essays when ideas are introduced into a new paragraph (i.e., a textual collocation), according to the statistical information from the BNC Sampler. The same line of reasoning may be applied to Spanish discourse in translation: A likely candidate for the expressions above, for example, may be the textual colligation *No se trata de* + infinitive, used in 82 percent of the instances at the beginning of sentences in the same type of argumentative text. As Bhatia et al. (2004: 205) state, this increase in linguistic discursive awareness of basic generic principles and lexico-grammatical resources enables the strengthening of text processing and production competence. Such are also the tenets for our project, considered potential strengths for the CBMT system.

References

Abir, E., Klein, S., Miller, D. and Steinbaum, M. 2002. " Fluent Machines' EliMT system." *Association of Machine Translation in the Americas* 2002: 216-219.

Bhatia, V.K., Langton, N.M. and Long, J. 2004. "Legal discourse: Opportunities and threats for corpus linguistics." In *Discourse in the Professions: Perspectives from Corpus Linguistics*, U. Connor and T.A. Upton (eds.). Amsterdam: John Benjamins, 203-234.

Biber, D., Csomay, E., Jones, J. and Keck, C. 2004. "A corpus linguistic investigation of vocabulary-based discourse units in university registers." In *Applied Corpus Linguistics. A Multidimensional Perspective*, U. Connor and T.A. Upton (eds.). Amsterdam: Rodopi, 53-72.

Burnard, L. and McEnery, T. 1999. *The British National Corpus sampler.* Oxford: Oxford University Press.

Cademártori, Y., Parodi, G. and Venegas, R. 2007. "El discurso escrito y especializado: Las nominalizaciones en manuales técnicos." In *Lingüística de Corpus y Discursos Especializados: Puntos de Mira*, G. Parodi (ed.). Valparaíso, Chile: Ediciones Universitarias de Valparaíso, 79-96.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T. and Frey, J. 2006. "Context-Based Machine Translation." *Association of Machine Translation in the Americas* 2006: 19-28.

Hoey, M. 2005. *Lexical priming. A new theory of words and language*. London: Routledge.

Lahuerta Martínez, A.C. 2004. "Discourse markers in the expository writing of Spanish university students." *Ibérica* 8: 63-80.

Morales, O.A., Cassany, D. and González-Peña, C. 2007. "La atenuación en artículos de revisión odontológicos en español: Estudio exploratorio." *Ibérica* 14: 33-58.

Storch, N. and Tapper, J. 1997. "Student annotations: What NNS and NS university students say about their own writing." *Journal of Second Language Writing* 6: 245-264.

SAME SAME BUT DIFFERENT — A CORPUS-DRIVEN APPROACH TO THE STUDY OF PARAPHRASES

Pernilla Danielsson[23]

Abstract

This paper presents an ongoing corpus-driven project focusing on identifying naturally occurring paraphrases in text. As such it differs from most previous studies of paraphrases which have focused on the quality of a paraphrase and on how to use paraphrases in the classroom. Paraphrases hold an important clue to the creation of meaning and a better understanding of how a word or a phrase can be expressed in other words and still contain similar information would make a breakthrough for any automated system for language understanding. The paraphrases identified in this study are further categorised into the following three categories; (a) Alternative lexis, (b) Alternative phrases, and (c) Alternative grammar. The tentative results presented here suggest that the paraphrases are evenly distributed between the three categories.

Keywords: corpus, naturally occurring paraphrases, alignment

Introduction

This paper intends to take a small step towards a corpus-based study of paraphrases. A paraphrase expresses a statement, a phrase or a single word, in some other words. They are used to clarify, explain, describe, define, transfer and reformulate an expression and, as such, they are vital for exploring natural language semantics. Paraphrases hold important information about how meaning is created in texts; or as stated in Teubert (2003): "Meaning is paraphrase". Yet, corpus linguists have so far neglected to study how paraphrases are linked to meaning. One reason why there has been so little research in this area is due to difficulties in identifying paraphrases in texts. This paper will show the results from the ongoing corpus-based project studying paraphrases in authentic text.

Paraphrases have been used in the classroom as a tool to capture the meaning and to practise rephrasing of a word or a statement. Rewriting exercises, where students are asked to fill in according to a pattern of "*an X is a Y*", are examples of classical educational paraphrases. In fact, the COBUILD dictionaries definitions are based on this classroom tradition on how to explain the meaning of words and phrases. Paraphrases in authentic texts are different though. Whereas humans seem to have a constant need to rephrase, clarify or define ourselves, the structures of how to do it have not been studied. There are obvious reasons for this. Normally, a paraphrase is not (easily) automatically identifiable in texts, which makes the study of paraphrasing a very lengthy and time-consuming task. To get round this problem, we are creating and using a corpus consisting of several versions of each text, here referred to as the Transfer corpus.

By undertaking this work, we hope to answer the question '*which types of paraphrase can be found in authentic texts?*' The question is directly linked to the data-driven approach. Most linguistic studies of paraphrases have so far either been based on what is expected to be a good, or bad, paraphrase. In terms of the automatic identification of paraphrases, studies have concentrated on a syntactic focal point. This project proposes to start from a lexical focus. Some studies have also been restricted by limiting themselves to a view of paraphrase that necessarily involves word-to-word synonymy; in other words, the unit of paraphrase is defined as the word, such as *shrub* is a paraphrase of *bush* and *city* is a paraphrase of town (Barzilay & McKeown 2001).

In our approach, we analyze language practice itself, instead of making assumptions based on existing conceptualizations. Once the paraphrases have been identified, they will be classified according to their alternative roles; Alternative lexis, alternative phrases, or alternative grammar. Contrary to our own belief when the project

[23] Dr Pernilla Danielsson is the Academic Director of the Centre for Corpus Research at Birmingham University. After receiving her PhD as a computational linguist from Gothenburg University, Sweden, she has spent the last 8 years working with corpus linguistics in the UK. Her main research areas include: identifying multi-word units of meaning, studying naturally occurring paraphrases in texts, identifying units of translation in parallel corpora, and using corpora for literary studies. She is currently writing a book on 'Exploring a Language Corpus' together with Judith Lamie, to be published by Cambridge University Press in 2009. Previous publications include 'Meaningful Texts', edited together with Geoff Barnbrook and Michaela Mahlberg, Continuum 2005 in which she also has an article on extracting meaningful units from a Chinese English parallel corpus.

Formatada: Espaçamento entre linhas: simples begun , the three categories have a fairly even distribution among the paraphrases identified and categorised so far.

Naturally occurring paraphrases in text

> *Paraphrases hold important information about how meaning is created in texts, yet little work has been done exploiting this resource in linguistic. Instead, the most active area of paraphrase research is in Machine Translation. Being able to automatically identify and mark-up authentic paraphrases in text could provide a most useful addition to any translation aids. Much of the current research focuses on using paraphrases to evaluate MT output, see for example Zhou et al (2006) and Owczarak et al (2006).*

The emphasis in this study lies on identifying *naturally occurring* paraphrases, which should be distinguished from other types of paraphrases. This study does not primarily take an interest in the data found in thesauri, nor does it include the usual teaching of paraphrasing in language education. The main reason for this is that whereas most thesauri offer a list of near-synonyms, which tend to be single word items. In the case of paraphrases used in language learning, they tend to be more of a simple reforming of the same expression ('the pen is on the table', 'on the table is the pen) and situations are made up in which students have to paraphrase statements which do not necessarily form a naturally sounding unit. In contrast to such paraphrases, the naturally occurring paraphrases embrace a wider category; When a source is paraphrased, the wording is changed and the meaning is both 'the same' and 'different'.

Using language corpora to identify paraphrases

A useful starting-point for the study of paraphrase is a set of texts that can be identified as paraphrases of each other on external rather than internal criteria. In this project texts that have more than one translation into English are compiled into a corpus, the Transfer Corpus. Similar corpora used in paraphrase research have been reported by Barzilay & McKeown (2001) and Iordanskaja et al (1991).

Other types of corpora that have been used look at paraphrases in summarised texts, i.e. where the shortened phrase is aligned with its counterpart in the longer text (Zhou et al 2006). These types of paraphrases differ from the one in this study as they also intend to be short.

Yet another type of corpus that may be used consists of revised versions of a text. Research on this type of data has been carried out by Falvey (1993), Utka (2004) and John (2005), although they have not been focusing on paraphrases per se, but instead on the revisions. Revisions offer a slightly alternative view on paraphrases as they also involve an evaluative feature, where the latest revision is considered the best version.

A fourth type of corpus that may be used for paraphrasal studies are the more traditional parallel corpora, consisting of source texts aligned with their target texts, again these have been used in MT research (Callison-Burch et al 2006) by searching for one and the same phrase in one language and assuming that the corresponding segments in the other language are paraphrases.

So far in this project we have only been able to study the first mentioned type of corpora, the monolingual aligned parallel texts. Hence, only a very small number of texts can be included in this type of corpus as few texts are translated into English more than once. The texts are aligned at sentence level. The reason behind choosing such a corpus can be found in the old scholarly argument that translation should be viewed as transferring texts from one language by *paraphrasing* it into another. As such, the whole text can be seen as a paraphrase. However, in this study we will focus on differential paraphrases. By identifying the differences between the texts, we may find valuable information about alternative phrasings of one and the same source segment, in other words *paraphrases*.

> *Will he be able to name* ***any****?*
>
> *Will he be able to name* ***one****?*

As mentioned above, the main obstacle for compiling a language resource like this is that only very few texts are translated into English several times. This imposes heavy restrictions on the available resources. In this study the chosen text is Plato's *Republic*, which has been the object of numerous translations into English. The full project also includes other texts with multiple translations such as Selma Lagerlof's 'Gösta Berling's Saga' and Dante's 'The Divine Comedy'.

(a) And about knowledge and ignorance in general; **see** whether you think that any man who has knowledge ever would wish to have the choice of saying or doing more than another man who has knowledge. Would he not rather say or do the same as his like in the same case?

(b) In any branch of knowledge or ignorance, do you think that a knowledgeable person would intentionally try to outdo other knowledgeable people or say something better or different than they do, rather than doing or saying the very same thing as those like him?

A paraphrastic relationship between two sentences.

Identifying paraphrases in naturally occurring text

The methodology of the study replicates that used in most corpus studies of translation. Initially, a search is conducted on the verbs *see, saw* and *seen*. In order to be captured in the search, a sentence in one of the texts will include a form of one of these verbs; the equivalent sentences in the parallel texts may or may not include the same verb. Sentences that are not formally identical are then identified as in a paraphrastic relationship. The degree of difference may vary, as illustrated in the examples below.

(a) '***everyone saw*** *that…'*

(b) ***'it was clear to all*** *that…'*

Illustration of the paraphrases '*everyone saw'* and '*it was clear that'*.

The example above illustrates a well-known concept in corpus linguistics, that what is equivalent in translation is not the word but the phrase or sentence. The equivalence here is not between the two words *saw* and *clear* but between the two phrases. Similarly in the example below, the word *what* is found in the same position as the word *something* in the second text. This is not the type of information found in a thesaurus, and without the context the paraphrastic relation would be lost.

(a) *then I saw* ***what I had*** *never seen before*

(b) *then I saw* ***something I'd*** *never seen before*

Illustrations of the paraphrases '*what'* and '*something'.*

Having performed a closer investigation of the corpus data, we found that the phrases lend themselves to a categorisation based on their types of differences. Three categories dominate our results:

(1) Alternative lexis

(2) Alternative phrases

(3) Alternative grammar

However, these categories do tend to overlap, and one and the same segment may be marked up as belonging to several categories. The example below is an illustration of 'alternative lexis' because the word *see* is present in (a) but not in (b). It illustrates alternative grammar because (b) is in interrogative mood whereas (a) consists of a declarative followed by an interrogative. It is also an example of alternative phrases because the phrase *see whether you think* in (a) could be replaced by *do you think* in (b). In addition, *any man who has knowledge* in (a) could be replaced by *a knowledgeable person* in (b), and so on through the example. The equivalence is not between one word and another but between a phrase and a phrase.

> (a) And about knowledge and ignorance in general; **see** whether you think that any man who has knowledge ever would wish to have the choice of saying or doing more than another man who has knowledge. Would he not rather say or do the same as his like in the same case?
>
> (b) In any branch of knowledge or ignorance, do you think that a knowledgeable person would intentionally try to outdo other knowledgeable people or say something better or different than they do, rather than doing or saying the very same thing as those like him?

Example of segments with a paraphrastic relationship, especially concerning '*see'* and *'do you think'*.

Formatada: Esquerda, Espaço Antes: 0 pto

Formatada: Inglês (Reino Unido)

Alternative Lexis

This is the simplest form of paraphrase; a word is exchanged for another word or a combination of a few words. This is the category that resembles the information captured in a thesaurus. Taking the outset from a single word, *the look up word*, the thesaurus will offer a selection of other single words, or in a few cases, a short phrase as possible alternatives.

Examples below illustrate some of these 'Alternative Lexis' items in context, where the word *see* is paraphrased into *look at*.

Formatada: Cor do tipo de letra: Preto, Italiano (Itália)

(a) Let us rise soon after supper and see this festival; there will be a gathering of young men, and we will have a good talk.

(b) After dinner, we'll go out to ***look at*** it. We'll be joined there by many of the young men, and we'll talk

Formatada: Normal, Justificado, Avanço: Esquerda: 0,95 cm, Sem controlo de linhas isoladas, Não ajustar o espaço entre texto asiático e texto em latim, Não ajustar o espaço entre texto asiático e números

Formatada: Tipo de letra: 10,5 pt, Cor do tipo de letra: Preto, Italiano (Itália)

Formatada: Tipo de letra: 10,5 pt, Não Negrito, Cor do tipo de letra: Preto, Italiano (Itália)

Formatada: Tipo de letra: 10,5 pt, Cor do tipo de letra: Preto, Italiano (Itália)

Paraphrases in context; *see* and *look at*.

Clearly, this categorisation allows us only to address the surface level of what is actually going on in a paraphrased sentence as the one above. A more in-depth analysis of what has changed between the two sentences would need to include much more.

Verb form/paraphrases	possible paraphrase	possible paraphrase
See	look at	determine
Saw	caught sight of	looked upon
Seen	appear	occur
Hear	listen to	listen
heard	mentioned	listened

Table 1: Some findings belonging to the category Alternative Lexis

The table above is used as an illustration of the findings within this category.

Alternative Phrases

Formatada: Tipo de letra: Itálico, Inglês (Reino Unido)

The category of Alternative Phrases takes into account the fact that the words in our study *see* and *hear* are often part of large units of meaning, i.e. multi-word units, such as *let's hear it.* A larger unit with a clear semantic value may very well still be paraphrased with only one word, however, in our data we find that two larger units often paraphrase each other. The concept of replacement is important here. In an example of Alternative Lexis (e.g. example 9), a single word in one sentence may replace a multi-word unit in the aligned segment. In Alternative Phrases, to make sense of the sentences, more than one word would have to be involved in the replacement. Here are some examples of phrases which could replace each other:

~~*HEAR*~~	*when I hear you say that*	~~*even as you were speaking*~~
	let us hear	*Continue to explain*
HEARD	*You have often heard me say*	*you know very well that I am going to say this*
~~*SA*~~*W*~~*W*~~	*everyone saw that*	~~*it was clear to all that*~~
SEE	*But see the consequence*	*then, it follows…that*

Table 2. Examples of findings from the category Alternative Phrases

In example (a) below, *see* forms a meaningful unit with *but … the consequence*. Evidence for this as a unit comes from the paraphrase of the whole unit rather than of the individual item *see*: *it follows that* (b). As before, the paraphrases have other consequences. The example (a) is an imperative, implying effort, whereas (b) construes a natural sequence of events.

(a) But see the consequence; Many a man who is ignorant of human nature has friends who are bad friends, and in that case he ought to do harm to them; and he has good enemies whom he ought to benefit; but, if so, we shall be saying the very opposite of that which we […]

(b) ***Then, it follows,*** Polemarchus***, that*** it is just for the many, who are mistaken in their judgment, to harm their friends, who are bad, and benefit their enemies, who are good.

Ex. Illustration of findings in the category Alternative Phrases.

Alternative Grammar

The category Alternative Grammar is used when the two sentences in the pair differ in terms of grammar. This includes realization or omission of optional elements, such as *that* in noun clauses or the object pronoun in verb phrases as illustrated below in the example.

he asked…

he asked him…

Optional object pronoun as paraphrase.

It also includes variation in tense, aspect, and voice as in:

SAW	*(and) we never saw (her)*	*(and) we didn't see (her)*
	what he saw before was an illusion	*what he'd seen before was incense…*
HEAR	*did you ever hear*	*have you ever heard*

Formatada: Normal, Avanço: Esquerda: 0 cm

Formatada: Tipo de letra: 10,5 pt, Inglês (Reino Unido)

Formatada: Tipo de letra: 10,5 pt, Italiano (Itália)

Formatada: Tipo de letra: 10,5 pt, Não Negrito, Não Itálico, Italiano (Itália)

Formatada: Tipo de letra: 10,5 pt, Não sublinhado, Italiano (Itália)

Formatada: Tipo de letra: 10,5 pt, Italiano (Itália)

	you ought to hear (them)	*(these things) must also be heard*

Table 3: Findings in the category Alternative Grammar.

Comparing the phrases above with their usage in a modern English Corpus may further this study. For example, the phrase '*did you hear that'* occurs in The Bank of English corpus (a 450 million corpus of present-day English) 52 times. Whereas the phrase '*have you ever heard*' occurs 197 times, almost four times as frequent.

The alternations between positive and negative statements is another phenomenon that we count as Alternative Grammar:

(a) ***Do you see that*** there is a way in which you could make them all yourself?

(b) ***Don't you see that*** there is a way in which you yourself could make all of them?

Ex. Paraphrase shift between positive and negative statement.

Conclusion

This study gives an indication of the types of paraphrases that can be identified in texts. Having classified the types of difference between paraphrases, and identified all the paraphrases in our corpus involving one of the words see *and* saw, *we are able to quantify the different types.*

There is a fairly even distribution between paraphrases in the three categories, which shows that traditional thesauri-type of information only suffice to offer a third of all the possibilities. This is perhaps not surprising, as it is known that good translators translate phrase-by-phrase or sentence-by-sentence rather than word-by-word. It also confirms findings in recent corpus linguistics that support the importance of phraseology, as opposed to separate concepts of lexis and grammar, to language. As differences between phrases might be said to include differences between lexis and grammar, our data shows how frequently these two aspects of language work together in expressing paraphrase. Our findings also suggest that our data is a good source for the identification and classification of paraphrases, with a substantial number of differences being identified of each type for each of the words investigated. With the methodology established and tested it now becomes possible to carry out more extensive investigations on a larger number of words using semi-automated techniques.

References

Barzilay, R & McKeown, K.R. *(2001) 'Extracting Paraphrase from a Parallel Corpus.' In: Proceeding of ACL 2001:50-57.*

Callison-Burch, C. Koehn, P. and Osborne, M*. (2006) Improved Statistical Machine Translation Using Paraphrases. In* Proceedings from the Human Language Technology Conference of the North American Chapter of the ACL, *pp. 17-24. New York: ACL.*

Danielsson, P. & Ridings, D. (1997) 'Practical Presentation of a "Vanilla" aligner.' In Reyle, U. and Rohrer, C. (eds.) Presented at the TELRI Workshop on Alignment and Exploitation of Texts. Institute Jozef Stefan. Ljubljana.

Falvey, P (1993) 'Towards a description of corporate text revision' PhD Thesis, University of Birmingham.

Iordanskaja, L., Kittredge, R. and Polguere, A. (1991) *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, chapter11. Amsterdam: Kluwer Academic Publisher.

John, S P (2005) The Writing Process and Writer Identity: investigating the influence of revision on textual and linguistic features of writer identity in dissertations. Unpublished PhD thesis, University of Birmingham, Birmingham, United Kingdom.

Owczarzak, K, Groves, D. Van Genabith, J. and Way, A. (2006) Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. IN *Proceedings of the Workshop on Statistical Machine Translation*, pp. 86-93. New York: ACL.

Teubert, W. (2003) 'Corpus Linguistics – A partisan view. Corpus Linguistics as a Theoretical Approach' In: International Journal of Corpus Linguistics. Vol5, No1.

Utka, A. 2004. English-Lithuanian Phases of Translation Corpus: Compilation and Analysis. In *International Journal of Corpus Linguistics 9:2: 195-224*.

Zhou, L. Lin Chin-Yew, Munteanu and Hovy, E. (2006). *ParaEval: Using Paraphrases to Evaluate Summaries Automatically. In Proceedings from the Human Language Technology Conference of the North American Chapter of the ACL, pp. 447-454. New York: ACL.*

Formatada: Tipo de letra: Negrito, Inglês (Reino Unido)

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

THE CORPUS DO PORTUGUÊS AND THE ROUTLEDGE FREQUENCY DICTIONARY OF PORTUGUESE: NEW TOOLS FOR LEARNERS AND TEACHERS

Mark Davies[24]

Ana Maria Raposo Preto-Bay[25]

Formatada: Português (Portugal)

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Abstract

In this paper we discuss two corpus-based resources of Portuguese that have recently become available, which hopefully are of real value to language learners. The first is the 45 million word Corpus do Português, which is freely available online and which offers many different types of searches that are oriented towards the language learner. These include searches by word, phrase, substring, lemma, and part of speech, as well as the ability to quickly and easily see the frequency across the major genres and time periods in the corpus. In addition, queries allow learners to compare the collocates of multiple words (to see the difference in meaning between the words), and between sections in the corpus (to see, for example, differences in word senses between genres). With one simple query, they can also see the frequency and distribution of all of the synonyms of a given word, and thus move far beyond a traditional thesaurus. The second tool is the recently-published Frequency Dictionary of Portuguese: Key Vocabulary for Learners (Routledge, 2007), which likewise has many features that are oriented towards language learners. Learners can easily find the most 5000 frequent lemmas in Portuguese, along with their English gloss, a sample sentence from the Corpus do Português, a translation of that sentence, and frequency and distributional information. In addition, there are "call-out boxes" with thematic vocabulary and frequency-based data on a wide range of grammatical phenomena that are often difficult for the Portuguese language learner.

Keywords: corpus-based, word frequency, dictionary, Portuguese, learners

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

The Corpus do Português

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

In terms of new resources available for learners of Portuguese, let us first turn to the *Corpus do Português*. This is a 45 million word corpus of Portuguese that was created by Mark Davies and Michael Ferreira (of Georgetown University) with generous funding from the United States National Endowment for the Humanities, and which was placed online in late 2006. The corpus contains 15 million words of historical Portuguese (1300s-1700s), 10 million words from the 1800s, and 20 million words from the 1900s. (It was the texts from the 1900s that served as the basis for the frequency dictionary.)

Of the 20 million words from the 1900s, two million words of text were taken from spoken Portuguese – either conversation (such as the *Linguagem Falada* project in Brazil or the *Projecto Corpus de Referência do Português Contemporâneo* from Portugal) or transcripts of interviews in newspapers and magazines. The written texts from the 1900s (18 million words) represent equally-sized sub-corpora from fiction, newspapers, and academic texts. In terms of the time period represented, virtually all of the texts from the 1900s are from 1970-2000, with the clear

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

[24] Mark Davies. Professor of Corpus Linguistics at Brigham Young University in Provo, Utah, USA. Co-creator of the Corpus do Português (www.corpusdoportugues. org), along with Michael Ferreira of Georgetown University. Co-author of the Frequency Dictionary of Portuguese: Core Vocabulary for Learners (Routledge, 2007) and the sole author of the Frequency Dictionary of Spanish: Core Vocabulary for Learners (Routledge, 2005). In addition to the Corpus do Português, the creator of several other large (100+ million word) online corpora, including the Corpus del Español, the Corpus of American English, BYU-BNC (the British National Corpus), and the TIME Corpus of Historical American English. Also the author of more than forty articles related to historical change and genre-based variation in syntax, as well as articles related to corpus design and construction.

Formatada: Justificado

[25] Ana Maria Raposo Preto-Bay. Assistant Professor in the Department of Spanish and Portuguese at Brigham Young University in Provo, Utah, USA. A native Portuguese, Ana graduated from the Universidade Clássica de Lisboa in English, German, and Modern Literature. She holds and MA in Linguistics from the University of Utah and a PhD in Instructional Psychology in the area of Language Acquisition from Brigham Young University. In addition to teaching Portuguese, she has taught in the Linguistics and English Departments where she worked in the areas of language acquisition and composition. In her research and published work she focuses primarily on issues dealing with literacy development— both in first and second language contexts—social and cognitive psychology, and foreign language teacher and program development.

Formatada: Tipo de letra: 9 pt, Inglês (Reino Unido) majority being from the 1990s. Finally, we should mention that the corpus for this century was evenly divided between texts from Portugal and Brazil, for each of the four classes of texts just mentioned.

In addition to serving as the basis for the frequency dictionary, the corpus is also freely available online (http://www.corpusdoportugues.org), and it contains a number of features that make it quite useful for language learners. In terms of basic features, learners can search for any word, phrase, or grammatical construction. In less that one second, they can see the frequency in both Portugal and Brazil, in each of the four major genres (spoken, fiction, newspapers and magazines, and academic texts), and they can also see (based on the historical data) whether the construction is increasing or decreasing in usage. In addition to seeing the overall frequency of the word, phrase, substring, or grammatical construction (or any combination of these), they can also see the frequency and distribution of each matching form. For example, users could search for the lemma *cujo* 'whose' and see that it is clearly decreasing in usage and that it is the most frequent in the more formal registers (as with *whose* in English). Likewise, they could search for the passive ([ser] [vk*]: a form of *ser* 'to be' + a past participle), and in a matter of less than two seconds they would see that it is increasing in usage, and that it is the most frequent in the formal registers. Finally, it is possible to compare the relative frequency of essentially anything across different sections of the corpus. For example, users can find adjectives that are more common in fiction than in academic (*pensativo, lívido, abafado, aflito* 'thoughtful, livid, stuffy, afflicted') or vice versa (*norte-americano, nuclear, mundial, químico* 'North American, nuclear, worldwide, chemical'), or verbs that are more common in Brazil than Portugal (*retornar*, *descartar*, *ressaltar*, *brigar* 'to return, disagree, emphasize, fight') or vice-versa (*regressar*, *aperceber*, *calhar*, *sublinhar* 'to return, perceive, happen, emphasize').

Users can also easily obtain useful semantic information from the corpus, via collocates. At the most basis level, they can simply input a word (such as *espesso* 'thick'), click on 'Context', and then see the most frequent collocates (sorted by Mutual Information score, if desired). For example, in the case of *espesso* they would find *pelagem, fumo, cauda, névoa*, and *treva* 'hair, smoke, tail, mist, darkness'). In the search form they can also select one section of the corpus (e.g. Portugal or Brazil, or any of the four main genres, or any historical periods) and compare the collocates, which provide valuable insight into differences in meaning and usage between these sections. For example, by selecting 'Fiction' vs. 'Academic', learners can see that the collocates for forms of *duro* 'hard' are somewhat more figurative in fiction (*olhos*, *palavras*, *trabalho* 'eyes, words, work') while in academic they are more literal (*rochas*, *materiais*, *fibras* 'rocks, materials, fibers'). Finally, since users can compare collocates across historical periods, they can see how the usage is changing over time. For example, they can easily compare the collocates of *mulher* 'woman' in the 1800s and 1900s, and see that in the 1800s the emphasis was often on 'moral qualities' (*desgraçada, indigna, divina* 'fallen, unworthy, divine'), while in the 1900s the collocates mainly refer to a fairly prosaic social category (*jovem, sozinha*, negra 'young, single, black').

Two other features of the corpus are specifically oriented towards language learners. First, the interface allows users to input two words and compare the collocates of the words. This information provides valuable insight into the difference in meaning between the two words – probably much more than could be obtained from a dictionary. For example, users can compare the collocates of *romper* and *quebrar*, which could be difficult for native English speakers to differentiate, since they are both translated as 'to break'. The corpus shows that the most frequent collocates with *romper* but not *quebrar* are *marcha, grito, lábio, nuvem*, and *fogo* 'pace, shout, lip, cloud, and fire', while those with *quebrar* but not *romper* are *cabeça, perna, coisa, nariz*, and *monotonia* 'head, leg, thing, nose, and monotony'. Finally, users can use the corpus as a type of 'thesaurus on steroids' to compare the frequency, use, and distribution of the synonyms of a given word. For example, they simply enter '[=limpo]' to find twenty different synonyms of *limpo* 'clean (ADJ)', and they can see which are found more in formal or informal genres, in Brazil or in Portugal, and which are increasing or decreasing in usage over time. In summary, there are many types of queries that allow language learners to quickly and easily gain insight into usage in Portuguese, which would in most cases be far beyond their level of intuition in the second language, or beyond that of most language learning materials.

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

The Frequency Dictionary of Portuguese: Core Vocabulary for Learners

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

In addition to creating a corpus that was 'user-friendly' for learners and teachers of Portuguese, there was also a desire to create a published work that could be used as an integral part of Portuguese language classes. Recognizing the value of frequency-based materials for language learners, we decided to create a frequency dictionary of Portuguese.

There were a handful of printed frequency dictionaries of Portuguese (Brown 1951, Duncan 1972, Kelly 1970, Nascimento 1987, and Roche 1975), as well as two or three in electronic format on the web. Nevertheless, none of these was based on a large, balanced corpus of Portuguese (in other words, with texts from a number of different genres). Therefore, the goal was to create a dictionary that would contain the 5000 most frequent lemmas in Portuguese – based on the data from the Corpus do Português, and with a number of features that were specifically oriented towards the language learner. We worked on this dictionary in 2006 and 2007, and it was published in late 2007 as the *Frequency Dictionary of Portuguese: Core Vocabulary for Learners*, which was published by Routledge in late 2007.

Before discussing this particular corpus-based dictionary, however, we might first address the general question of the value of a frequency dictionary for language teachers and learners. Why not simply rely on the vocabulary lists in a course textbook? The short answer is that although a typical textbook provides some thematically-related vocabulary in each chapter (foods, illnesses, transportation, clothing, etc), there is almost never any indication of which of these words the student is most likely to encounter in actual conversation or texts. In fact, sometimes the words are so infrequent in actual texts that the student may never encounter them again in the "real world", outside of the test for that particular chapter.

While the situation for the classroom learner is sometimes difficult with regards to vocabulary acquisition, it can be equally as frustrating for independent learners. These individuals may pick up a work of fiction or a newspaper and begin to work through the text word for word, as they look up unfamiliar words in a dictionary. Yet there is often the uncomfortable suspicion on the part of such learners that their time could be maximized if they could simply begin with the most common words in Portuguese, and work progressively through the list.

The bottom line, then, is that frequency dictionaries can be a valuable tool for language teachers. It is often the case that students enter into an intermediate language course with deficiencies in terms of their vocabulary. In these cases, the teacher may often feel frustrated, because there doesn't seem to be any systematic way to bring less advanced students up to speed. With a frequency dictionary, however, the teacher could assign students to work through the list and fill in gaps in their vocabulary, and they would know that the students are using their time in the most effective way possible.

The Routledge Frequency Dictionary of Portuguese: Core Vocabulary for Learners (Davies and Preto-Bay, 2007) is designed to meet the needs of a wide range of language students and teachers. The main index contains the five thousand most common words in Portuguese, starting with such basic words as *o* and *de*, and quickly progressing through to more intermediate and advanced words. Because the dictionary is based on the actual frequency of words in a large 20 million word corpus (collection of texts) of many different types of Portuguese texts (fiction, non-fiction, and actual conversations), the user can feel comfortable that these are words that one is very likely to subsequently encounter in the "real world".

The following information is given for each of the 5000 entries in the dictionary:

rank frequency (1, 2, 3, ...), headword, part of speech, English equivalent, dialect, sample sentence, translation, range count, raw frequency total, indication of major register variation

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

As a concrete example, let us look at the entry for *bruxa* "witch":

4522 *bruxa* *nf* *witch*
A caça às bruxas é muitas vezes acompanhada de histeria – Witch hunts are often accompanied by hysteria
35 | 235 –ac

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

This entry shows that word number 4522 in the rank order list is [bruxa], which is a feminine noun [nf] that can be translated as "witch" in English. We then see an actual sentence or phrase from the corpus, which shows the word in context, as well as a translation of this sentence into English. The two following numbers show that the word occurs in 35 of the 100 equally-sized blocks from the corpus (i.e. the range count), and that this lemma occurs 235 times in the corpus. Finally, the notation [–ac] indicates that the word is much more common in the fiction register than would otherwise be expected. In summary, then, each of the 5000 entries provides the language learners with information about the frequency of the word, its meaning (via the glosses and the sample sentence), and some indication of the distribution of the word in the different genres.

One of the criticisms of frequency dictionaries is that they are just “lists of words”, and that there is no semantic grouping of any of the words. To address this criticism in part, we placed throughout the main frequency-based index are approximately thirty “call-out boxes”, which serve to display in one list a number of thematically-related words. These include lists of words related to the body, food, family, weather, professions, nationalities, colors, emotions, verbs of movement and communication, and several other semantic domains.

In addition to vocabulary that is tied to a particular semantic category, however, we also focused on several topics in Portuguese grammar that are often difficult for beginning and intermediate students. For example, there are lists that show the most common diminutives, superlatives, and derivational suffixes to form nouns, the most common verbs and adjectives that take the subjunctive, which verbs most often take the “reflexive marker” *se*, which verbs most often occur almost exclusively in the imperfect and preterit, and which adjectives occur almost exclusively with the two copular verbs *ser* and *estar* or the semi-copular *ficar*. Finally, there are even more advanced lists that compare the use of nouns, verbs, adjectives, and adverbs across registers, and show which words are used primarily in spoken, fiction, newspapers, or academic texts. Related to this is a list showing which are the most frequent words that have entered the language in the past 100-200 years.

Aside from the main frequency listing, there are also indexes that sort the entries by alphabetical order and part of speech. The alphabetical index can be of great value to students who for example want to look up a word from a short story or newspaper article, and see how common the word is in general. The part of speech indexes could be of benefit to students who want to focus selectively on verbs, nouns, or some other part of speech. Finally, there are a number of thematically-related lists and lists related to common grammatical problems for beginning and intermediate students, all of which should enhance the learning experience. The expectation, then, is that this frequency dictionary will significantly maximize the efforts of a wide range of students and teachers who are involved in the acquisition and teaching of Portuguese vocabulary.

In summary, all of these features of the corpus-based frequency dictionary, as well as the *Corpus do Português* itself, represent linguistic tools that can greatly facilitate the learning of Portuguese by speakers of other languages.

References

Brown, Charles Barrett. 1951. *Brazilian Portuguese idiom list, selected on the basis of range and frequency of occurrence*. Nashville: Vanderbilt University Press

Davies, M., and A. Raposo Preto-Bay. 2007. *A Frequency Dictionary of Portuguese : Core Vocabulary for Learners*. London : Routledge.

Duncan, J.C. 1972. *A Frequency Dictionary of Portuguese Words*. Unpublished dissertation. Stanford University.

Kelly, J.R. 1970. “A computational frequency and range list of five hundred Brazilian Portuguese words”. *Luso-Brazilian Review* 7:104-13

Nascimento, M. Bacelar do, et al. 1987. *Portugues Fundamental. Metodos e Documentos*. Lisbon, INIC.

Roche, Jean. (1975) *Sobre o vocabulário da poesia portuguesa*. Paris : Fundação Calouste Gulbenkian

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Inglês (Reino Unido)

Formatada: Tipo de letra: Negrito, Inglês (Reino Unido)

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

CORPUS ANALYSIS IN AN ESP COURSE FOR INTERNATIONAL RELATIONS AND DIPLOMATIC STUDIES STUDENTS[26]

Silvia de Candia[27]

Giulia Riccio[28]

Abstract

Helping students identify significant features of those discourse types that are relevant to their studies can be considered one of the main goals in language teaching, especially when the students' main field of study is other than languages. Believing that corpus analysis can be very helpful in such a context, the authors carried out the seminar discussed in this paper, within the framework of an ESP course attended by students of International Relations and Diplomatic Studies.

The methodology adopted is based on the Corpus-Assisted Discourse Studies (CADS) approach, as outlined by Partington (2004). The seminar was aimed at enhancing the students' awareness of the main features of those discourse types that are more relevant to diplomatic studies – e.g. United Nations Security Council Resolutions, US State Department Reports – through the creation of ad hoc *small corpora and their analysis in terms of specific linguistic features and links existing between discourse strategies and socio-political issues. The seminar included a theoretical introduction to Corpus Linguistics, a practical introduction to the use of AntConc, and a final paper produced by the students as the outcome of their analysis of the corpora that they had autonomously created.*

Following Gavioli's claim (2005: 31-32), the main goal of this teaching experience was to enable the students to identify the main features of specific discourse types, rather than to focus on their findings, which did not necessarily have to be "scientifically interesting" (2005: 32). This study presents an evaluation of the whole experience, including an assessment of the students' autonomous corpus analysis.

Keywords: ELT, ESP, CADS, diplomatic discourse, methodology acquisition

Introduction

While the use of Corpus Linguistics in the teaching of English for Special Purposes (ESP) was generally limited, until few years ago, to syllabus design and the selection of didactic materials, nowadays this methodology is more frequently taught to be directly exploited by students in order to enhance their knowledge of specific discourse

[26] The authors discussed and conceived the article together. More specifically, Silvia de Candia is responsible for sections 2 (The seminar: aim and structure), 5 (Autonomous corpus analysis) and 6 (Assessment of the teaching/learning experience), Giulia Riccio for sections 1 (Introduction), 3 (Materials and methodology) and 4 (Reading meaningful data). The authors wish to acknowledge the professional and personal support of Professor Vanda Polese for organising, coordinating and providing invaluable help and advice in all the stages of the seminar described and in drafting the final paper. The authors also wish to thank Dr. Marco Venuti for providing advice and encouragement during the whole experience, and for revising the paper.

[27] Silvia de Candia is a PhD student in English for Special Purposes at the Linguistic Section of the Department of Statistics, Faculty of Political Sciences of the University of Naples Federico II, Italy. Her PhD research aims at analysing the discourse of broadcast, with special regard to the language of British and Italian TV news programmes. Since 2007 she has been working on the InTune project – an integrated research coordinated by the University of Siena, Italy – which investigates the identity, representation and standards of good governance in the European citizenship through the analysis of a corpus of media texts collected in 2006-2007. The research deals with English, French, Italian and Polish languages. Her focus is on the British and Italian TV sub-corpora of the IntUne corpus, which consist of TV news programmes collected from February to April 2007.

[28] Giulia Riccio is a PhD student in English for Special Purposes at the Linguistic Section of the Department of Statistics, Faculty of Political Sciences of the University of Naples Federico II, Italy. Her PhD research project focuses on the CADS analysis of discourse strategies enacted by the George W. Bush Administration, as realised in the White House press briefings. She is the author of a chapter (White House press briefings as a message to the world) in the book Wordings of War: Corpus-assisted discourse studies on the Iraq conflict, edited by Paul Bayley and John Morley and published by Routledge (forthcoming) and co-author, with Marco Venuti, of a chapter (Discovering patterns in the discourse of foreign relations. Corpus analysis in an ESP course for International Relations students) in the book Using Corpora to Learn About Language and Discourse, edited by Linda Lombardo and published by Peter Lang (forthcoming).

Formatada: Tipo de letra: Negrito, Maiúsculas

Formatada: Tipo de letra predefinido do parágrafo, Tipo de letra: Negrito, Superior à linha, Maiúsculas

Formatada: Tipo de letra: Negrito, Maiúsculas

Formatada: Inglês (Reino Unido)

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: simples

Formatada: Justificado, Espaçamento entre linhas: simples types (Brodine 2001; Gavioli 2005). Starting from the assumption that this approach is particularly helpful if integrated in a wider ESP learning context, the authors discuss a corpus-based teaching experience held within the framework of an English Language course addressed to students of International Relations and Diplomatic Studies. The present paper thus attempts to demonstrate how the application of Corpus Linguistics methodology and, in particular, of the Corpus-Assisted Discourse Studies (*CADS*) approach can yield positive results in such a teaching context.

This paper is structured into five sections. Firstly, the context in which the seminar took place will be briefly outlined; secondly, the theoretical background will be introduced. Subsequently, the way the students were guided to the interpretation of corpus data will be described, followed by an overview of their reports about autonomous corpus analysis. Finally, an assessment of the whole learning/teaching experience will be provided, together with some concluding remarks.

The seminar[29]: aim and structure

The seminar was carried out within an English Language course whose teaching approach combined Discourse Analysis and Corpus Linguistics. The attending students were in their first and second year of the second level degree course (9 and 6 credits respectively) in International Relations and Diplomatic Studies at the Faculty of Political Science of the University of Naples Federico II in the academic year 2007/08. The English Language course attended by the students mainly focused on political and diplomatic discourse. The aim of the course was to explore language behaviour and discourse strategies in such texts as international treaties, declarations and conventions, fact sheets and reports released by Foreign Ministries, Security Council Resolutions and other documents issued by the United Nations.

The seminar was mainly aimed at:

- making students aware of the main aspects of Corpus Linguistics theory and methodology;
- enabling them to use corpus processing tools for their own studies;
- teaching them the methodology of observation and interpretation of corpus data;
- extending their knowledge of the main features of diplomatic discourse, by focusing in particular on the relationship between typical patterns, meanings and strategies, rather than merely on linguistic features.

A total of twelve students attended the seminar: eight were in their first year of the degree course and four in their second year. The two groups had already passed one and two English Language exams respectively. Furthermore, the latter group had already taken a seminar on Corpus Linguistics in the previous academic year and had already experienced and applied corpus analysis methodology (see Riccio and Venuti forthcoming). Due to the small number of students it was possible to provide them with all the assistance and directions required in such activities.

The seminar was held in the computer laboratory of the Department of Statistics at the University of Naples Federico II. It consisted of 14 hours, divided into seven two-hour sessions, and it was structured into three steps:

- a theoretical background session about Corpus Linguistics methodology and corpus processing tools;
- four practical sessions in which the students were guided through the collection and analysis of corpora of political and diplomatic texts;
- two final sessions during which the students presented and discussed preliminary findings about the corpora they had assembled, in order to receive advice and feedback about their autonomous work.

Materials and methodology

The methodology adopted in this seminar was mainly based on the Corpus-Assisted Discourse Studies (henceforth *CADS*) approach. As outlined by Partington (2004; forthcoming), this approach is aimed at exploring specific

[29] This article includes a report on the seminar activities carried out by Silvia de Candia and Giulia Riccio within the English Language course coordinated and taught by Professor Vanda Polese.

Formatada: Tipo de letra: Itálico

Formatada: Tipo de letra: Itálico

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra predefinido do parágrafo, Tipo de letra: Itálico, Superior à linha

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Avanço: Esquerda: 0 cm, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Avanço: Esquerda: 0 cm, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

discourse types through quantitative and qualitative analytical techniques, in order to uncover meanings non-obvious to the naked eye. This is achieved by means of investigating the relationship between the use of specific discourse features and the strategies enacted by the producer of the text. *CADS* analysts thus assemble specialised corpora in line with their research purposes by focusing on discourse types with which they are already familiar. The *CADS* approach was adopted in this teaching experience because of its relevance to the purposes of the seminar – and of the course as a whole – which, as mentioned above, included enhancing the students' awareness of the main features of specific discourse types, through the creation and analysis of *ad hoc* small specialised corpora.

The corpus processing software that was selected for use in this seminar is AntConc, developed at Waseda University by Laurence Anthony[30]. AntConc is a cross-platform, freeware programme, which was originally designed for didactic purposes. It includes the main tools to carry out corpus analysis: a concordancer, a word list and a keyword list generator, as well as cluster and collocate producers.

AntConc was chosen first of all because it is freely downloadable and available for different operating systems, which allowed the students to install it on their own computers at home, no matter the type of machine with which they worked. Secondly, this programme was selected because of its user-friendliness and intuitive interface – two fundamental aspects in a learning activity of this kind.

Reading meaningful data

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

After the initial session, which – mainly following Baker (2006) – introduced the basics of Corpus Linguistics theory and practice, during the four sessions that constituted the core of the seminar the students were guided to develop the ability to extract meaningful and relevant information from corpus data and to interpret it correctly. As Gavioli (2005: 71) points out, what is at stake in such contexts is "enabling students to ask appropriate questions and to 'read' and interpret the data to get sensible answers", rather than teaching them how to use the software.

Three issues, which are emphasised by Gavioli (2005: 71) and are also relevant to the *CADS* approach, represented the main points of concern at this stage. First, the students needed to acknowledge that collecting and analysing authentic unedited material could allow them to take a different learning perspective. Secondly, they had to be guided through the selection of interesting corpus data, so as to narrow down the range of features to be examined. Last but not least, they had to bear in mind that they were dealing with specific discourse types, and that corpus data should not be analysed out of context.

The aforementioned issues were tackled during these hands-on sessions by assisting the students in their exploration of various specialised corpora with the AntConc tools. They started practising the tools and gradually familiarised themselves with word lists, concordances, clusters and collocates. They were provided access to two specialised corpora: the 2001-05 White House press briefings corpus[31] and a very small corpus containing all the UN Security Council Resolutions regarding Iran and North Korea dating back to 2006-07.

The students explored the two corpora as though they were in the middle of an "unknown land", as Bernardini puts it (2002: 166). One example will be provided here to illustrate how the students gradually achieved awareness of what to look for in a corpus. By looking at the word list of the Iran and North Korea Resolutions corpus, the students were required to identify those frequent lexical items that were not obviously related to the corpus' topic. Therefore, rather than focusing on such items as *Iran* (97 occurrences), *nuclear* (82 occurrences) and *resolution* (74 occurrences), the students were guided to identify words like *measures* (42 occurrences) and *activities* (35 occurrences) as potentially interesting lexical items for the analysis of the discourse strategies adopted in these Resolutions. A brief analysis of the concordance, clusters and collocates of both terms revealed that they had a very specific connotation in the texts. *Activities*, indeed, was found in such phrases as *proliferation sensitive nuclear activities* (10 occurrences), *enrichment-related and reprocessing activities* (6 occurrences) and *heavy water-related activities* (4 occurrences), all of which related to nuclear proliferation, and most frequently co-occurred with *suspend* and *suspension*. *Measures*, on the other hand, was found in such phrases as *appropriate*

[30] http://www.antlab.sci.waseda.ac.jp/software.html (last accessed 15 May 2008)

[31] This corpus, which had been previously assembled by Giulia Riccio for research purposes, was selected on the grounds of the specificity of the topics covered, dealing with the US domestic and foreign policy.

measures, *additional measures* (5 occurrences each), *necessary measures* (3 occurrences) and *transparency measures* (2 occurrences) and mainly co-occurred with *imposed* (14 occurrences). It was thus observed that, in these resolutions, the word *activities* refers to something which the Security Council regards as potentially dangerous and therefore believes should be blocked, while *measures* refers to sanctions that are likely to be inflicted if such activities are not stopped. This brief analysis enabled the students to understand how any single lexical item actually plays a specific role in the texts examined.

Once the students had become familiar enough with the AntConc tools, they were asked to collect *ad hoc* small corpora (about 50,000 words) of diplomatic texts. They were subsequently asked to carry out an analysis on the assembled corpora and to report the results in written papers, in order to discuss them during their final examination. Each student had to work autonomously but could ask for feedback and advice during the two final sessions of the seminar. The second-year students, who had already been introduced to Corpus Linguistics during the previous academic year, were also invited to take a step further and carry out a comparative analysis. It was suggested that they could build a corpus including two subcorpora and compare them, possibly also running a keywords comparison through the AntConc tool (see Baker 2006: 121-149).

Autonomous corpus analysis

Ten out of the twelve students who attended the seminar signed for the exam in January and February 2008. Each of them wrote a report about the corpus analysis they had autonomously carried out.

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

The students were recommended to design homogeneous specialised corpora in terms of discourse type, time span and/or topic and/or text source (e.g. US State Department). As a whole, this prerequisite was satisfied by all the students. The size of their corpora ranged from about 30,000 to 135,000 tokens. Four students focused on United Nations Security Council Resolutions – covering different time spans and topics – while two of them analysed reports of the UN Secretary General on specific issues, and one student dealt with the UN General Assembly Resolutions on Palestine. The three remaining corpora were made up respectively of: US Department of State Country Reports; official English translations of speeches and press conferences delivered by the then Russian President Vladimir Putin; declarations and conventions on Human Rights.

One of the aspects that emerged from the papers was that the students preferred to use data from frequency lists, collocates and clusters to carry out their investigations. Their observations were generally supported by giving examples from single texts rather than by showing concordance lines. Although most of the students did use concordances in their analysis, they probably found it easier to report collocate and cluster data in the papers, due to their better readability.

As regards the analysis of lexicogrammatical features, most students decided to start from the exploration of different modal auxiliary verbs – a feature that they had already encountered during their English Language courses. The students observed that *shall* was the most frequent modal auxiliary verb in the UN Security Council Resolutions and in Human Rights conventions, but with different nuances of meaning. While in the Resolutions it was found to have a prescriptive value and to state rules and decisions to be respected, in the Human Rights conventions it was shown to co-occur with phrases referring to human beings and to express rights to which they are entitled. Other students found *will* and *should* to be more typical of UN General Assembly Resolutions and UN Secretary General reports. While *will* was found to express intention and commitment on the part of the text producer, *should* was found to be used for giving advice and recommendations. What emerged was that the prevalence of different modal auxiliary verbs in different types of diplomatic texts reflects the different nature and legal status of each of them. Through this analysis, the students proved to be able to investigate the use of specific linguistic features and make claims about their functions and their relationship with the discourse type in which they occurred.

It is beyond the scope of this paper to describe in detail the findings obtained by each student. However, it is worth mentioning that, by analysing their corpora, most of them were able to identify non-obvious meanings attached to some specific lexical items in each discourse type. For example, the use of the item *groups* in Secretary General reports about children and armed conflicts – which was one of the most frequent lexical words in the corpus – was shown to specifically refer to local armies (*armed groups*: 144 occurrences; *militia groups*: 35 occurrences) and thus to be negatively connoted. Another student discussed phraseology in the UN General Assembly Resolutions about Palestine, in order to detect the ways in which “language is *conventionally* used” in those particular texts (Gavioli 2005: 33). She observed that the repeated use in the corpus of the phrase *to express grave concern* regarding the Palestinian conflict indicated that the General Assembly, due to the nature of its Resolutions, could only take a stance on the issue rather than state a decision. In the same way, she observed that the phrase *peaceful settlement* was regularly used to outline the goal that needed to be achieved.

As regards grammatical items, some students discovered interesting patterns of pronoun usage in their corpora. While those who investigated the discourse of UN Resolutions found that the study of pronouns was not relevant to their analysis, those examining other types of diplomatic texts focused on the use of masculine and feminine third person pronouns and on what is suggested by the choice of using either of them. In particular, one student observed that, in most declarations and conventions on Human Rights, *he* and *his* are used to refer both to men and women, with the only exception of the EU Charter, where the formula *he or she* is always used, and of the Arab and Islamic charters, where female pronouns and possessives are used when referring to specific rights of women.

As a whole, these papers proved that the students succeeded in identifying salient linguistic features in the investigated discourse types, although their findings were not necessarily interesting from a scientific point of view (Gavioli 2005: 32).

Assessment of the teaching/learning experience

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

The overall assessment of the seminar is rather satisfactory. As mentioned above, the students' papers reported a number of meaningful observations about the discourse types investigated, which are likely to be useful for the students in view of their future jobs or studies. However, as suggested by Gavioli (2005: 31-32), such a teaching experience should be assessed by focusing more on the learning process than on the product of the analysis. In this perspective, the ability to build a corpus and to use AntConc was successfully achieved by all the students, although competence in actual corpus analysis was developed differently by each of them.

However, it needs to be pointed out that not all the students were evenly motivated and attendance was not stable. Moreover, the limited time available represented a point of concern. A higher number of sessions would have helped the students reinforce their analytical skills by focusing on less straightforward aspects of the discourse types. Furthermore, problematic areas might have been more easily identified in order to provide the students with further guidance and advice. As a result, the students would have been enabled to focus more on the methodology than on the technical aspects of the work – an inclination that was indeed observed, despite strong emphasis placed on methodology acquisition since the seminar's first session.

When, at the end of their oral examination, the students were asked to provide feedback about their learning experience, they acknowledged that they had benefited from applying Corpus Linguistics to their own studies. In particular, they thought that the analysis had enabled them to uncover discourse features of which they were previously unaware.

From a teaching perspective, the experience was motivating despite the difficulties mentioned above. More specifically, the seminar has shown that it is indeed possible to apply corpus analysis and the *CADS* approach to enhance the students' awareness of specialised discourse types, even when their main field of study is other than languages.

Conclusions

The present paper has attempted to discuss the feasibility of a corpus-based ESP teaching experience with students of International Relations and Diplomatic Studies. The authors believe that the application of *CADS* in such contexts can be fruitful only if the students have sufficient linguistic background to be able to carry out at least essential metalinguistic analysis and to concentrate on methodological aspects. Therefore, students who have developed language awareness are more likely to be highly motivated, involved and 'interested', and to formulate significant observations about lexico-grammatical features of specific discourse types.

Provided that the students have an adequate linguistic background, it would be desirable for the *CADS* approach to be more and more exploited for ESP courses in the future. As emphasised by Johns (1994: 299), indeed, using corpora in the classroom can be fruitful for both teachers and students, since, as a rule, "things unobserved and unsuspected" are discovered during such activities.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Bernardini, S. 2002. "Exploring New Directions for Discovery Learning." In *Teaching and Learning by Doing Corpus Analysis*, B. Kettemann, G. Marko (eds). Amsterdam: Rodopi, 165-182.

Brodine, R. 2001. "Integrating Corpus Work into an Academic Reading Course." In *Learning with Corpora*, G. Aston (ed.). Bologna: CLUEB, 138-176.

Gavioli, L. 2005. *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.

Johns, T. 1994. "From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning." In *Perspectives on Pedagogical Grammar*, T. Odlin (ed.). Cambridge: Cambridge University Press, 293-313.

Partington, A. 2004. "Corpora and Discourse, a Most Congruous Beast." In *Corpora and Discourse*, L. Haarman, J. Morley, A. Partington (eds). Bern: Peter Lang, 11-19.

Partington, A. Forthcoming. "The Armchair and the Machine: Corpus-Assisted Discourse Studies." In *Corpora for University Language Teachers*, C. Taylor Torsello, K. Ackerley, E. Castello (eds). Bern: Peter Lang.

Riccio, G. and Venuti, M. Forthcoming. "Discovering patterns in the discourse of foreign relations. Corpus analysis in an ESP course for International Relations students." In *Using Corpora to Learn About Language and Discourse*, L. Lombardo (ed.). Bern: Peter Lang.

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

POSITIVE AND NEGATIVE EVALUATION IN NATIVE AND LEARNER SPEECH

Sylvie De Cock[32]

Formatada: Tipo de letra: 10,5 pt, Negrito, Inglês (Reino Unido)

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: 10,5 pt, Negrito, Inglês (Reino Unido)

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Abstract

This paper reports on an investigation of attitudinal stance (Biber et al. 1999, Conrad and Biber 1999) in native and learner speech that will be presented at TALC 2008. Attitudinal stance conveys speakers' attitudes, likes or dislikes, or evaluations of events or personal experiences for example. The focus of the study is on the adjectives that native speakers and advanced EFL learners use to express both positive and negative evaluation (Hunston and Sinclair 1999). The adjectives under investigation (e.g. good, great, nice, wonderful, bad, awful, terrible) fit into Biber et al.'s (1999) evaluative/emotive subcategory of descriptors. Native speakers' and learners' use of evaluative adjectives are analysed using the Louvain Corpus of Native English Conversation and the French, German and Chinese components of the Louvain International Database of Spoken English Interlanguage. The analysis concentrates more specifically on the preferred syntactic and collocational patterns in which these adjectives are used and possible practical implications of some of the findings for ELT are outlined. After providing a description of the data and the adjectives focused on in the study, this short report presents and discusses selected qualitative findings from a contrastive study of evaluative adjectives in native speaker speech and in the spoken productions of French-speaking advanced EFL learners.

Keywords: Attitudinal stance, evaluative adjectives, syntactic and collocational patterns, learner corpora, spoken English

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Introduction

This paper reports on an investigation of attitudinal stance (Biber et al. 1999, Conrad and Biber 2000) in native and learner speech that will be presented at TALC 2008. According to Conrad and Biber (1999: 57) attitudinal stance conveys 'speakers' attitudes, feelings, or value judgements.' The focus of the study is on the adjectives that native speakers and advanced EFL learners use to express both positive and negative evaluation (Hunston and Sinclair 1999). Beside range and frequency, the investigation of positive and negative evaluative adjectives in native speaker speech and in the spoken productions of advanced EFL learners from Chinese, French and German mother tongue backgrounds focuses more specifically on the preferred syntactic and collocational patterns in which these adjectives are used. The aim of this short report is threefold as it sets out (1) to describe the corpora used in the study, (2) to discuss the selection of the adjectives under investigation and (3) to zoom in on a number of qualitative results from a contrastive study of evaluative adjectives in native speaker speech and in the spoken productions of French-speaking advanced EFL learners.

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Introducing LINDSEI and LOCNEC

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

In the study that will be presented at TALC 2008 learners' and native speakers' use of evaluative adjectives are analysed using the Chinese, French and German components of the Louvain International Database of Spoken English Interlanguage (henceforth LINDSEI) and a comparable native speaker corpus, the Louvain Corpus of Native English Conversation (henceforth LOCNEC). The LINDSEI project was launched in 1995 at the Centre for English Corpus Linguistics, Université catholique de Louvain, as the spoken counterpart of the International

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

[32] Sylvie De Cock is a lecturer in English language and linguistics at the Université catholique de Louvain and at the Facultés universitaires Saint-Louis (Belgium). She is also a member of the Centre for English Corpus Linguistics (Université catholique de Louvain), directed by Sylviane Granger. She has been involved in the compilation of the Louvain International Database of Spoken English Interlanguage (LINDSEI) since 1996 and collected and compiled the Louvain Corpus of Native English Conversation as part of her PhD project (De Cock 2003). Her main research interests include corpus linguistics, learner corpora, phraseology, pedagogical lexicography and the study of English for Specific Purposes (Business English).

Corpus of Learner English (ICLE, Granger 1998). LINDSEI is made up of informal interviews with advanced EFL learners and currently contains data from learners from 12 different mother tongues (i.e. Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Norwegian, Polish, Spanish and Swedish). The learners who contributed data to LINDSEI are labelled as advanced on the basis of an external criterion: they are third and fourth year students of English at university. The Louvain Corpus of Native English Conversation (LOCNEC) is actually something of a misnomer as it contains informal interviews with British university students. The corpora used in the study total between approximately 70,000 and 110,000 words of interviewee speech.

The informal interviews in LINDSEI and LOCNEC are of similar length and follow the same set pattern: the main body of the interviews takes the form of an informal and open discussion mainly centred around topics such as university life, hobbies, foreign travel or plans for the future. Each interview starts with one of three topics (topic 1: an experience that taught them a lesson, topic 2: a country that impressed them, topic 3: a film or play they liked/disliked), which the students were given a few minutes to choose and think about. This was designed to make the interviewees, and especially the learners, feel at ease. The students were, however, specifically asked not to make any notes as it was intended for the spoken productions to be as spontaneous as possible. Each interview concludes with a short picture-based story-telling activity.

Results from a separate analysis of recurrent sequences of words conveying attitudinal stance in the BNC spoken component and in the corpora used in this study seem to suggest that LINDSEI and LOCNEC lend themselves particularly well to the study of a group of speakers' shared repertoire of preferred ways of expressing evaluations (De Cock 2007). The informal interviews contained in LINDSEI and LOCNEC are packed with frequently recurring expressions of personal attitudes and feelings (e.g. *I love*, *I really like*, *I really enjoy(ed)*, *I enjoy(ed) it*, *(yes) I like it*), which can be closely related to the types of topics discussed and the informal character of the interviews.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Identifying positive and negative evaluative adjectives

The positive and negative evaluative adjectives under investigation were identified on the basis of frequency lists of word forms from the corpora (using Wordlist, WordSmith Tools 4.0). Two main criteria were used in the selection process: the first relates to frequency and the second to meaning. Were taken into consideration in the investigation reported on here the evaluative adjectives that recur at least five times in either LOCNEC and/or in any of the LINDSEI components. The frequency threshold adopted goes some way towards ensuring that the adjectives selected are not the result of individual usage. With respect to meaning, the adjectives selected for inclusion in the analysis had to be considered as prototypically evaluative (i.e. their overt and only purpose is to evaluate, Channel 1999) and to fit into Biber et al.'s (1999) evaluative/emotive subcategory of descriptors. Examples of such prototypical evaluative adjectives include *good*, *great*, *nice*, *wonderful*, *bad*, *awful*, or *terrible*. Adjectives such as *big* and *little* were excluded from the analysis because, although they can be used to express speakers' subjective positive or negative evaluations of things or events as in *a nice little house*, *poor little thing* or *a big mistake*, they tend to be used prototypically to indicate size. The selection process was followed by a disambiguation process in which any instances of the word forms not used as adjectives were removed from the working data after careful examination of the items in context (using Concord, WordSmith Tools 4.0). The instances of *good* and *best* in (1) and (2) were for example discarded:

(1) they . just er . try their **best** to help help us (LINDSEI_Chinese)

Formatada: Avanço: Esquerda: 0 cm, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

(2) I didn't think of that . erm . yeah for a while . but not for **good** because erm . I love my . country (LINDSEI_German)

It is noteworthy that some of the most frequently recurring evaluative adjectives in all the corpora used can be traced to one specific part of the interview format in LOCNEC and LINDSEI. The adjective *beautiful* is a case in point as it is almost exclusively used when the native speakers and especially the learners carry out the short picture-based story telling task at the end of the interview. This is illustrated by the following extract from the German component of LINDSEI (B = the learner interviewee; A = the interviewer):

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

> <B> erm there are two people . er sitting in a room . one is a model . and the[i:] other one is a painter .. and: erm . yeah he draws a picture of her .. and: when he's finished he shows the picture to the woman .

and it's really good . I mean it's exactly the woman . that he painted but the woman .. doesn't really like herself [<laughs> <\B>

<A> [<laughs> <\A>

<B> and: em tells him to make her more **beautiful** .. well erm .. the painter looks like . he's .. erm kind of surprised and doesn't really understand why she's <\B>

<A> [<laughs> <\A>

<B> [. acting like that but . well erm . I don't know why he does it but he changes the picture and now she's looking really really **beautiful** . and: erm . in the[i:] end she: shows it to her .. friends . her female friends .. and: seems to be very happy <\B>

Zooming in on some preferred syntactic patterns

The selected findings discussed in this section illustrate differences in the preferred syntactic patterns in which some evaluative adjectives are used in the native speaker corpus and in the French component of LINDSEI (henceforth LINDSEI_French). The focus is on a number of patterns that appear to be preferred in the native speaker corpus.

Findings from a comparison of the use of the evaluative adjective *good* in attributive vs. predicative position in the native speaker corpus and in LINDSEI_French provide a good illustration of differing preferred syntactic environments. While in the native speaker corpus *good* tends to be used more frequently in predicative position (203 instances, in approximately 58% of the cases; see examples 3 and 4), the preferred position in the learner corpus is inside an NP before the head noun as in (5). *Good* occurs in predicative position in 48% of the cases in LINDSEI-French (51 instances). A closer examination of the native speaker data reveals that the adjective *good* is used predicatively in a number of frequently recurring sequences which are not recurrent in the learner corpus. These sequences include *it's good*, *it was good*, *(and / yeah) it was really good*, *it was quite good*, *that was good*, *it's really good*.

(3) <B> oh right er .. I might have to look into that that sounds really **good** <\B> (LOCNEC)

(4) (...) wrote a script for a play at college . and I got to see that acted out and it was **really** good (...) (LOCNEC)

(5) <B> it was a very **good** experience for me . yes because they . made a lot of critics (...) (LINDSEI-French)

The study reveals that the use of the evaluative adjective *bad* follows a similar trend. It is used in predicative position in 75% of the cases in the native corpus (47 instances) and the sequence *it's not too bad* clearly emerges as frequently recurring in the data (9 instances, see example 6). By contrast, *bad* is used predicatively in only 52% of the cases in LINDSEI_French (12 instances) and it does not occur as part of any recurring sequence.

(6) <B> yeah you've got like a bit of countryside around and so it's not too **bad** yeah <\B> (LOCNEC)

Differences in preferred syntactic patterns can also be seen to relate to patterns of adjective complementation. The use of the adjective *nice* in predicative position in the native speaker corpus and in LINDSEI_French is an illustration of this. While examples 7 and 8, where nice is followed by a to-infinitive clause, represent almost 20% of the uses of the adjective in predicative position in LOCNEC (25 instances), this pattern of use represents a mere 4% of the uses of the adjective in predicative position in LINDSEI_French (3 instances) .

(7) <B> [I don't know it'd sort of be **nice** *to have our own house in a way* <\B> (LOCNEC)

(8) <B> it's **nice** *to walk round and you know fancy something* <\B> (LOCNEC)

Another preferred syntactic environment for evaluative adjectives in the native speaker corpus is the sentential relative clause, as in examples (9) to (13). While a number of evaluative adjectives can be seen to occur in these clauses (e.g. *brilliant*, *excellent*, *great*, *amazing*, *interesting*, *impressive*, *bad*), the adjectives *good* and *nice* appear to be particularly comfortable and frequent in such an environment (c. 15 instances for each adjective). Unlike

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

LOCNEC, LINDSEI_French contains few sentential relative clauses and only a handful of these contain evaluative adjectives (see example 14).

> (9) <B> first of all the way that they speak English which is **great** for me cos I speak no Dutch at all <\B> (LOCNEC)
>
> (10) <B> you know you've always .. I've got other people to hitch with which is **good** <\B> (LOCNEC)
>
> (11) we run them up and down our little railway and children come and ride on it which is really **good** <laughs> <\B> (LOCNEC)
>
> (12) <B> and again it's only five minutes' walk from my house which is **nice** and there's a bar <\B> (LOCNEC)
>
> (13) <B> they all come and visit me cos they think it's great having a student life so close to <X> so a lot of them travel up at weekends and that [which is quite **nice** <B> (LOCNEC)
>
> (14) (...) they don't ask me to pay so . this is sometimes for just one article for thirty pages of something er I don't have to pay which is really eh **interesting** <\B> (LINDSEI_French)

Evaluative sentential relative clauses would in fact be a particularly good candidate for inclusion in ELT reference materials or textbooks. As well as providing learners with a wider range of ways of expressing attitudinal stance in interactions (see Tao and McCarthy 2001), giving more prominence to this type of clauses could also help learners cope with the pressures of on-line processing in unplanned spoken discourse. The use of sentential relative clauses is indeed consistent with the 'clause chaining style' or clause 'add-on strategy' that has been shown to be particularly well-suited to the constraints of real-time planning (Biber et al 1999).

Looking ahead

This short report on an investigation of positive and negative adjectives in native speaker speech and learner speech has essentially focused on differences in some of the preferred syntactic patterns in which evaluative adjectives are used in the native speaker corpus and in LINDSEI_French. The full report, which will also include results from LINDSEI_Chinese and LINDSEI_German, will give a general quantitative picture of the use of positive and negative evaluative adjectives (range and frequency) and will discuss the results of a contrastive analysis of the preferred syntactic and collocational patterns in which the adjectives are used in the four varieties under investigation. The paper will conclude by outlining a number of possible practical implications of some of the findings for ELT.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. and **Finegan, E.** 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Conrad, S. and **Biber, D.** 1999. "Adverbial marking of stance in speech and writing." In *Evaluation in Text. Authorial Stance and the Construction of Discourse*, S. Hunston and G. Thompson (eds). Oxford: Oxford University Press, 56-73.

Channel, J. 1999. "Corpus-based analysis of evaluative lexis." In *Evaluation in Text. Authorial Stance and the Construction of Discourse*, S. Hunston and G. Thompson (eds). Oxford: Oxford University Press, 38-55.

De Cock, S. 2003. *Recurrent sequences of words in native speaker and advanced learner spoken and written English*. Unpublished PhD thesis, Université catholique de Louvain.

De Cock, S. 2007. "Routinized building blocks in native speaker and learner speech: Clausal sequences in the spotlight." In *Spoken Corpora in Applied Linguistics*, M.C. Campoy and M.J. Luzón (eds). Bern: Peter Lang, 217-233.

Granger, S. 1998. "The computerized learner corpus: a versatile new source of data for SLA research." In *Learner English on Computer*, S. Granger (ed.). London and New York: Addison Wesley Longman, 3-18.

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: 10,5 pt, Francês (França)

Formatada: Tipo de letra predefinido do parágrafo, Tipo de letra: 10,5 pt, Negrito, Francês (França)

Formatada: Tipo de letra predefinido do parágrafo, Tipo de letra: 10,5 pt, Francês (França)

Hunston, S. and **Sinclair, J.** 1999. A local grammar of evaluation. In *Evaluation in Text. Authorial Stance and the Construction of Discourse*, S. Hunston and G. Thompson (eds). Oxford: Oxford University Press, 74-101.

Tao, H. and **McCarthy, M.J**. 2001. "Understanding non-restrictive *which*-clauses in spoken English, which is not an easy thing." *Language sciences* 23; 651-677.

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

AN ACADEMIC FORMULAS LIST (AFL):
CORPUS LINGUISTICS, PSYCHOLINGUISTICS, AND EDUCATION

Nick C. Ellis[33]

Rita Simpson-Vlach[34]

Abstract

This research creates an empirically derived, pedagogically useful list of formulaic sequences for academic speech and writing, comparable to the Academic Word List (Coxhead 2000), called the Academic Formulas List (AFL). The AFL includes formulaic sequences identified as (1) frequent recurrent patterns in corpora of written and spoken language, which (2) occur significantly more often in academic than in non-academic discourse, and (3) inhabit a wide range of academic genres. We assess the instructional and psycholinguistic validity of these formulas in order to prioritize them using an empirically derived measure of utility that is educationally and psychologically valid and operationalizable with corpus linguistic metrics. The formulas are classified according to their predominant pragmatic function for descriptive analysis and in order to marshal the AFL for inclusion in English for Academic Purposes instruction.

Keywords: An Academic Formulas List (AFL), Corpus Linguistics, Psycholinguistics, Education, English for Academic Purposes

Introduction

The specific aim in this research is to create an Academic Formulas List (AFL), a pedagogically useful list of formulaic sequences for academic speech and writing comparable to the Academic Word List (hereafter AWL, Coxhead 2000). Corpus linguistic analyses of written and spoken academic discourse allow us to identify recurring, high-frequency lexical bundles, phrases, or formulas, and research has shown that these are important characteristics of academic registers (Biber, Conrad et al. 2004; Simpson 2004). Cognitive scientific analyses also inform us that knowledge of these formulas is crucial for fluent processing. Second language acquisition researchers and EAP practitioners need a prioritized list of the most important formulas characterizing academic discourse, which as of yet has not been available.

Our research therefore triangulates the construct of 'formula' from corpus linguistic, psycholinguistic and educational perspectives.

Corpus extraction of the AFL

Three, four, and five word formulas occurring at least 10 times per million words were extracted from corpora of 2.1M words of academic spoken language from MICASE (Simpson, Briggs, Ovens, & Swales, 2002) and selected academic spoken BNC files (British National Corpus, 2006), 2.1M words of academic written language from Hyland's (2004) research article corpus, plus selected academic writing BNC files, 2.9M words of non-academic

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

[33] Nick Ellis is a Research Scientist and Professor of Psychology at the University of Michigan. His research interests include language acquisition, cognition, reading in different languages, corpus linguistics, cognitive linguistics, and applied psycholinguistics. Currently his research focuses on second language acquisition, particularly (1) explicit and implicit language learning and their interface, (2) usage-based acquisition and the probabilistic tuning of the system, (3) vocabulary and phraseology, (4) language and brain, (5) the advanced language learner, (6) applications of psychological theory in language testing and instruction, (7) learned attention and language transfer, (8) emergentist accounts of language acquisition.

[34] Rita Simpson-Vlach was a Research Associate at the English Language Institute of the University of Michigan, where she served as project director of the Michigan Corpus of Academic Spoken English from its inception until 2006. Most recently she has been a lecturer in the Department of Linguistics and Language Development at San José State University. Her research interests lie mainly in the area of corpus linguistics and EAP, specifically in the use of corpora for pragmatic and discourse analyses and for use in EAP teaching materials development.

speech from the Switchboard (2006) corpus, and 1.9M words of non-academic writing from the FLOB and Frown corpora gathered in 1991 to reflect British and American English over 15 genres (ICAME, 2006).

The program Collocate (Barlow, 2004) allowed us to measure the frequency of each n-gram along with the mutual information (MI) score for each phrase. MI is a statistical measure commonly used in the field of information science designed to assess the degree to which the words in a phrase occur together more often than would be expected by chance; it is a measure of how much they cohere or are found in collocation. A higher MI score means a stronger association between the words, while a lower score indicates that their co-occurrence is more likely due to chance. High frequency n-grams occur often. But this does not imply that they have clearly identifiable or distinctive functions or meanings; many of them occur simply by dint of the high frequency of their component words, often grammatical functors. High MI n-grams, in contrast, are those with much greater coherence than is expected by chance, and this tends to correspond with distinctive function or meaning as well as grammatical well-formedness as a complete phrase.

The total number of formulas appearing in any one of the four corpora at the threshold level of 10 per million was approximately 14,000. In order to determine which formulas were more frequent in the academic corpora than in their non-academic counterparts, we used the log-likelihood (LL) statistic to identify the formulas which were statistically more frequent, at a significance level of p<.01, in the academic corpora than in their non-academic counterparts. We separately compared academic speech vs. non-academic speech, resulting in over 2000 items, and academic writing vs. non-academic writing resulting in just under 2000 items.

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Instructional Validation of Academic Formulas

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Our investigation of educational validity of these academic formulas used a representative sample of 108 of them, 54 from the Speech list and 54 from the Written list. These were chosen by stratified random sampling to represent three levels on each of three factors: *n*-gram length (3, 4, 5), Frequency band (High, Medium, and Low; means 43.6, 15.0 and 10.9 per million respectively), and MI band (High, Medium, and Low; means 11.0, 6.7, and 3.3 respectively). There were two exemplars in each of these cells. Example items are shown in Table 1.

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

		Mutual Information		
		Low (3.3)	Medium (6.7)	High (11)
Frequency (n per million)	Low (10.9)	that the only the length of the in the context of the	happens is that and so on but as in the case of	circumstances in which it has been shown of the court of appeal
	Medium (15.0)	and at the the value of the the way in which the	that may be the relationship between the it is not possible to	see for example a wide variety of it should be noted that
	High (43.6)	the content of is one of the in the case of the	a kind of the extent to which at the beginning of	in other words a great deal of it can be seen that

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin, Manter com o seguinte

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin, Manter com o seguinte

Table 1: Sample formulaic sequences factorially crossing *n*-gram Length (3, 4, 5), Frequency (low, medium, high), and Mutual Information (low, medium, high)

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

The stratified sample of 108 *n*-grams in total constituted the stimuli for the Instructor judgments of formulaicity and the Psycholinguistic Processing experiments. We asked experienced EAP instructors and language testers at the English Language Institute of the University of Michigan to rate these formulas, given in a random order of presentation, for one of three judgments using a scale of 1 (disagree) to 5 (agree):

A. whether or not they thought the phrase constituted ‘a formulaic expression, or fixed phrase, or chunk’. There were 6 raters with an inter-rater □ = 0.77.

B. whether or not they thought the phrase had ‘a cohesive meaning or function, as a phrase’. There were 8 raters with an inter-rater □ = 0.67

C. whether or not they thought the phrase was ‘worth teaching, as a bona fide phrase or expression’. There were 6 raters with an inter-rater □ = 0.83

Formulas which scored high on one of these measures tended to score high on another: *r* AB = 0.80, *p* < .01; *r* AC = 0.67, *p* < .01; *r* BC = 0.80, *p* < .01). The high alphas of the ratings on these dimensions and their high intercorrelation reassured us as to the reliability and validity of these instructor insights. We then investigated which of Frequency or MI better predicted the insights. Correlation analysis suggested that while both of these dimensions contributed to instructors valuing the formula, it was MI which most influenced their prioritization: *r* Frequency/A = 0.22, *p* < .05; *r* Frequency/B = 0.25, *p* < .05; *r* Frequency/C = 0.26, *p* < .01; *r* MI/A = 0.43, *p* < .01; *r* MI/B = 0.51, *p* < .01; *r* MI/C = 0.54, *p* < .01. A multiple regression analysis predicting instructor insights regarding whether an *n*-gram was worth teaching as a bona fide phrase or expression from the corpus metrics gave a standardized solution whereby teaching worth = □□0.56 MI + □ 0.31 Frequency.

The high intercorrelations of the instructor ratings suggest a latent factor of formulaicity underlying their judgments. The significant associations between the corpus metrics of *n*-gram frequency and MI, and the various instructor judgments of *n*-gram formulaicity, identifiably of function, and teaching-worth suggest a successful triangulation of instructor insights and corpus metrics: In other words, these corpus-derived measures do serve to identify *n*-grams that instructors judge to be clearly identifiable formulas which are worth teaching. Both *n*-gram frequency and MI factor into this prediction, but it is the MI of the string – the degree to which the words are bound together – that is the major determinant.

Psycholinguistic Validation of Academic Formulas

Four experiments then determined which of these factors affected the accuracy and fluency of processing of these formulas in native language speakers of English and in advanced second language learners of English (all students at a large North American University). The language processing tasks were: (1) rate of reading and rate of spoken articulation, (2) speed of reading and acceptance in a grammaticality judgment task where half of the items were real phrases in English and half were not, (3) speed of comprehension and acceptance of the formula as being appropriate in a sentence context, (4) binding and primed pronunciation: the degree to which reading the beginning of the formula primed recognition of its final word. These tasks were selected to sample an ecologically valid range of language processing skills: spoken and written, production and comprehension, form-focused and meaning-focused. Processing in all experiments was affected by the various corpus-derived measures: length, frequency, MI, and source, but to very different degrees in the different learners. For native speakers it is predominantly the MI of the formula which determines its processability. For non-native learners of the language it is predominantly the frequency of the formula which determines its accuracy and fluency of processing. These findings have important implications for the psycholinguistic validity of corpus-derived formulas, their acquisition, and their instruction.

The Academic Formulas List (AFL)

The resultant AFL includes formulaic sequences identified as (1) frequent recurrent patterns in corpora of written and spoken language, which (2) occur significantly more often in academic than in non-academic discourse, and (3) inhabit a wide range of academic genres. It lists formulas that are common in academic spoken *and* academic written language, as well as those that are special to academic written language alone and academic spoken language alone. The AFL further prioritizes these formulas using an empirically derived measure of utility that is educationally and psychologically valid and operationalizable with corpus linguistic metrics.

The final stage of the analysis involved grouping the formulas into categories according to their primary discourse-pragmatic functions in academic speech and writing. For purposes of space and time, as well as the

Formatada: Inglês (Reino Unido)

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Inglês (Reino Unido)

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Inglês (Reino Unido)

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Inglês (Reino Unido)

anticipated pedagogical applications, we did not group all the formulas from all three lists, but again included only those from the Core AFL list and the top 200 from the Written AFL and the Spoken AFL lists. These functional categories—determined after examining the phrases in context using a concordance program—are not meant to be taken as definitive and exclusive, since many of the formulas have functions that span multiple categories, but rather as indications of the most common or salient function the phrases fulfill in academic contexts.

In our presentation we illustrate the AFL and give an overview of the functional analysis, providing examples to illustrate some of the more important formula functions in context.

References

Barlow, M. 2004. *Collocate*. Houston: Athestan Publications.

Biber, D., Conrad, S., & Cortes, V. 2004. “If you look at …”: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371-405.

British National Corpus. 2006. from http://www.natcorp.ox.ac.uk/.

Coxhead, A. 2000. A new Academic Word List. *TESOL Quarterly, 34*, 213-238.

Hyland, K. 2004. *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor: University of Michigan Press.

ICAME. 2006. from http://icame.uib.no/.

Simpson, R. 2004. 'Stylistic features of academic speech: The role of formulaic expressions' in T. Upton and U. Connor (eds.): *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins.

Simpson, R., Briggs, S., Ovens, J., & Swales, J. M. 2002. *The Michigan Corpus of Academic Spoken English*. [Electronic Version]. The Regents of the University of Michigan from http://www.hti.umich.edu/m/micase.

Switchboard. 2006, August 5, 2006. *A User's Manual*. from

http://www.ldc.upenn.edu/Catalog/docs/switchboard/

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Centrado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: 10,5 pt, Inglês (E.U.A.)

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

THE PEDAGOGIC VALUE OF CORPORA: A CRITICAL EVALUATION

Lynne Flowerdew[35]

Abstract

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

In the phraseological approach to corpus analysis, the lexical item has primacy, with its core meaning and semantic prosody as obligatory categories, and collocation, colligation and semantic preference considered as optional categories. The use of corpora in Data-Driven Learning (DDL) has been invaluable in highlighting this phraseological nature of language. However, several drawbacks of using corpora in DDL have been raised in the literature in recent years. This paper discusses three issues. DDL has been promoted as 'discovery learning', but the value of this kind of incidentalist learning has been questioned. Corpus-driven learning, usually associated with the inductive approach, may not necessarily suit the learning style of all students. Students may have difficulty interpreting and authenticating the corpus data for their own purposes, for which some kind of 'pedagogic processing' is necessary.

Keywords: DDL, inductive, pedagogic processing, discovery learning,

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Introduction

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Corpus linguistics is usually associated with a phraseological approach to analysis, which takes a syntagmatic, as opposed to a paradigmatic, view of language. Following Sinclair (1999, 2004) the lexical item has primacy, with its core meaning and semantic prosody as obligatory categories, and collocation, colligation and semantic preference considered as optional categories. An interweaving of some or all of these categories gives what Sinclair refers to as an 'extended unit of meaning'. Analysis of recurring patterns in concordance output has revealed how language can follow certain tendencies according to Sinclair's notion of an 'extended unit of meaning' rather than being bound by hard-and-fast rules. The use of corpora in Data-Driven Learning (DDL) has been invaluable for highlighting this phraseological nature of language. Nothwithstanding the advantages of this approach, during the last few years some accounts in the literature have adopted a more critical stance, drawing attention to potential drawbacks of using corpora in DDL.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Key issues in applying corpus linguistics to pedagogy

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

This paper reviews the following three key issues in the debate on applying corpus linguistics to pedagogy.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

DDL is promoted as 'discovery learning' or 'serendipitous learning', but the value of this kind of incidental learning has been questioned.

Corpus-driven learning is usually associated with an inductive, i.e. 'discovery-based' rather than a rule-based, deductive approach to learning. This approach may not necessarily be the most appropriate choice for some students.

Corpus data are decontextualised. Such data have to be transformed from samples of language to authentic examples to fit the students' local context. Some kind of 'pedagogic processing' is therefore necessary.

The above issues will be discussed with reference to some of my own research on expert and learner corpora and experience in facilitating DDL activities in an undergraduate academic writing environment. The discussion points will be exemplified using a suite of online tools, *My Words* (http://mywords.ust.hk), developed at a tertiary

[35] Lynne Flowerdew coordinates communication skills courses at the Hong Kong University of Science and Technology. Her main research interests include corpus linguistics, discourse analysis, genre analysis, EAP/ESP syllabus design and methodology.

Formatada: Avanço: Esquerda: 0 cm, Espaçamento entre linhas: simples

Formatada: Espaçamento entre linhas: simples institution in Hong Kong. These tools comprise not only a variety of sub-corpora but also dictionaries, thesauri, and a collocational tool, *JustTheWord*, which interfaces with the BNC (Flowerdew 2006).

'Serendipitous' or 'incidental' learning?

The following is an example of what the author considers to be successful serendipitous learning (Flowerdew in press, 2008b). She notes that in a course on report writing one student query related to whether the active or passive voice was used in the following sentence:

This project *focuses / is focused* on the incidence of mosquitoes on campus.

A search on 'focus' was conducted in an institutionally-compiled 7 million-word corpus of reports in the "My Words" suite of programs referred to earlier, which gave the results shown in Figure 1 below.

Pattern Left sort Right Sort			*Frequency Sort*
NOUN + VERB + PREP:	*e.g. "study focuses on"*	*Show results*	*292*
VERB + VERB + PREP:	*e.g. "has focused on"*	*Show results*	*231*
TO + VERB + PREP:	*e.g. "to focus on"*	*Show results*	*95*
ADV + VERB + PREP:	*e.g. "not focused on"*	*Show results*	*94*
PRON + VERB + PREP:	*e.g. "we focus on"*	*Show results*	*57*
CONJ + VERB + PREP:	*e.g. "that focus on"*	*Show results*	*56*
DET + VERB + PREP:	*e.g. "which focus on"*	*Show results*	*38*
VERB + VERB + ADV	*e.g. "has focused almost"*	*Show results*	*34*
NOUN + VERB + ADV	*e.g. "efforts focused primarily"*	*Show results*	*33*
TO +VERB + ADV	*e.g. "to focus more"*	*Show results*	*31*

Search for 'focus' (all word forms) (in press, Flowerdew 2008b)

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Besides the fact that the students were able to glean the different meanings between the active and passive forms of 'focus' by examining the verb in a wider context, accessed via 'Show results', this search encouraged students to engage with the corpus results beyond their original query. Students' scrutiny of the concordance output prompted one student to ask: 'Why are there so many occurrences of 'focus' in the present perfect'? This kind of comment which I have termed a 'triggered query' because it is activated by something the student has alighted on in the corpus data, unprompted by the teacher (Flowerdew 2007), echoes Swain's (1998) concept of 'noticing'. Swain (1998: 66) remarks that there are several levels of 'noticing', one of which is that: 'Learners may simply notice a form in the target language due to the frequency or salience of the features themselves'. An examination of the wider context of the present perfect forms of 'focus' revealed that this tense was used when previous research was introduced, to set up a critical evaluation of this work signaled by 'However', replicating Swales' (1990) CARS (create a research space) model.

This type of browsing is thus in the spirit of Bernardini's philosophy as the "learner as traveler" (Bernardini 2004). Although Swales (2002; see also Lee & Swales 2006) has voiced some misgivings about this type of serendipitous learning advocated by Bernardini (2000, 2002), referring to it as 'incidentalist', an example such as the one above illustrates that this ad hoc browsing can encourage students to process corpus data in a productive way. However, as Flowerdew (2008a) notes, sometimes such browsing results in time-wasting and going down a few blind alleys, but preliminary investigations indicate that when successful, this type of browsing encourages students to formulate further 'triggered queries'.

Inductive or deductive approach?

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Both Gavioli (2005) and Meunier (2002) have noted the drawbacks of an inductive approach, in which students extrapolate the rules, or patterning, from examples.

Despite their advantages, DDL activities have some drawbacks…. The various learning strategies (deductive vs. inductive) that students adopt can lead to problems. Some students hate working inductively and teachers should aim at a combined approach (see Hahn 2000 for a combined approach). (Meunier 2002: 135)

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

I would like to put forward two possible reasons as to why an inductive approach may not always be appropriate. First, as suggested in Flowerdew (2008b), an inductive approach may not appeal to students with different cognitive styles. Field-dependent students who thrive in cooperative, interactive settings and who would seem to enjoy discussion centering on extrapolation of rules from examples may benefit from this type of pedagogy. However, field-independent learners who are known to prefer instruction emphasizing rules may not take to the inductive approach inherent in corpus-based pedagogy. It is interesting to note that Vannestål & Lindquist (2007: 343) state that some of the students in their inductive corpus-based grammar course commented that '…they preferred the more traditional way of reading about grammatical rules in the book and did not feel that they learned anything by doing corpus exercises'.

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Secondly, whether an inductive or deductive approach is adopted would very much seem to depend on the nature of a particular enquiry. If the enquiry is based on a hard-and-fast grammar rule (for example, the difference between *for* and *since* in time expressions; see Tribble & Jones 1990), then the differences are quite clear cut. However, if the enquiry focuses on an aspect of phraseology, students may find it difficult to extrapolate the *tendencies* associated with patterns in language (Hunston & Francis 2000), as they may be confronted with conflicting examples which do not follow a particular pattern in all cases.

One area that posed difficulty for my students was that of ergativity. As noted by Celce-Murcia (2002), overpassivisation of ergative verbs is an aspect that poses particular problems for advanced learners:

With the verbs 'increase' and 'decrease' [the ergative] *tends* [my italics] to be used when the inanimate subject is objectively or subjectively measurable (rather than an animate agent/dynamic instrument object – both of which favor active voice – or a patient subject – for the passive voice.(Celce-Murcia 2002: 146)

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Students found it difficult to work out from a close reading of concordance lines the correct choice of verb in the following sentence because of the probabilistic nature of language when viewed syntagmatically:

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

With a very crowded schedule, students' level of motivation was decreased / has decreased.

Vannestål & Lindquist (2007) have commented on the difficulty students have in interpreting corpus data and this aspect seems to be a particularly thorny issue when phraseology comes into play. It would seem then that it is in order to supply prompts or hints to enable students to work out the tendencies of phraseological patterns. For example, in the case of the use of the ergative students could be given a prompting question such as: 'Do you notice any difference in the subjects for '…was decreased' and 'has decreased'?

Corpora are useful for phraseological enquiries (cf. Granger & Meunier (eds.) 2008, Meunier & Granger (eds.) 2008) as the language which falls between lexis and grammar is often not easily retrievable from grammars or dictionaries. However, some intervention in the form of clues or hints may be necessary, i.e. 'pedagogic processing' (see following section). Conversely, while hard-and-fast grammar rules may be easier for students to glean from corpora, a corpus, or indeed a particular sub-corpus, may not be the best, or most efficient, resource for consultation.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

'Decontextualisation' or 'Recontextualisation'?

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Widdowson's (2004) argument on the decontextualised nature of corpus data are well-rehearsed in the literature (see Flowerdew 2008a; Braun 2005; Kaltenböck & Larcher 2005; McEnery et al 2006), but it is worth reviewing them again briefly. Both Aston (1995) and Widdowson (1998, 2002) have drawn attention to the decontextualised nature of corpus data, with Widdowson commenting that corpus data are but a sample of language, as opposed to an example of authentic language, because it is divorced from the communicative context in which it was created: 'the text travels but the context does not travel with it.'

Widdowson maintains that it may not be expedient to transfer corpus data directly to pedagogic materials on account of the cultural or contextual inappropriacy of the corpus data (see Cook 1998; Widdowson 1991, also cited in Seidlhofer 2003, for a discussion on the issue of prescription vs. description, regarding the transfer of corpus data to pedagogy). Widdowson therefore advocates adopting some kind of 'pedagogic processing', as do other corpus linguists such as Braun (2007) and McCarthy (2001) in order to transform samples of language into pedagogically-accessible examples.

In order to integrate the type of pedagogic processing Widdowson is referring to so as to enable students to authenticate the corpus data for their own contextual writing environment, Flowerdew (2008b) has adopted student peer response activities, which draw on Vygotskian socio-cultural theories of co-constructing knowledge through collaborative dialogue and negotiation (see O'Sullivan (2007) who gives a very insightful exposition on the role of cognitive and social constructivist theories to foster corpus consultation literacy). In these peer-to-peer interaction groups, weaker students were intentionally grouped with more proficient ones to foster productive dialogue through 'assisted performance', thus drawing on another aspect of socio-cultural theory. The author reports some success with this approach of incorporating group discussion activities revolving around the corpus data as a form of pedagogic mediation, resulting is consciousness-raising of register awareness, not only for the task in hand, but also what might be appropriate phraseologies for other contexts. Students were therefore encouraged to engage in 'collaborative metatalk' (Swain 1998:68) to 'use language to reflect on language use'. This shows there is a case for incorporating contextual information in written texts to aid the transfer of corpus data to pedagogy.

Suggestions for other types of pedagogic mediation of corpora have been given by Braun (2005) for inclusion of video activities, by Milton (2006) for didactic written hints built into the software, and by Vannestål and Lindquist (2007) for peer teaching. Although there are a few accounts in the literature regarding the 'pedagogic mediation' of corpus data, these are few and far between, indicating this is an area ripe for further discussion and expansion.

A brief discussion of the above three key issues concerning the application of corpus linguistics to pedagogy reveals there is still much to debate and develop in DDL.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

References

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Aston, G. 1995. "Corpora in language pedagogy: Matching theory and practice." In *Principle and Practice in Applied Linguistics*, G. Cook & B. Seidlhofer (eds). Oxford: Oxford University Press, 257-270.

Bernardini, S. 2000. "Systematising serendipity: Proposals for concordancing large corpora with language learners." In *Rethinking Language Pedagogy from a Corpus Perspective*, L. Burnard & T. McEnery (eds). Frankfurt: Peter Lang, 225-234.

Bernardini, S. 2002. "Exploring new directions for discovery learning." In *Teaching and Learning by Doing Corpus Analysis*, B. Kettemann & G. Marco (eds). Amsterdam: Rodopi, 165-182.

Bernardini, S. 2004. "Corpora in the classroom: An overview and some reflections on future developments." In *How to Use Corpora in Language Teaching*, J. Sinclair (ed.). Amsterdam: John Benjamins, 15-36.

Braun, S. 2005. "From pedagogically relevant corpora to authentic language learning contents." *ReCALL*, 17/1: 47-64.

Braun, S. 2007. "Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora." *ReCALL*, 19/3: 307-328.

Celce-Murcia, M. 2002. "On the use of selected grammatical features in academic writing." In *Developing Advanced Literacy in First and Second Languages*, M. Schleppegrell & C. Colombi (eds). Mahwah, N.J.: Lawrence Erlbaum, 143-157.

Cook, G. 1998. "The Uses of Reality: A reply to Ronald Carter." *ELT Journal*, 52/1: 57-63.

Flowerdew, L. 2006. "Texts, tools and contexts in corpus applications for writing." Paper presented in invited academic session "Current Trends in Corpus Linguistics Research". *40th Annual TESOL Convention*, Tampa, Florida, March 16, 2006.

Flowerdew, L. 2007. "The Pedagogic Value of Corpora: A Critical Evaluation." Invited lecture at the Hong Kong Association for Applied Linguistics, Hong Kong, March 5, 2007.

Flowerdew, L. 2008a. *Corpus-based Analyses of the Problem-Solution pattern: A phraseological analysis*. Amsterdam: John Benjamins.

Flowerdew, L. in press 2008b. “Corpus linguistics for academic literacies mediated through discussion activities.” In *The Oral-Literate Connection: Perspectives on L2 Speaking, Writing and other Media Interactions*, D. Belcher & A. Hirvela (eds). Ann Arbor, MI: University of Michigan Press.

Gavioli, L. 2005. *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.

Granger, S. & Meunier, F. (eds) in press, 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

Hunston, S. & Francis, G. 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Kaltenböck, G. & Larcher, B. 2005. “Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching.” *ReCALL*, 17/1: 65-84.

Lee, D. &. Swales J.M. 2006. “A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self compiled corpora.” *English for Specific Purposes*, 25/1: 56-75.

McCarthy, M. 2001. *Issues in Applied Linguistics*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-Based Language Studies*. London: Routledge.

Meunier, F. 2002. “The pedagogic value of native and learner corpora in EFL grammar teaching.” In *Computer Learner Corpora, Second Language Acquisition and Foreign Language* Teaching, S. Granger, J. Hung, & S. Petch-Tyson (eds). Amsterdam: John Benjamins, 119-141.

Meunier, F. & Granger, S. (eds) 2008. *Phraseology in Foreign Language Teaching*. Amsterdam: John Benjamins.

Milton, J. 2006. “Resource-rich Web-based feedback: Helping learners become independent writers.” In *Feedback in Second Language Writing*, K. Hyland & F. Hyland (eds). Cambridge: Cambridge University Press, 123-139.

O’Sullivan, I. 2007. “Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy.” *ReCALL*, 19/3: 269-286.

Seidlhofer, B. (ed.) 2003. *Controversies in Applied Linguistics*, (Section 2: Corpus Linguistics and Language Teaching). Oxford: Oxford University Press.

Sinclair, J.McH. 1999. “The lexical item.” In *Contrastive Lexical Semantics*, E. Weigand (ed.). Amsterdam: John Benjamins, 1-24.

Sinclair, J.McH. 2004. “The search for units of meaning.” In *Trust the Text*. London: Routledge, 24-48.

Swain, M. 1998. “Focus on form through conscious reflection.” In *Focus on form in Classroom Second Language Acquisition*, C. Doughty & J. Williams (eds). Cambridge: Cambridge University Press, 64-81.

Swales, J.M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, J.M. 2002. “Integrated and fragmented worlds: EAP materials and corpus linguistics.” In *Academic Discourse*, J. Flowerdew (ed.). Harlow, UK: Longman, 150-64.

Tribble, C. & Jones, G. 1990. *Concordances in the Classroom*. Harlow, UK: Longman.

Vannestål, M. & Lindquist, H. 2007. “Learning English grammar with a corpus: Experimenting with concordancing in a university grammar course.” *ReCALL*, 19/3: 329-350

Widdowson, H.G. 1991. “The description and prescription of language.” In *Georgetown University Round Table in Language and Linguistics*, J. Alatis (ed.). Washington, DC: Georgetown University.

Widdowson, H.G. 1998. “Context, community and authentic language.” *TESOL Quarterly*, 32/4: 705-716.

Widdowson, H.G. 2002. “Corpora and Language Teaching Tomorrow.” Keynote lecture delivered at the *Fifth Teaching and Language Corpora Conference*, Bertinoro, Italy, July 29, 2002.

Widdowson, H.G. 2004. *Text, Context, Pretext*. London: Blackwell.

Formatada: Inglês (Reino Unido)

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: Negrito, Inglês (Reino Unido)

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

COLLOCATIONS AND IDIOMS IN ELT TEXTBOOKS: THIRST FOR EFFICIENT METALANGUAGE

Céline Gouverneur[36]

Abstract

The aim of this paper is to shed light on the way collocations and idioms are introduced and referred to in ten recent English for General Purposes (EGP) textbooks, at the intermediate and advanced levels. Methodologically speaking, the analysis will mainly be carried out automatically by querying the TeMa corpus, a corpus of ***Te****xtbook* ***Ma****terial (Meunier and Gouverneur forthcoming). After giving a general overview of the use of the two terms in the whole corpus, a detailed analysis of the ten textbook series will be provided. First, a purely quantitative analysis will provide information on the occurrences of the metalinguistic terms collocations and idioms. Second, the contexts of use of those two terms will be examined. Third, I will investigate whether the two terms are defined or explained be it in the student's book, the workbook or the teacher's book. The next step will consist in matching the metalinguistic terms and the lexical items they refer to in order to cast light on what the textbooks actually refer to when using the terms 'collocation' and 'idiom'. Finally, comparisons will be drawn between the metalanguage used at the intermediate and advanced levels in order to identify potential level-specific differences.*

Keywords: collocations, idioms, metalanguage, textbook corpus, EFL

Metalanguage research

Metalanguage, also called metatalk (Faerch 1985), has commonly been defined as "language used to analyse or describe language" (Johnson & Johnson 1998, Basturkmen et al. 2002). Metalanguage is used in three different types of instructional contexts:

(1) in reference tools, such as linguistic encyclopaedias, grammars, dictionaries, which deal with language as an object;

(2) in teaching materials, notably in the guidelines, to explain the tasks to be performed and to refer to the targeted linguistic units;

(3) in the classroom, by teachers and/or learners (Basturkmen et al. 2002).

The focus of this paper is on the second type of instructional context, namely metalanguage used in teaching material. The importance of that type of metalanguage in language learning has been emphasized by Widdowson (2003: 135-136), who claims that it is "designed to teach language" and "to engage learners in the process of learning". According to him, this is done essentially by providing two types of information: *explanation* and *exemplification*.

Prior research on metalanguage

Only a handful of researchers have investigated metalanguage in Instructed Second Language Acquisition (ISLA) and classroom instruction so far, and most of them have focused on grammatical metalanguage (see Borg 1999, Brufit et al. 1996, Lyster and Ranta 1997, Seedhouse 1997). One study, however, by Basturkmen et al. (2002), is of particular relevance in the framework of this paper. It analyses the metalanguage used by teachers and students

[36] Céline Gouverneur is a research fellow at the Université catholique de Louvain, Belgium. She works at the Centre for English Corpus Linguistics where she is currently working on her PhD thesis. Her research focuses on phraseology in foreign language learning and teaching, and more precisely on the treatment of word combinations in EFL textbooks. In the framework of her thesis, she has built up the TeMa corpus, a corpus of **Te**xtbook **Ma**terial, on which her research is based.

Formatada: Tipo de letra: Negrito, Inglês (Reino Unido)

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Avanço: Esquerda: 0 cm, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Avanço: Esquerda: 0 cm, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 2, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin, Manter com o seguinte

Formatada: Tipo de letra: Itálico

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado

during focus on form episodes (FFEs), i.e. in teaching contexts similar to those examined in this paper. The results show that the main focus of the learning activity, i.e. on form or on meaning, has significant influence on the amount and type of metalanguage used. Metalinguistic terms are more frequent and more technical in activities with a strong focus on form, while they appear to be scare and less technical in learning activities which focus exclusively on meaning. The present paper studies metalanguage used in form-focused learning episodes, more particularly here in vocabulary exercises. We could therefore expect the terms used to refer to the targeted lexical items to be significantly frequent and technical in nature. This hypothesis will be tested in the study.

Phraseological metalanguage

Using pedagogically-appropriate metalanguage is no easy task. As rightly pointed out by Widdowson (2003: 142), metalanguage has to be suitable to the learner's level to effectively enhance language learning, for, "if the way language is prescribed, exemplified, and explained does not correspond with the way learners actually learn, then it surely becomes an imposition which, far from facilitating the learning process, actually makes it more difficult."

The task is particularly challenging when it comes to word combinations. The last three decades have witnessed revived interest in phraseological research in a number of interrelated linguistic subdisciplines, such as corpus linguistics, psycholinguistics, second language acquisition, lexicology and lexicography, language teaching, material design and NLP tools design. Such scientific enthusiasm has resulted in the emergence of a profusion of terms which all refer to word combinations, though from a variety of perspectives. Consequently, phraseology has become "a field bedevilled by the proliferation of terms and conflicting uses of the same term" (Cowie 1998: 210), so that textbook writers have a multitude of possible options to choose from.

Zooming in on collocations and idioms

Within the phraseological spectrum, collocations and idioms are probably the two multi-word units which have been most researched. Linguists belonging to what has been called the traditional approach to phraseology were particularly fond of idioms. Following the Russian tradition, researchers such as Cowie (1988, 1994), Burger (1998), Gläser (1998) or Mel'čuk (1995, 1998) have attempted to set up defining criteria for classifying idioms, and more importantly, distinguishing them from other types of multi-word units. Idioms were also dealt with extensively by Anglo-American lexicographers (see for instance Moon 1998). While collocations were also dealt in the traditional approach to phraseology, notably by Palmer (1933) or Cowie (1988, 1994), it has really become a buzzword since the advent of the distributional or frequency-based approach to phraseology.

Researchers interested in phraseology, be they linguists, corpus linguists, or psycholinguists, have soon convinced teaching specialists of the relevance of integrating word combinations in the foreign language curriculum. Phraseology can no longer be said to be absent from the learning and teaching scene. Most ELT textbooks nowadays claim that they deal with (and sometimes even focus on) 'collocations' and 'idioms' (Meunier and Gouverneur 2007). Nevertheless, while a number of studies have attested the increasing importance of phraseology in foreign language teaching materials (Koprowski 2005), little research has been conducted into the metalanguage used by textbook writers to introduce phraseology to learners. The aim of this paper is to study the way collocations and idioms are introduced and referred to in ten recent English for General Purposes (EGP) textbooks, at the intermediate and advanced levels.

The TeMa Corpus

The data analysed in this study comes from the guidelines subcorpus of the TeMa corpus (see Meunier and Gouverneur forthcoming for a detailed description of the corpus). When available, up to four different types of textbook language have been subsumed under the term *guidelines*:

(1) the headings of the vocabulary section or of the vocabulary exercise

e.g. Phrases with do (2214)

Wordspot Idioms with laugh, cry and tears (2114, 2124)

All's well that ends well! Idioms and collocations (5114)

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

(2) the instructions to the exercises

e.g. Match the phrases in A with their opposites in B. (2214)

Use one of the above collocations in each of the following sentences. Sometimes you will need to change the form of the words in the collocation. The first one has been done for you as an example. (9114)

(3) the introductory comments on the relevance of the focus of the exercise

e.g. The importance of collocations

A collocation is when two words (or sometimes more than two) are seen together frequently. In English, these are very common, and it is a sign of a good English speaker to be able to use collocations well. There are several exercises to do with collocations in this book. Decide whether the collocations (in bold) in the sentences below are possible or not. Use a good English-English dictionary if you have one.

All these extracts can be considered as constituting an integral part of the message material writers address to learners before they do the exercises. The guidelines compiled in the subcorpus are all linked to *focus on forms (FoFs)* episodes, i.e. exercises which are language-focused and which focus explicitly on vocabulary and/or multi-word units more particularly. The exercises were selected on the basis of the textbook writers' description of the syllabus content and organisation in terms of vocabulary learning and practice. Table 1 provides a breakdown of the guidelines subcorpus, in number of words per volume: student's book (SB) and workbook (WB).

Textbook	**Intermediate**		**Advanced**		**Total**
	SB	**WB**	**SB**	**WB**	
English Panorama	/	/	9,335	/	**9,335**
Clockwise	4,191	492	3,263	819	**8,765**
Matters	2,320	1,422	2,066	2,494	**8,302**
Cutting Edge	3,448	1,131	2,254	756	**7,589**
New Cambridge	1,435	1,272	3,230	/	**5,937**
Inside Out	2,210	785	1,947	966	**5,908**
New Headway	1,319	438	1,712	1,471	**4,940**
Advance your English	/	/	1,839	1,033	**2,872**
Accelerate	642	436	193	1,306	**2,577**
Initiative	/	/	2,489	/	**2,489**
All textbooks	**15,565**	**5,976**	**28,328**	**8,845**	**58,714**
	21,541		**37,173**		

Table 1: Breakdown of the guidelines subcorpus

Quantitative analysis

The first step in the corpus analysis consisted in extracting the relevant metalinguistic terms from the guidelines. A wordlist of the whole guidelines corpus was drawn up using WordSmith Tools 4 (Scott 2004). Given the small

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: Negrito, Inglês (Reino Unido)

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

size of the corpus (i.e. 58,714 words), the minimum frequency level was set to one occurrence to ensure that any relevant metalinguistic term could be detected. The resulting wordlist contained 3,123 tokens and was lemmatised. *Collocation* and *idiom* qualified for inclusion in our bank of terms as technical terms, as opposed to non-technical terms (Basturkmen et al. 2002: 5). Non-technical items such as *words*, *expressions* or *phrase* represent the large majority of metalinguistic terms used, with technical words representing 20% of the terms.

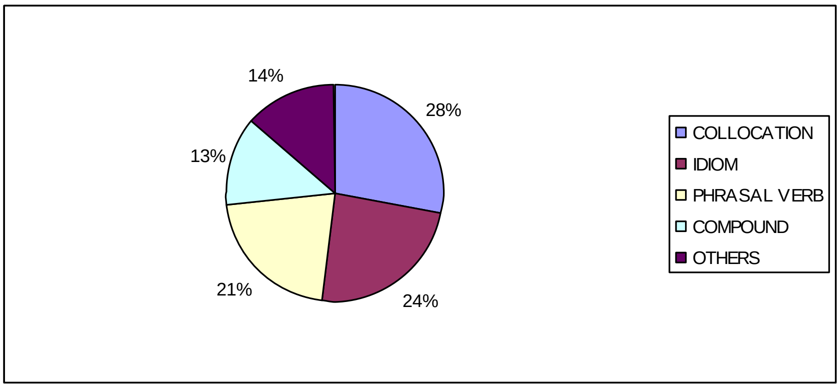


Figure 1: Proportions of technical terms

Figure 1 shows the proportions of the various technical terms encountered. The percentages show that *collocation* and *idiom* are the two most frequent lemmas used, followed by *phrasal verb* and *compound*. The last portion is shared among the remaining explicit terms: *proverb*, *saying*, *metaphor*, *cliché*, *simile*, each representing less than 3%. These preliminary figures illustrate the importance of collocations and idioms in present day EFL textbooks. Extreme caution is however required at this stage as figures provide no insight into the exact use of those terms.

Distribution across textbook series

A second stage consisted in observing the distribution of the terms idioms and collocations in the ten textbook series. This was done using relative frequencies.

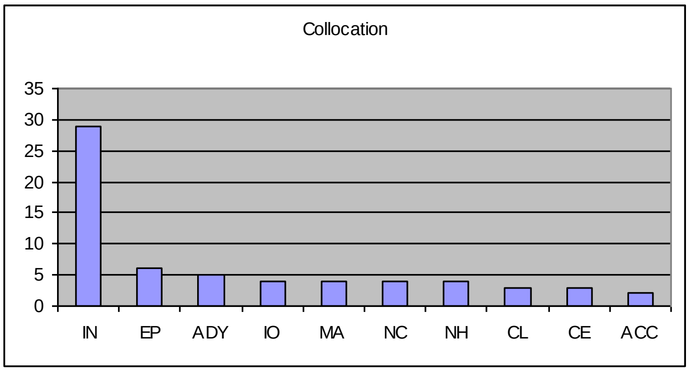


Figure 2: Distribution of the lemma *collocation* across textbooks

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

At first sight, it appears from Figure 2 that the term collocation is almost evenly distributed across the various series (presented here with their codes), except for Initiative, which seems to contain a far greater proportion of the term *collocation.*

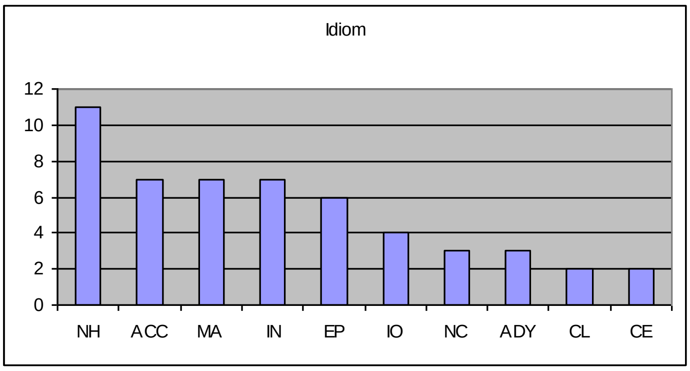


Figure 3: Distribution of the lemma *idiom* across textbooks

Figure 3, which represents the distribution of the lemma *idiom* across textbooks, displays a more varied and contrasted picture.

Contexts of use

The first step in the qualitative analysis consisted in identifying the contexts in which the terms are used. This was done through cluster and concordance analyses for both terms. They revealed that the terms *idiom* and *collocation* are used in the three different kinds of context presented above: section headings, exercises guidelines, and metalinguistic comments.

As far as headings are concerned, clusters mainly reveal that the term *collocation* is often preceded by the structure it consists of, e.g. *verb/noun collocations, adjective + noun collocations, verb-adverb collocations.* The term idiom is almost always associated with another type of multi-word units, such as for instance *Idioms and fixed phrases, Phrasal verb or idiom, Idioms and sayings.*

Exercise guidelines: matching metalinguistic terms and lexical items

The next step in the analysis consisted in matching metalinguistic terms used in the guidelines to the exercises and the lexical items they actually refer to. Both inter-textbook and intra-textbook analyses were carried out. The inter-textbook analysis revealed that the terms *collocation* and *idiom* were used in a similar way across most textbook series. Collocations were commonly perceived as frequently occurring (two-) word combinations with a certain degree of fixedness and collocability restrictions. The lexical items referred to as *collocations* were both lexical and grammatical collocations (Lewis 2001). Idioms were mainly defined in terms of their non-compositional meaning and fixedness of form. The lexical items labeled *idiom* were shown to be mainly figurative idioms. Both *collocation* and *idiom* were however also frequently used as cover terms referring to a variety of word combinations. Improvements can be suggested in this respect. The intra-textbook analysis revealed that the ten textbook writers follow a particular logic in their use of metalinguistic terms and make their own metalinguistic choices, mainly according to proficiency level. All textbook writers were shown to have recourse to too many non-technical terms whilst the use of technical terms would have been appropriate.

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Metalinguistic comments

In the learning and teaching tips, the terms *collocation* and *idiom* often appear in definitions and metalinguistic comments which textbook authors include before the exercises. Figures 4 and 5 display a sample of the definitions and comments available for the two terms.

1/ Clockwise Intermediate Teacher's book
Word combinations (or 'collocations') are words that frequently go together.
Collocations are words that frequently go together, e.g. heavy smoker, get married, in particular. Some words need a companion word to be used in certain ways, for example can't + afford, look + at (and other dependent prepositions), phrasal verbs (get up).
2/ Cutting Edge Intermediate Teacher's book
collocations – common word combinations – including:
nouns + verbs (work long hours, have a drink)
adjectives + nouns (old friends, good news)
adverbs + verbs (work hard, will probably)
verbs + prepositions/particles, including phrasal verbs (think about/grow up)
adjectives + prepositions (famous for, jealous of)
other combinations of the above (go out for a meal, get to know)

3/ New Headway Advanced teacher's book
Adverb collocations. Adverbs modify verbs and adjectives. Often usage has resulted in some adverbs collocating specifically with certain verbs and adjectives. For example, we say deeply worried and not sorely worried. This is because there is a semantic link between the adverb and the verb/adjective. Emotions can be deep, so we say deeply affected, deeply regret. Similarly, there are semantic links with collocations such as *freely admit*, *desperately anxious*, *highly recommend*. While some adverbs collocate with some verbs or adjectives because of a link in meaning, with many others there are no rules that dictate why certain combinations are possible or not. Here are a few possibilities. Extremely angry, difficult, important, sorry…. These collocations can only be learnt through memorizing and practising.

4/ English Panorama Student's book
Whenever you read something English, you learn something, consciously or unconsciously, about collocations or words that have a close association with each other.
The strength of collocations becomes clear when we look at phrases which are not collocations. We do not say thawing pot or mixing pot, for instance, only melting pot. Note that phrases like thawing pot or mixing pot might be possible for a particular humorous or literary effect, but they are unlikely to be used in normal speech or writing; in normal speech and writing we tend to stick to standard collocations.
Often words may be used in a range of contexts but they acquire a more specialised meaning when collocated in a particular way or when used in a particular context.

5/ Initiative Student's book
The importance of collocations A collocation is when two words (or sometimes more than two) are seen together frequently. In English, these are very common, and it is a sign of a good English speaker to be able to use collocations well. There are several exercises to do with collocations in this book.
There are many types of collocations in English. In the box below, there are nine types:
1 adjective + noun
2 verb + noun
3 noun + verb
4 verb + adverb
5 verb + preposition
6 noun + adjective
7 adjective + adjective
8 adverb + adjective
9 adjective + preposition

Figure 4: definitions and comments for collocations

Formatada: Justificado, Avanço: Esquerda: 1,25 cm, Espaçamento entre linhas: simples

Formatada: Normal, Avanço: Esquerda: 1,25 cm, Espaçamento entre linhas: simples

Formatada: Justificado, Avanço: Esquerda: 1,25 cm, Espaçamento entre linhas: simples

Formatada: Tipo de letra: 10,5 pt, Francês (França)

Formatada: Normal, Avanço: Esquerda: 1,25 cm, Espaçamento entre linhas: simples

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

1/ Matters Intermediate student's book

Idiomatic expressions are a combination of two or more words which go together and have a special meaning. (4214)

2/ Accelerate Advanced Student's book

Fixed order idioms

A number of expressions contain two nouns which come in a fixed order. It would be totally unnatural to change the order of the nouns, even though the meaning would not be affected, eg man and wife. (0123)

3/ English Panorama Student's book

Vocabulary: idioms

Idioms have their origins in various areas of life and sometimes a knowledge of the origin can help you to understand – and remember the meaning of the idiom. For example, to be blinkered means to have a very narrow view. The idiom originates in horse riding where blinkers are leather squares fixed to the sides of a horse's eyes to prevent it form seeing sideways.

Figure 5: include caption

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Figures 4 and 5 illustrate how definitions vary in length and details according to textbook series and level, with advanced textbooks presenting finer aspects of the phenomena. We also observe that some textbook writers prefer including comments in the teacher's book while some others address their metalinguistic comments to learners.

Conclusions

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

The study conducted in this paper has shed light on the actual use of the terms *collocation* and *idiom* in well-known EFL textbooks. The results obtained demonstrate that material writers are aware of the importance of integrating collocations and idioms in the language syllabus. Whilst some of them use the technical terms in an efficient way when defining and explaining collocations and idioms, many still have recourse to general terms (such as *expressions* or *combinations*) where the informed use of a technical term would enhance learning. Others still use idiom and collocations interchangeably or as a superordinate term to refer to a wide range of multi-word units. The study has yielded encouraging results and opened ways for further research avenues.

Formatada: Normal, Justificado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

References

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Inglês (Reino Unido)

Formatada: Inglês (Reino Unido)

Basturkmen, H., Loewen, S. and Ellis, R. 2002. "Metalanguage in Focus on Form in the Communicative Classroom." In *Language Awareness* 11/1: 1-13.

Borg, S. 1999. "The use of grammatical terminology in the second language classroom: A

qualitative study of teachers' practices and cognitions." In *Applied Linguistics* 20: 95–126.

Brumfit, C., Mitchell, R. and Hooper, J. 1996. 'Grammar', 'language' and 'practice'. In *Teaching and Learning in Changing Times*, M. Hughes (ed.). Oxford: Blackwell.

Cowie, A.P. 1994. "Phraseology." In *The Encyclopedia of Language and Linguistics* Asher, R. (ed.) Vol.6. Oxford and New York: Pergamon, 3168-3171.

Cowie, A.P. 1998. “Introduction”. In *Phraseology: Theory, Analysis and Applications*, Cowie, A.P. (ed.) Oxford University Press, Oxford, 1-20.

Faerch, C. 1985. “Metatalk in FL classroom discourse”. In *Studies in Second Language Acquisition* 7: 184-199.

Gläser, R. 1998. “The stylistic potential of phraseological units in the light of genre analysis.” In *Phraseology. Theory, Analysis, and Applications*, Cowie A.P. (ed.) Oxford University Press: Oxford, 125–143.

Granger, S. and Paquot, M. in press. “Disentangling the phraseological web”. In *Phraseology: An Interdisciplinary Perspective*, Granger S. and F. Meunier (eds) Amsterdam & Philadelphia : Benjamins.

Johnson, K. and Johnson, H. (eds) (1998) *Encyclopaedic Dictionary of Applied Linguistics: A Handbook for Language Teaching*. Oxford: Blackwell.

Koprowski, M. 2005. “Investigating the usefulness of lexical phrases in contemporary coursebooks.” In *ELT Journal* 59/4: 322- 332.

Lewis, M. 2001. *Teaching collocation*. LTP.

Lyster, R. and Ranta, L. 1997. “Corrective feedback and learner uptake: Negotiation of

form in communicative classrooms.” In *Studies in Second Language Acquisition* 19: 37–66.

Mel’čuk, I. 1995. “Phrasemes in language and phraseology in linguistics.” In *Idioms: Structural and Psychological Perspectives*, Everaert, M., E.J. Van der Linden & A. Schenk (eds.) Hillsdale: Lawrence Erlbaum Associates, 167–232.

Melcuk, I. 1998. “Collocations and Lexical functions.” In *Phraseology. Theory, Analysis and Applications*, Cowie, A. (ed.) Oxford. OUP.

Meunier, F. and Gouverneur, C. forthcoming. “New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material.” In *Corpora and Language Teaching*, K. Aijmer (ed.). Benjamins.

Meunier, F. and Gouverneur, C. 2007. “The treatment of phraseology in ELT textbooks.” In *Corpora in the Foreign Language Classroom. Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC6), University of Granada, Spain, 4-7 July, 2004. [Language and Computers Series 61]*, Encarnación H., Quereda L. and Santana J. (eds) Amsterdam/New York: Rodopi, 119-139.

Palmer, H.E. 1933. *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

Scott M. 2004. *WordSmith Tools 4*. Oxford: Oxford University Press.

Seedhouse, P. 1997. “The case of the missing ‘no’: The relationship between pedagogy and

Interaction”. In *Language Learning* 47: 547–583.

Widdowson, H. G. 2003. *Defining Issues in English Language Teaching*. Oxford: Oxford University Press.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Formatada: Inglês (Reino Unido)

Formatada: Inglês (Reino Unido)

Formatada: Tipo de letra: Negrito, Inglês (Reino Unido)

Formatada: Normal, Centrado, Nível 1, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

LINGUISTIC DETERMINANTS OF ENGLISH FOR ACADEMIC PURPOSES

Christoph Haase[37]

Abstract

Formatada: Normal, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

This contribution will introduce the application of statistical techniques to the quantitative evaluation and profiling of academic texts in English. Establishing standards for good academic writing practice is notoriously difficult for practitioners as well as for teachers of academic writing classes at university level. A major problem is the appropriate balance between a subjective display of the individual research and an objective display of the results obtained. The problem is exacerbated by the fact that there are no unique and specific linguistic markers of this balance even though some corpus-based suggestions were made by Hyland, 1994 or Aijmer, 2005. Our approach represents the outcome of a corpus-based research project in which specialized academic texts are juxtaposed with popular academic texts. The main difference measured here is thus not in content but in mode of display on different levels of subjectivity and objectivity.

The talk introduces a new parallel corpus to the study of English for Academic Purposes (EAP). It tries to connect linguistic approaches of a classification of point-of-view markers with normative aspects of teaching EAP. Furthermore, our study tries to assess the potential of the texts collected and the tools devised for developing a model of academic writing. For that end, a corpus with texts from diverse natural sciences (physics and biosciences) has been compiled. Data from a preliminary study of hedge expressions in academic and popular-academic texts is investigated and subjected to qualitative as well as quantitative analysis. We will define a quantifiable feature of hedge expressions – propensity- and a quantifiable feature for semantic complexity – semantic depth. The talk includes the presentation of a corpus that is fully annotated and rated for expressions of subjectivity.

Finally, the contribution makes suggestions how to incorporate this theoretical insight into academic writing course modules (cf. Chambers 2005).

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Keywords: Corpus linguistics, English for Academic Purposes, Hedge expressions, Stance, Vagueness in language, Academic writing

Linguistic determinants in the academic discourse situation

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

A common goal in the study of academic discourse is the linguistic evaluation and exemplification of cultural determinants of academic writing. A basic assumption shared by many linguists working in the Cognitive linguistic paradigm is that the different use of language in different societies is due to systematic cultural differences. These systematic differences reflect cultural values or hierarchies of those values which can be explained via independent determinants. The larger goal of the study presented here is to investigate these determinants using large collections of academic texts in English from different cultural backgrounds under application of quantitative methods. The study further applies these results to enhance standards of academic writing in the respective regions. Therefore an initial hypothesis could be that culture-dependent hedges or markers of subjectivity can be found through in-depth analysis of large numbers of lexical items, e.g. in text corpora.

The problem of identifying these items is the first methodological point that this study tried to solve. It thus combines in a novel way approaches from lexicostatistics and re-applies its results to a subsequent semantic analysis.

[37] Christoph Haase studied at first physics, later English and German with particular focus on linguistics at the Ernst-Moritz-Arndt University in Greifswald, Germany. His studies abroad involved generative linguistics at the University of Oviedo in Oviedo, Spain and cognitive science as a visiting scholar at Carleton University in Ottawa, Canada. Christoph came to Chemnitz in 2000 to work on the Internet Grammar. His research interests revolve around cognitive linguistics, lexical semantics, universals and typology, grammaticalization and the philosophy of language.

Since all fundamental human concepts can be considered innate, there is a strong universalist point of view in this research. Therefore, what is needed is not only a list of linguistic expressions of subjectivity but also combinations of these expressions.

In most analyses of the lexical semantics of the identified items there is considerable lack in methodological rigor due to the inherent vagueness of the semantics of the conceptual categories. This problem is further exacerbated when larger linguistic objects like texts are concerned, especially large quantities of texts in the corpora suggested above. A contrastive analysis of cultural determinants needs therefore to focus on linguistic markers relevant for the expression of subjectivity. Finding these markers however is not a matter of simple frequency counts, as commonly employed in corpus linguistics. Frequency alone is not a satisfying indicator to identify a culturally relevant distribution of these markers.

Hedge expressions

Hedges are a metaphorical device devised by Lakoff in order to lexicalise properties of delimiting the scope of an utterance via vagueness i.e. they distance the speaker from the utterance, blur quantities, attributes and specifications given in the utterance and relativize notions of truth. The classic example are the Lakoffian hedges of *sort of, kind of, like* etc. proposed in his seminal article on hedges in 1973. Hedges can therefore be used to estimate the adherence or commitment of a speaker or producer of a text to his/her utterance, the amounts, causes and applications in question and they can distance the listener to fully commit to the semantic content of the utterance.

In a discursive situation, speaker A and listener B act in a contract of a diffusion of knowledge from A to B. The contract they enter into is that A and B both share the knowledge that A knows that B does not know everything that A knows (about X). Taking recourse to hedges enables therefore both to cross borders which are primarily borders of knowledge. Speaker A uses conventionalized strategies to express himself/herself comprehensibly. Thus, hedging is a means to modify the propensity of a statement, with propensity defined as the degree of probability of a statement to hold true.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Hedge forms and functions

Hedges follow primarily pragmatic lexicalization patterns and cut therefore across syntactic classes, i.e. there is no definite, automatically identifiable class of a hedge. In a refined system suggested by Martín-Martín (2005:78), only stance adverbs, approximators, approximatives and fuzzy quantifiers exist.

In the diffusion of knowledge in academic prose hedges therefore serve multiple purposes. At least three cases apply:

1) The author A of an academic text knows that reality has more than “ideal” cases of e.g. Newtonian mechanics because it is blurred at micro-levels of description. He/She can therefore hedge an utterance by saying “Ideally,…” etc. A possible caveat is that the intended listeners share this knowledge and do not need the explicit hedge marking.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

2) A different case occurs when the author B of a popular academic text (who at the same time is an informed listener of 1) knows that his/her listeners do not share the knowledge of 1) so the “hedgy” precision of 1) has an entirely different function i.e. the hedginess here is used as simplification.

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

3) The layperson listener C of 2) has a contract that the academic contents will be processed and virtually spoon-fed via hedges and metaphors. C knows that C does not know what A or B know so C expects lexicalisation patterns that coincide with C’s knowledge by transformation of source domains (specific knowledge) to target domains (generic knowledge).

Hedging enables it therefore both to cross borders which are primarily borders of knowledge. It is further an instrument of protection in order to distance themselves from the outcome and results of an opinion advanced (cf. Makaya & Bloor 1987: 59).

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

The Corpus of Scientific and Popular Academic English (SPACE)

Content and Structure

The study was carried out by using authentic texts from the SPACE corpus compiled at Chemnitz University of Technology. A few brief details should suffice here. A more comprehensive description can be found in Schmied, 2007 and Haase, 2008.

The corpus has a binary structure of a) academic texts and b) popular-academic texts. All texts in a) have a direct correlate in b) because science journalists sift through new and groundbreaking publications in major preprint servers in order to preprocess this information for the academically interested layperson. The academic texts were all retrieved from three preprint servers for academic publications, *arXiv.org*, *Proceedings of the National Academy of Sciences* and *Public Library of Science - Medicine* (PLoS) where these texts are essentially published freely and without peer-reviewing. All popular-academic texts were taken from the *New Scientist* magazine.

subcorpus	**descriptors**	**words**
arXiv	physics, astrophysics, theoretical computer science, quantum mechanics	161,864
New Scientist – physics		40,694
Proceedings of the National Academy of Science (PNAS)	biochemistry, genetics, genetical engineering, microbiology	267,105
New Scientist - biosciences		30,499
Public Library of Science – Medicine (PLoS)	medicine, virology, clinical psychology, public health	217,254
New Scientist – medicine		17,050
total		**734,466**

Tab. 1: Scope and structure of the SPACE corpus

The word counts show the imbalance between the academic and the popular academic texts; the former, representing the outgrowths of real research, are always longer, the popular texts can be long cover stories or very short research news items.

Academic vs. popular-academic strategies: Corpus evidence

A brief glance at the texts makes obvious that the academic texts are accessible to specialists in their fields but to a large extent exclude laypersons simply due to the terminology used in these texts. Standard strategies of the popular-academic versions are therefore

1) syntactic compression and semantic simplification, (e.g. compare the titles of one text used in test B: the academic text is titled “Topical DNA oligonucleotide therapy reduces UV-induced mutations and photocarcinogenesis in hairless mice” (0090PN) whereas its popular-academic counterpart is titled “Suntan lotion primes the skin's defences” (0090NS)).

2) a “lack of lexical differentiation” as observed by Lorenz (1998: 58) in second language learners and

3) an overuse of stylistic devices like amplifiers (*completely*, *absolutely*) and boosters (*very*, *highly*, *immensely*). However, all these strategies are also applicable to the study of texts written by nonnative speakers. In that sense, a similarity between authors of popular academic texts and second language learners can be proposed in which a lack of lexical differentiation is mirrored by lack of scientific differentiation. The conjecture for this study is thus to investigate second language learners’ (e.g. German and Czech learners’) behavior in the processing of academic and popular-academic texts.

In a previous project, the entire corpus was lexically analyzed in its POS (part-of-speech)-tagged version under use of the tool Treetagger and the Penn Treebank tagset. In a semantic analysis, the corpus was manually

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

annotated for author commitment. This means that hedge expressions like *probably, normally, suggests that, some evidence for* etc. were assigned a value between 1 and 10 by a human processor with 1 indicating an extremely low probability/propensity and 10 indicating certainty or extremely high probability/propensity.

An example hedge tag looks as follows:

partially_AV0_M_6

POS tag:	AV0 (adverb)
positional marker:	M for medial
propensity score:	6

A short popular-academic text that was used for testing with learners is shown below with hedge tagging:

0090NS Suntan lotion primes the skin's defences

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

It **might be_VM_M_4** possible to develop suntan lotions that kick-start the skin's protective mechanisms against cancer before you hit the beach. The key ingredient **could be_VM_M_6** a fragment of DNA just two bases long, called a TT dimer, that mimics one of the signs of DNA damage from ultraviolet light. Barbara Gilchrest's team from Boston University and colleagues in the Netherlands exposed hairless mice to a mild ultraviolet radiation, the equivalent of half an hour of afternoon sun. They found that genes involved in DNA repair were **more_AV0_M_8** active in mice that had the TT dimmer rubbed on their skin before exposure. And only 22 per cent of the treated mice developed skin cancers after 24 weeks compared with 88 percent of untreated mice. (Proceedings of the National Academy of Sciences, DOI:10.1073/pnas.0306389101). People who want a tan may **not even need to_VM_M_5** go out in the sun. Mouse skin does not produce melanin but earlier tests on guinea pigs **suggest that_VV_M_7** the TT dimer also triggers the tanning response. The team has not yet begun testing it on people.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Hypotheses on subjectivity

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

An overall hypothesis that was confirmed in previous projects was that the observed and intuitive differences between academic and popular academic texts can be quantified. This quantification is easy in all lexical classes via automatic annotation but difficult for semantic analysis. However, the manual processing suggests that quantification is also possible

1) in terms of the propensity of their hedge expressions and subjectivity markers and

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

2) in terms of the semantic depth of their content words. Previous studies have attained similar goals: Stable quantitative results were obtained e.g. by Hidalgo by investigating indicators of writer stance, especially by measuring frequencies of modal adjectives and modal adverbs (Hidalgo 2006:126).

Furthermore,

3) Academic texts are less subjective and show less hedge markers

4) Popular-academic texts are more subjective and use more general content words

5) Academic text processing of non-native speakers reflects the difference between academic and popular-academic texts.

Conclusion

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

The data discussed here display trends that could be solidified with more research, quantitatively and qualitatively. Concluding, it can be summarized that especially the subjectivity in academic discourse finds diversified linguistic expressions. There are versatile determinants of the linguistic shape of commitments. This means that subjectivity can be lexicalized in different ways according to 1. origin (academic or popular academic), 2. readership (scientists or educated laypersons) and 3. English as a first or second language. The data investigated shows a) a higher semantic complexity of academic texts, b) corpus and analyzer can be used in teaching academic writing and that this can c) be used to sensitivize students to the pragmatic and semantic markers of academic English and lastly, it enables them to d) investigate and quantify the results of their own academic work.

References

Aijmer, K. 2005. "Evaluation and Pragmatic Markers." In *Strategies in Academic Discourse*, E. Tognini-Bonelli and G. Del Lungo Camiciotti (eds.). Amsterdam: John Benjamins, 83-96.

Chambers, A. 2005. "Integrating Corpus Consultation in Language Studies." *Language Learning & Technology* 9/2: 111-125.

Haase, C. 2008. "Subjectivity and Vagueness in Academic Texts. Scientific vs. Popular-Scientific English." *Topics in linguistics* 1: 45-52.

Hidalgo, L. 2006. "The Expression of Writer Stance by Modal Adjectives and Adverbs in a Comparable Corpus of English and Spanish Newspaper Discourse." In *Corpus Linguistics. Applications for the Study of English*, A.M. Hornero, M.J. Luzon and S. Murillo (eds.). Bern/Berlin: Peter Lang, 125-140.

Hyland, K. 1994. "Hedging in Academic Writing and EAP Textbooks." *English for Specific Purposes* 13/3: 239-256.

Lakoff, G. 1973. "Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts." *Journal of Philosophical Logic* 1: 458-508.

Lorenz, G. 1998. "Overstatement in Advanced Learners' Writing: Stylistic Aspects of Adjective Intensification. In *Learner English on Computer*, S. Granger (ed.). London: Longman, 53-66.

Makaya, P. and **Bloor, T.** 1987. "Playing Safe with Predictions: Hedging, Attribution and Conditions in Economic Forecasting." *Written Language* 55: 55-70.

Martín-Martín, P. 2005. *The Rhetoric of the Abstract in English and Spanish Scientific Discourse. A Cross-Cultural Genre-Analytic Approach*. Bern/Berlin: Peter Lang.

Schmied, J. 2007. "Complexity and Coherence in English Student Writing, Especially in Hypertext Learning Systems." In *Complexity and Coherence: Approaches to Linguistic Research and Language Teaching*, J. Schmied, C. Haase and R. Povolná (eds.). Göttingen: Cuvillier, 13-30.

Formatada: Centrado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Inglês (Reino Unido)

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

EXPLORING THE MARKING OF STANCE IN ARGUMENTATIVE ESSAYS WRITTEN BY EFL LEARNERS AND NATIVE SPEAKERS OF ENGLISH

Anna-Maria Hatzitheodorou[38]

Marina Mattheoudakis[39]

Abstract

*This paper examines the projection of attitude in learner corpora that contain essays in English written by Greek and American university students. Within a suggested framework for the analysis of stance, specific stance indicators are discussed: boosters, hedges, and attitude markers that show affect. The findings of the study reveal that Greek university students deploy different rhetorical conventions to indicate their stance from their American counterparts. Greek learners use more stance indicators than native speakers in argumentative writing. In addition, compared to native speakers, Greeks overuse boosters (*it is true that, it is a fact that*) and attitude markers (*unfortunately*), and underuse hedges; hence, their more emphatic writing. Factors that may influence the ways in which Greek learners structure their arguments are L1 and L2 instruction in argumentation, transfer from L1, and learners' development as L2 writers. It is suggested that L2 writing instruction should be informed by contrastive rhetoric.*

Keywords: learner corpora, argumentative writing, stance

Introduction

Previous research in learner English (Altenberg & Tapper 1998; Tankó 2004; Hatzitheodorou & Mattheoudakis in print, a.o.) has relied on grammatical categories to investigate how learners structure their discourse. This paper aims to adopt a more stance-oriented perspective to look at students' writing and investigate the interface between coherence and stance/evaluation (cf. Neff et al. 2004 and Leńko-Szymańska 2006). Since evaluation and coherence become more transparent in longer stretches of discourse, we chose to examine and compare stance indicators in Greek and American students' writing by relating them to more context so that conceptual and meaning relations among sentences would be highlighted.

Categories of stance

Hyland's (2005) as well as Biber and Finegan's (1989) research on stance deal with organization of discourse and presentation of evaluation. According to Hyland (2005: 176-77), stance comprises four main elements: (a) hedges, (b) boosters, (c) attitude markers, and (d) self-mention (Figure 1). As Hyland and Tse (2004: 169) and Hyland (2005: 178-180) point out, writers employ hedges to withhold full commitment to a proposition (e.g., *might, perhaps*) and boosters to emphasize the writer's certainty in the proposition (e.g., *definitely*). Attitude markers are used to pull readers into agreement (e.g. *unfortunately, surprisingly*). Finally, self-mention makes explicit reference to author(s), e.g., *I, we.*

[38] Dr. Anna-Maria Hatzitheodorou is a Teacher of English for Specific/Academic Purposes at the Centre for Foreign Language Teaching (CFLT), Aristotle University of Thessaloniki. She holds an M.A. in English from Ohio State University, Columbus, Ohio and a Ph.D. in Written Discourse Analysis (Comprehension and production of written discourse in a university EFL context) from Aristotle University of Thessaloniki. She teaches English for Specific/Academic Purposes at the Law School. She has presented her research work in several national and international conferences. Her main research interests lie in the areas of written discourse analysis, academic discourse and genre, corpora and their applications.

[39] Dr. Marina Mattheoudakis is an Assistant Professor at the Department of Theoretical and Applied Linguistics, School of English, Aristotle University of Thessaloniki. She holds an M.A. in T.E.F.L. from the University of Birmingham, UK and a Ph.D. in Lexicology (Problems related to Greek-English lexical loans) from Aristotle University of Thessaloniki. She teaches courses in second language acquisition and methodology of language teaching. She is one of the scientific coordinators of the teacher training courses held in the department. She has presented her research work in several national and international conferences. Her main research interests lie in the areas of second language learning and teaching, corpora and their applications.

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Direita: -0,59 cm

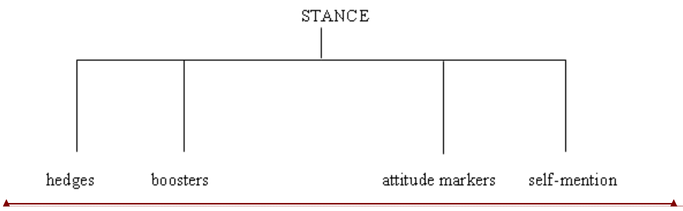


Figure 1. Categorization of stance features according to Hyland (2005)

Formatada: Tipo de letra: 10,5 pt, Italiano (Itália)

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: 10,5 pt, Italiano (Itália)

Formatada: Tipo de letra: (predefinido) Arial, 10,5 pt

Formatada: Centrado, Avanço: Primeira linha: 0 cm, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin, Sem controlo de linhas isoladas, Não ajustar o espaço entre texto asiático e texto em latim, Não ajustar o espaço entre texto asiático e números

Formatada: Tipo de letra: (predefinido) Arial, 10,5 pt, Negrito, Inglês (Reino Unido)

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

As we aim to account for instances of attitudinal stance not normally included in academic discourse, we enrich Hyland's framework with elements taken from Biber and Finegan's (1989: 98) model. In the latter, stance features are divided into two pragmatic functions, namely, *affect* and *evidentiality*. Affect includes both positive and negative markers expressing the author's personal feelings and attitude (e.g. *happily*, *sadly*) (cf. Ochs 1989). Evidentiality covers grammatical categories that express the author's certainty (e.g., *impossible*, *obvious*) or doubt (*perhaps*) (Figure 2).

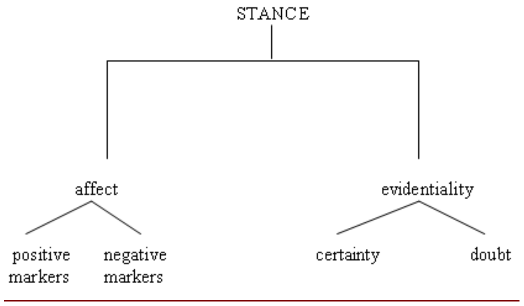


Figure 2. Categorization of stance features according to Biber and Finegan (1989)

Formatada: Tipo de letra: 10,5 pt, Italiano (Itália)

Formatada: Tipo de letra: 10,5 pt

The two stance frameworks presented in this section will need to be combined into a new model to better serve the purposes of our research (see section 5 below).

Aims and research questions

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

This paper has a two fold-purpose: to present an adapted framework for stance analysis as well as to focus on coherence and projection of stance in the essays of Greek advanced learners of English and American students. To this aim, relevant data from two corpora are presented and discussed within a contrastive rhetoric framework. The following research questions will be addressed:

(a) how do Greek and American students structure their arguments and indicate their stance?

(b) when and how do these groups of students adopt an assertive or a self-effacing attitude?

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

After we have sketched out a tentative profile of Greek students as writers, we will attempt to identify the factors influencing the rhetorical conventions that Greek students use.

Methodology: subjects and materials

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

The participants in this study were 176 Greek native speakers at the 3rd and 4th year of their university studies at the School of English, Aristotle University of Thessaloniki in Greece. The data used were drawn from the Greek Corpus of Learner English (henceforth GRICLE), which we compiled following the guidelines of the International Corpus of Learner English (henceforth ICLE) (Granger et al. 2002). GRICLE is the Greek written component of ICLE; the size of the corpus used for this study is 177,500 words. Each student was required to produce two argumentative essays of at least 500 words each on a given set of topics. Two other corpora were used as control of the native writer's norm: (a) the American collection of LOCNESS (Louvain Corpus of Native English Essays) (size of corpus: 149,580 words), and (b) the American collection of the PELCRA project (Polish and English Language Corpora for Research and Applications). The latter is a subcorpus compiled by Leńko-Szymańska (Leńko-Szymańska 2006) and includes argumentative essays written by American first- and second-year students (size of the American subcorpus: 25,467 words).

Proposing an adapted model of stance

In order to answer the research questions presented in section 3 above, we compared the writings of Greek and American students. For our analysis, we adapt the two frameworks of stance presented in section 2 above and propose an alternative categorization of stance features; this follows Hyland's categorization (i.e., hedges, boosters, attitude markers, self-mention), but further subdivides attitude markers into two distinct features: *affect* and *opinion*. We make use of Biber's affect category as this allows us to account for expressions that would normally be found in argumentation in general, but not in academic discourse, which is the focus of Hyland's model. The new sub-category created is that of opinion; this includes lexical items that introduce the writer's cognitive attitude to the proposition stated (e.g. verbs such as *I think*, *I agree*) (Figure 3). With regard to Biber's evidentiality, it does not need to be included in our model, as its features are covered by Hyland's categories of hedges and boosters.

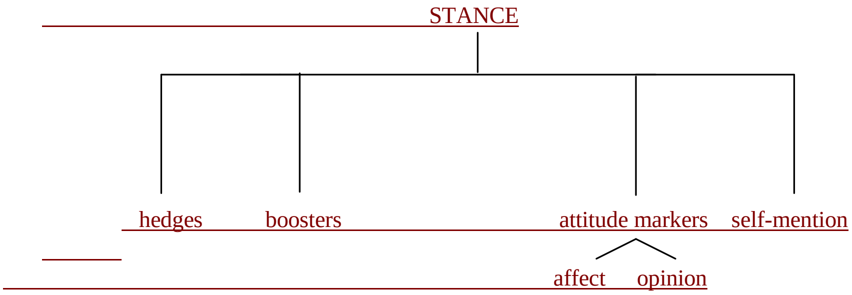


Figure 3. Categorization of stance features according to the model proposed in this study

This study examines the frequency and functions of specific hedges, boosters, and attitude markers showing affect evidenced in the Greek learner and native corpora; it also discusses these stance indicators in the light of contrastive rhetoric. Firstly, we carried out frequency counts of specific hedges, boosters, and attitude markers indicating affect that were expected to be found in argumentative writing. In our qualitative analysis, we selected those stance indicators whose occurrence frequency in the two corpora was markedly different, and we explored the grammatical accuracy and rhetorical functions they perform.

Results

Quantitative analysis

Boosters in GRICLE get the lion's share as their use is much more extensive than that of hedges and attitude markers (458 occurrences of boosters, 73 of hedges, 97 of attitude markers). Moreover, boosters are much more frequent in GRICLE than in the native corpora (458 vs 159 occurrences respectively). Conversely, hedging is a more common rhetorical choice in native speakers' writing (167 occurrences in the native corpora vs. 73 in GRICLE). Finally, attitude markers indicating affect are twice as many in GRICLE than in the native corpora. *Unfortunately* features extensively in GRICLE (83 occurrences vs 26 in the native corpora), while the rest of the attitude markers are either very rarely used or not used at all in either of the two corpora (e.g. *luckily, it is shocking, it is fortunate*, etc.) (Table 1).

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: Negrito, Itálico

Formatada: Tipo de letra: Itálico, Italiano (Itália)

Formatada: Normal, Nível 2, Espaçamento entre linhas: Múltiplo 1,1 lin, Manter com o seguinte

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Espaçamento entre linhas: Múltiplo 1,1 lin

Stance indicators	***GRICLE (177,490)***	***LOCNESS&PELCRA (175,047)***
Boosters		
Of course	153	34
It is true	71	5
No doubt/undoubtedly	63	13
Indeed	46	15
Certainly	37	22
It is obvious	32	9
It is a fact	21	1
Truly	14	24
Clearly	9	27
It is evident	8	1
It is clear	4	8
Hedges		
Perhaps	8	46
Likely	5	37
Possibly	8	22
Probably	47	52
It is possible	5	10
It is probable	0	0
Attitude markers-affect		
Unfortunately	83	26
Hopefully	4	13
Fortunately	4	2
Luckily	1	2
Happily	3	3
It is shocking	1	0
It is fortunate	0	0
It is surprising	0	0
It is amazing	1	1

Table 1. Comparative results regarding the frequency of stance indicators

Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada
Formatada

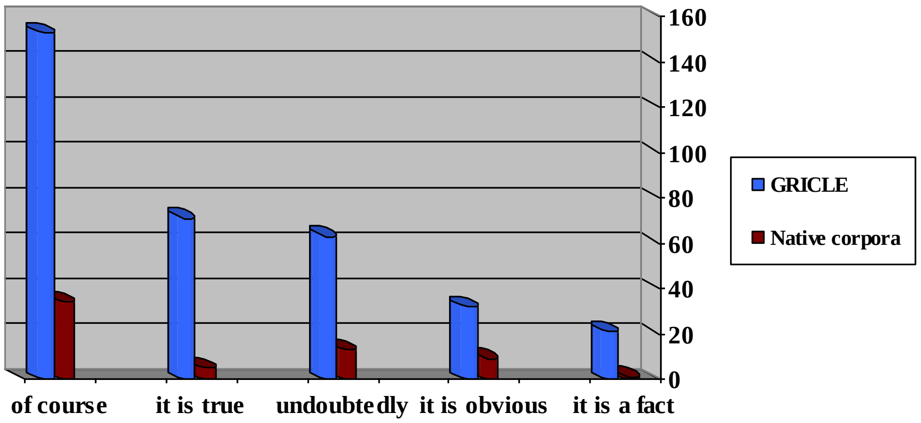


Chart 1. Specific boosters in GRICLE and the native corpora

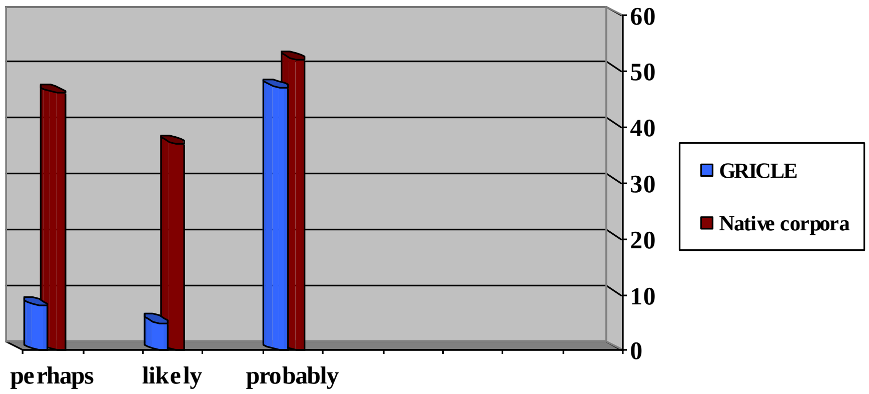


Chart 2: Specific hedges in GRICLE and the native corpora

Qualitative analysis

Boosters

Greek learners, compared to native speakers, show a tendency to be more emphatic in their argumentative writing by overusing boosters, and in general, by underusing hedges, they avoid mitigating their claims (cf. Hinkel 2002: 126).

With regard to boosters, Greek learners prefer lexical chunks to adverbs to emphasize their arguments. In general, there is grammatically correct use of boosters. Rhetorically, however, they often perform functions that diverge from conventional writing. Besides expressing the writer's certainty in the proposition, boosters in GRICLE are also used to perform the following functions:

(a) state commonly accepted ideas,

(b) project a personal opinion as an objective truth,

(c) introduce / agree with the topic, and

(d) provide emphasis.

Formatada: Normal, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Avanço: Pendente: 0,63 cm, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin, Tabulações: 1,27 cm, Listar tabulação

Formatada: Normal, Avanço: Pendente: 0,63 cm, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Hedges

Hedging features more often in the native corpora than GRICLE. This finding is in line with the Anglo-American rhetorical convention, whereby overstatements are generally discouraged so as to allow for alternative arguments to be expressed (cf. Hyland, 2005: 131-2). In general, hedges do not present particular interest because, apart from their numerical differences in frequency, they perform similar functions in both corpora.

Attitude markers

Regarding attitude markers, their use is equally limited in both corpora, the only exception being the adverb *unfortunately* (cf. section 6.1). What is worth noting is that this adverb collocates very strongly with *but* (1 in 4 occurrences) in GRICLE. The Greek translation equivalent is quite commonly used in Greek discourse; *αllá* ('but') and *dystychós* ('unfortunately') often co-occur to aid connectivity.

Discussion

The factors that seem to account for the differences presented above between GRICLE and the native corpora combine elements related to (a) learners' instruction in both English and Greek, (b) transfer from their native language, and (c) learners' development as L2 writers (cf. Granger 2004: 135).

Instruction in L1 and L2

State English language instruction in Greece spans over 7 to 9 years and is further intensified by learners' attending private language courses. This aims to enable learners to obtain language certificates because these are highly valued. As regards writing, learners are extensively exposed to and required to produce various text types. However, as the ultimate goal of learning the language is obtaining a language certificate at an early age, there is not adequate time for learners to master the L2 style of writing, which is very different to that of L1; this task becomes even more difficult due to Greek students' young age and lack of cognitive maturity. Alongside EFL instruction, the high school curriculum in Greece involves rigorous instruction in argumentation in Greek, as the writing of an argumentative essay is a required component of university entrance exams.

Therefore, when entering the English Department, Greek students have a depository of knowledge on how to write argumentation due to their exposure to the genre of argumentation for over 6 years of their schooling in their native language. However, intensive instruction in writing in English that would further enhance their writing expertise is not feasible at the university level due to large student audiences. Given the context of learning presented above, learners' written production, as we witness it in our corpus, is a combination of their exposure to L2 during both secondary and tertiary education with their knowledge of L1 writing acquired during their high school years; hence, the presence of both L2 and L1 styles of writing in the corpus.

Transfer from L1

In order to examine the possible influence of L1 on Greek learners' rhetorical choices, we looked at argumentative writing by Greek skilled writers in the Hellenic National Corpus (HNC). As far as boosters are concerned, we examined the translation equivalents of the lexical chunks *it is true that, it is a fact that, it is obvious that* that were very often used by Greek learners. Our analysis indicated that those chunks are used in Greek to perform similar functions to the ones performed in GRICLE. Thus, it is reasonable to assume that learners transfer rhetorical conventions from their native language. If these choices are culturally induced, it is possible that learners may be misled into believing that they can transfer Greek rhetorical conventions to L2 writing.

Such findings point towards the need to look at data in the light of contrastive rhetoric and to raise learners' awareness of differences between L1 and L2 rhetorical conventions. Contrastive rhetoric has developed into intercultural rhetoric research because of the growing emphasis in recent years on the social aspects of writing (Connor 2004); these become transparent if we look at both the national or ethnic culture and what Holliday (1999) has referred to as "small cultures". National culture would explain the rhetorical choices made by particular national or ethnic groups; an example would be the tendency of Greek writers to boost their statements. Small cultures,

Formatada: Normal, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Justificado, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: Itálico, Italiano (Itália), Kerning em 16 pt

Formatada: Normal, Justificado, Nível 1, Espaçamento entre linhas: Múltiplo 1,1 lin, Manter com o seguinte

Formatada: Normal, Justificado, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Normal, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin according to Holliday (1999: 237), refer to “small social groupings or activities wherever there is cohesive behaviour”; an example would be the English language learner culture. The small culture example may extend beyond national divisions and therefore the English language learner culture may include groups of people in various national cultures. National and small cultures, as both Atkinson (2004) and Holliday (1999) have claimed, interact and overlap in their manifestations. This means that language learners’ choices, when using the L2, will be influenced both by their national or ethnic culture as well as by the various small cultures to which they may belong, e.g., classroom culture, youth culture, etc.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Learners as developing writers

Turning now to our study, the small culture involves advanced learners of English who have been exposed to systematic instruction in argumentative writing for a substantial number of years and now in the final years of their studies attempt to become members of the academic community. The specific characteristics found in the writings of this social group are not only a result of their respective national culture but also the product of their development as L2 writers. It is important to bear in mind that Greek learners of English, as all foreign language learners, are what we would call “developing writers” (cf. Granger 2004). This means that their skills in academic writing are in a continuous process of development in both English and Greek and thus the choices they make, as these are witnessed in our corpus, may be necessary stages they have to pass through to reach writing expertise. Accordingly, Greek learners’ overuse of boosters may be a discourse strategy adopted to enable them to organize their thinking and enhance the force of their arguments.

Formatada: Justificado, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Justificado, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Pedagogical implications

In light of the findings of this study, we suggest that EFL instruction in Greece should be informed by the principles of contrastive rhetoric. Applying such principles in the L2 instructional context can help us raise both teachers’ and learners’ awareness of differences between L1 and L2 rhetorical conventions.

Teachers may need to be trained in recognizing similarities and differences in the ways stance is projected by L1 and L2 writers. Therefore, it is necessary to provide them with the opportunity to access both native and learner corpora as well as train them in using corpora as a teaching tool; in this sense, corpora may be used to detect differences between native and learners’ writing, to provide systematic feedback, as well as to design and adapt teaching materials according to learners’ needs (cf. Farr 2008: 39).

As for students, they need to become aware of the use and functions of stance indicators so as to become rhetorically literate. Knowledge of rhetorical conventions and discourse is part of students’ metalinguistic awareness and this is closely related to their cognitive maturity and level of L2 proficiency. If the development of writing skills is a combined result of exposure to written texts, explicit instruction and cognitive maturity, then we can confidently suggest that learners’ exposure to academic text types together with explicit instruction on L2 rhetorical conventions will enhance their L2 writing skills (cf. Neff et al. 2007). At the same time, such exposure is expected to foster their cognitive maturity.

Through our analysis of the students’ essays, it became transparent that what is urgently needed is for students to realize that stance indicators should fit into the particular context and also reflect conceptual and meaning relations between text parts. Since there existed gaps in students’ reasoning, we would argue that what is mainly needed is instruction focused on thinking patterns**,** as well as enriching and boosting students’ content knowledge.

Formatada: Tipo de letra: Não Negrito, Italiano (Itália)

Formatada: Normal, Justificado, Espaçamento entre linhas: Múltiplo 1,1 lin

While writing expertise is closely related to cognitive maturity, we believe that, since awareness of discourse is an ongoing process, it should start from a very early stage in L2 acquisition and learners should be given the opportunity to be exposed to discoursal features through various teaching – preferably corpus-based – materials. These may be corpus extracts and concordances that will allow learners to look at connectivity within longer textual chunks and sensitize them to the L2 rhetorical conventions. Such material may also be combined with awareness-raising, corpus-driven activities aiming to provide learners with training and practice in the recognition and use of semantic and textual relations. This is actually what Johns and King (1991: iii) have referred to as ‘data-driven learning’ (DDL). So far, DDL has mostly promoted the use of native speaker data in the classroom in order to expose learners to the authentic target language structures and patterns. However, Granger and Tribble (1998) have also suggested the combined use and comparison of native and learner data in the classroom. This comparison is expected to help learners notice the differences between their own and native speakers' forms in a predefined problematic area and thus improve their output (cf. Meunier, 2002).

In this paper, we have argued for the need to sensitize learners to how propositional content is organized and presented in the writing practices of English as an L2. If students are more consciously attuned to how elements of writing (purpose, audience, genre) determine linguistic and rhetorical choices, the acquisition of academic literacy is promoted.

References

Altenberg, B. and **Tapper, M.** 1998. "The use of adverbial connectors in advanced Swedish learners' written English." In *Learner English on Computer*, S. Granger (ed.). London: Longman, 80-93.

Atkinson, D. 2004. "Contrasting rhetorics/contrasting cultures: Why contrastive rhetoric needs a better conceptualization of culture." *Journal of English for Academic Purposes* 3/4: 277-289.

Biber, D. and **Finegan, E.** 1989. "Styles of stance in English: Lexical and grammatical marking of evidentiality and affect." *Text* 9/1: 93-124.

Connor, U. 2004. "Intercultural rhetoric research: Beyond texts." *Journal of English for Academic Purposes* 3/4: 291-304.

Farr, F. 2008. "Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users." *Language Awareness* 17/1: 25-43.

Granger, S. 2004. "Computer learner corpus research: Current status and future prospects." In *Applied Corpus Linguistics: A Multidimensional Perspective*, U. Connor and T.A. Upton (eds.). Amsterdam: Rodopi, 123-145.

Granger, S., Dagneaux, E. and **Meunier, F.** 2002. *The International Corpus of Learner English/Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. and **Tribble, C.** 1998. "Learner Corpus data in the foreign language classroom: form-focused instruction and data-driven learning." In *Learner English on Computer*, S. Granger (ed.). London & New York: Addison Wesley Longman, 199-209.

Hatzitheodorou, A-M. and **Mattheoudakis, M.** in print. "The Greek Corpus of Advanced Learner English (GRICLE): An electronic database of written discourse." To appear in the *Proceedings of the 30th International Conference on Functional Linguistics*, University of Cyprus, October 18-21, 2006.

Hinkel, E. 2002. *Second Language Writers' Text*. Mahwah, NJ: Lawrence Erlbaum.

Holliday, A. 1999. "Small cultures." *Applied Linguistics* 20/2: 237-264.

Hyland, K. 2005. *Metadiscourse: Exploring Interaction in Writing*. London: Continuum.

Hyland, K. and **Tse, P.** 2004. "Metadiscourse in academic discourse: A reappraisal." *Applied Linguistics* 25/2: 156-177.

Johns, T. and **King, P.** (Eds.) 1991. "Classroom concordancing." *English Language Research Journal* (New Series) 4. Special Issue. Birmingham: University of Birmingham.

Leńko-Szymańska, A. 2006. "The curse and the blessing of mobile phones – A corpus-based study into American and Polish rhetorical conventions." In *Corpus linguistics around the world*, A. Wilson, D. Archer and P. Rayson (eds.). Amsterdam-New York: Rodopi, 141-154.

Meunier, F. 2002. "The pedagogical value of native and learner corpora in EFL grammar teaching." In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung and S. Petch-Tyson (eds.). Amsterdam: John Benjamins, 119-141.

Neff, J.A., Dafouz, E., Gallego, J., Rica, J.P., and **Bunce, C.** 2007. "Academic literacy in university English studies in Spain: Incorporating grammatical and rhetorical accuracy." Paper presented at the *14th International GALA*

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Tipo de letra: Negrito, Itálico, Inglês (Reino Unido)

Formatada: Normal, Nível 4, Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Inglês (Reino Unido)

Formatada: Inglês (Reino Unido)

Formatada: Inglês (Reino Unido)

Formatada: Francês (França)

Formatada: Inglês (Reino Unido)

Formatada: Inglês (Reino Unido)

Formatada: Inglês (Reino Unido)

Formatada: Inglês (Reino Unido)

conference, "Advances in Research on Language Acquisition and Teaching, Thessaloniki, Greece, December 14-16, 2007.

Neff, J.A., Ballesteros, F., Dafouz, E., Martinez, F., Rica, J.P., Diez, M. and **Prieto, R.** 2004. "Formulating writer stance: A contrastive study of EFL learner corpora." In *Applied Corpus Linguistics: A multidimensional Perspective*, U. Connor and T.A. Upton (eds.). Amsterdam: Rodopi. 73-89.

Ochs, E. 1989. "Introduction." *Text* 9/1: 1-5.

Tankó, G. 2004. "The use of adverbial connectors in Hungarian university students' argumentative essays." In *How to use corpora in language teaching*, J. Sinclair (ed.). Amsterdam: John Benjamins, 157-181.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

A CORPUS STUDY OF INTELLECTUAL DEMANDS IN ENGLISH LANGUAGE CLASSROOMS: A CROSS-LINGUAL PERSPECTIVE

He, Ane[40]

Walker, Elizabeth[41]

Wang, Lixun[42]

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Abstract

This presentation introduces the methodological framework of a proposed corpus study on intellectual demands in English-as-second-language (ESL) classrooms. The study impetus comes from our concern that, given the need for very high ESL proficiency levels in Hong Kong (for higher education and business), English lessons appear intellectually inadequate (He 2006). Moreover, students' maturity and intellectual capacity observed in mother tongue (Cantonese Chinese) classrooms are generally ignored. Such classroom practices disregard the established principles of language as a thinking tool (e.g., Vygosky 1986) and of language development (e.g., Halliday 1994), regardless of whether the language is L1 or L2.

Intellectual demand is defined in this study as thinking opportunities provided for students in classrooms, including recall, recognition, retrieval and differentiation of information, abstract reasoning, planning, analysis, and judgment. Intellectual demand cannot be observed directly. It is mediated through the language of classroom discourse.

We intend to assess the intellectual demands of English lessons through a corpus approach, with data of Chinese classrooms as benchmark. A corpus with two sub-corpora (one in English, one in Chinese) will be compiled, each containing 40 enacted lessons. Patterns of intellectual demand are expected to emerge from analysis of the frequency lists and the collocational/associated patterns of these lessons. Specifically, we will explore lexical items (e.g. nouns in terms of concreteness/abstraction) to infer the subject focus of a lesson and its intellectual demand on students. We will also, following systemic functional grammar (Halliday 1994), explore process types realized in verbs to reveal the physical, mental, and cognitive experiences in which students are involved in the lessons, such as using language for doing (material processes), feeling and perceiving (mental processes), saying (verbal processes), or naming and describing (relational/existential processes). In addition, we will analyze personal pronouns and modal verbs to explore interpersonal meaning in classroom interaction, reflecting the intellectual demand of regulative aspects of lessons (Christie 2002, after Bernstein). Chinese corpus data will be collected and analyzed in the same way to obtain a parallel data-set, making possible a cross-lingual perspective within the constraints arising from differences between the languages.

Based on the argument above that language ***is*** *thought, levels of thinking will be judged by coding the language patterns identified in the corpus analysis with reference to a list of linguistic indicators of the intellectual hierarchy (low, moderate, and high) as exemplified in Bloom (1979), as well as other equivalents as they emerge during the corpus analysis.*

The proposed corpus study is expected to develop understanding of the interrelationship between language and thinking and also provide empirical evidence on how L2 teaching could take advantage of L1 competency. We

[40] He, Ane: an Assistant Professor in the Department of English at the Hong Kong Institute of Education. She teaches comparative language studies, discourse analysis for language teachers etc. at undergraduate and postgraduate level. Her research interests include corpus linguistics and classroom discourse analysis. She and her colleague created the Corpus of Classroom Language Teaching (CELT) in Hong Kong.

[41] Walker, Elizabeth: an Associate Professor in the English Department at the Hong Kong Institute of Education. Her research interests and publications concern the development of foreign language teaching and bilingual teaching of academic subjects. She and He Ane created the Corpus of Classroom Language Teaching (CELT) in Hong Kong.

[42] Wang, Lixun: an Assistant Professor in the English Department at the Hong Kong Institute of Education. His research interests include corpus linguistics, computer-assisted language learning, and online learning. He created an English-Chinese Parallel Corpus and developed the online E-C Concord program with his colleague (http://ec-concord.ied.edu.hk).

hope for advice from conference attendees regarding the cross-lingual analytical framework and corpora construction for two such distantly-related languages.

Keywords: corpus study, intellectual demand in second language classrooms, cross-lingual study

Introduction

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

This study introduces the methodological framework of a proposed corpus study on intellectual demands in English-as-second-language (ESL) classrooms. The study is inspired by the philosophical work on the nature of language, especially on language and thought; and the research on cross-lingual transfer. The study was also a result of our concern over the inadequacy of English lessons in Hong Kong, given the need for very high ESL proficiency levels required in the local context for higher education and business.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Language and thought

Discussion of language education requires an understanding of the nature of language. There have been two broad camps on this issue. Piaget and his followers maintain a theory of independence. That is, the abstract logical operation of cognition is independent of language. For them, "language is assumed to exist as a separable system… internally as a mental organ, or… externally as a means of expression and reception of language-neutral information" (Nelson 1996: 4). As such, language is recognized as an important communicative tool through which thoughts are expressed, but no link is made between the tool and thought. Vygotsky (1986), in contrast, holds a tool-mediated theory of mind, which asserts that language is not irrelevant to thought; and that higher forms of human mental activity are dependent on symbolic tools, with language being the most critical. For Vygotsky, language plays various roles in knowledge construction. As a medium, language expresses thought in terms of inner speech; and as a mediating tool, language is used to organize thought and knowledge systems (Nelson 1996: 21). In congruence with Vygotsky, Halliday (1994) states that language is a semiotic, meaning-making system. It construes human beings' experience by mediating their internal and external worlds, enacts interpersonal relationships and enables flow of information to coincide with the flow of events (Halliday 2002: 390). With a concern for the education context, Halliday (1993) argues "language is not a domain of human knowledge… Language is the essential condition of knowing, the process by which experience becomes knowledge (p. 94, original emphasis). It is in this sense that Bernstein's claim that "educational failure is primarily linguistic failure" is valid (Halliday 1973: 3, cited in Byrnes 2006: 4).

First and second language – cross-lingual relationships

When the mediation of language in thinking is ignored, there is neglect of the knowledge base and communicative skills acquired in the mother tongue, which learners bring to a L2/FL classroom. L2/FL learners are conventionally taken as inadequate language users, whose utterances are frequently related to 'error'. Not until recent years has the concept of the native speaker as a language model been seriously challenged (e.g. Kirkpatrick 2007; Braine 1999, 2005). Implicit in the expression 'privilege of the non-native speaker' (Kramsch 1997) is a similar concern with the interrelationship between learners' command of mother tongue and the development of a L2/FL, namely, cross-lingual transfer. Byrnes (2006) questions the common practice that L2 learners already literate in a L1 should follow the same sequence of language development as L1 learners.

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

The notion of 'the privilege of the non-native-speaker' (Kramsch 1997) derived from a sociocultural viewpoint. From a psycholinguistic perspective, Bialystok (2001) argues "second language learners need not relearn the fundamental principles of language structure... Second language acquisition is facilitated because a language template is available" (p. 127). Research studies (see Durgunoglu 2002) indicate positive cross-lingual transfer in several specific areas, such as functional awareness and meaning-making strategies. Cross-lingual transfer occurs not only between two alphabetical languages but also between an alphabetical and a non-alphabetical language such as Chinese (Geva & Wang 2001). Other researchers note that learners replicated L1 strategic behavior in the L2 across a large range of meta-cognitive skills (Hardin 2001) and exhibited better metalinguistic awareness (Goldin-Meadow 1999, cited in Bialystok 2001).

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Language teaching in Hong Kong

Language as a communicative tool is a familiar concept to the Hong Kong community. This can be seen from the specified content and recommended methodologies in a series of school curricula issued by the government in recent years (Walker, Tong & Mok 2000). However, language as thinking tool and as a representational system mediating a human being's internal and external worlds seems to be neglected at various levels of education reform. Although thinking skills are mentioned in the prescribed curricula, the interrelationship between language and thinking is never made clear. Language competence in these reforms has been defined in behavioral terms and relies primarily on stimuli in the immediate environment or on intuition. This tends to encourage immediate response at the expense of inward reflection by learners (Tarvin & Al-Arishi 1991: 23). By focusing exclusively on functional skills, and ignoring the mediating role of language in thought and knowledge construction, the abovementioned curricula and practice might contribute eventually to a "diminishment of language learners" (Tomlinson 1986: 34, cited in Pennycook 1994: 171).

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

The researchers' observation of classroom practice suggests that linguistic issues have been over-prioritized at the expense of intellectual demand. For example, perusal of widely used junior secondary English textbooks in Hong Kong reveals that English learning frequently occurs in the same contexts from early primary through to secondary 4, such as ordering food; arranging parties; going shopping. Moreover, when ubiquitous concepts such as 'present tense', or 'conditional clauses' are taught, the taught critical features of these concepts remain the same across levels. Alternatively, some teaching prioritizes communication without having students attend at all to the language system through which the communication occurs. Once students' maturity and intellectual capacity observed in mother tongue (Cantonese Chinese) classrooms are generally ignored, classroom practices as we observed, did not seem to have taken into consideration the established principles of language as a thinking tool (e.g., Vygosky 1986) and of language development (e.g., Halliday 1994), regardless of whether the language is L1 or L2. All this means that positive student motivation and development in linguistic understanding are unlikely, and that students risk experiencing both content-less and language-less English lessons (Walker, Tong & Mok 2000).

The studies reviewed above serve as a springboard for the current research. The established link between language and cognition makes it possible for the researchers to observe intellectual demands in classroom language use. The acknowledgment of the potential impact of L1 development on L2/FL helps the researchers to explore the issue of intellectual demand from a cross-lingual perspective. A lack of methodology in assessing intellectual demand motivates experimentation with a corpus approach. The current research therefore is a response to the calls for further investigation in English classrooms with an ultimate goal of contributing, theoretically and practically, to a better understanding of language teaching in schools.

The proposed study aims to explore the intellectual demands of English lessons from a cross-lingual perspective. It is not concerned with the establishment of absolutely accurate levels of intellectual demand per se, as demands in L2 teaching are very unlikely to equal those of the mother tongue, at least in junior secondary school. The ultimate goal of the research is to help teachers help their students to get as close as possible in English to what the students can already express in their mother tongue. Therefore, the Chinese lesson data will provide an essential benchmark for the purposes of the study, and the focus of the research will be on the intellectual demand placed on students in the teaching, not on individual students' cognitive capacities. Thus, no attempt will be made to match English and Chinese teaching by obtaining parallel Chinese and English lessons delivered to the same groups of students.

Intellectual demand is defined in this study as thinking opportunities provided for students in classrooms, including recall, recognition, retrieval and differentiation of information, abstract reasoning, planning, analysis, and judgment. Intellectual demand cannot be observed directly. It is mediated through the language of classroom discourse.

We intend to assess the intellectual demands of English lessons through a corpus approach, with data of Chinese classrooms as benchmark. A corpus with two sub-corpora (one in English, one in Chinese) will be compiled, each containing 40 enacted lessons. Frequency of words and collocation (co-occurring language) are fundamental to corpus analysis. Frequency information is taken as an indication of what aspect of a situation the discourse community considers to be key or most salient. In corpus analysis, collocation and association patterns are considered essential (Biber, Conrad & Reppen 1998: 6). These collocates and association patterns are believed to be able to reveal how a particular linguistic feature is systematically associated with particular words or with grammatical features in the immediate context (ibid).

Patterns of intellectual demand are expected to emerge from analysis of the frequency lists and the collocational/associated patterns of these lessons. With reference to the framework in He (2004, 2006) and He & Walker (2005), nouns and verbs are taken as indices of intellectual demand in English lessons. We will categorize, based on word frequency count, concrete nouns, abstract nouns; nouns within the immediate context of the classroom setting, and those beyond as an indication of the focus of a lesson and the effect of the intellectual demand it places on students. We will also categorize verbs in process types as in systemic functional grammar (Halliday 1994) to reveal the physical, mental, and cognitive experiences in which students are involved in the lessons, such as using language in doing (as in material processes), feeling and perceiving (as in mental processes), saying (as in verbal processes), or naming and describing (as in relational/existential processes). In addition to the analysis of nouns and verbs in the frequency lists, we will examine four- and five-word clusters for the purpose of capturing the patterns which, although unable to reach the top of the word lists, occur, nevertheless, in larger chunks of words in the lessons. Based on a belief that the teachers' managerial talk (managing the content and student behavior) contributes to intellectual demand, especially when it is in an L2/FL, we will analyze personal pronouns and modal verbs to explore interpersonal meaning in classroom interaction, reflecting the intellectual demand of regulative aspects of lessons (Christie 2002, after Bernstein). Chinese corpus data will be collected and analyzed in the same way to obtain a parallel data-set, making possible a cross-lingual perspective within the constraints arising from differences between the languages.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

A key methodological issue is how intellectual demand will be assessed. Assessment will be through inference from the patterns of language use identified in the analysis, based on the argument above that language **is** thought and judgment of levels of thinking (conceptual understanding) is, and can only be derived from examination of language. We will judge the levels of thinking in this study by coding the language use with reference to a list of linguistic indicators of the intellectual hierarchy (low, moderate, and high) in accordance with long established practice in determining levels of intellectual demand as exemplified in Bloom (1979). The hierarchy presented below is developed from the conceptual work of Mohan (1986) in language classrooms, Biggs and others (e.g., Biggs 1996) in other disciplines, and Panizzon and Bond (2006) in science. Typical wording indicating the thinking related to each of the hierarchical categories may include those specified in the list given, as well as other equivalents as they emerge during the corpus analysis. The list of the indicators is as follows.

Lower level intellectual demand

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Language use in this category involves students in unidimensional thinking. Students are required to recall and recognize previously learned concepts and principles, as follows.

- Recall or recognize a fact
- Reproduce previously seen materials
- Follow simple procedures
- Draw
- Describe
- Define
- Retrieve/locate/identify information from texts, tables, graphs, figures, etc.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

Moderate level intellectual demand

Language use in this category involves multidimensional thinking. Students are required to choose among alternatives in executing tasks that require a response beyond the simple retrieval of information by considering more than one single step and more than one single variable, as follows.

- Represent
- Select
- Compare
- Interpret
- Identify main theme
- Identify organizational patterns

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

High level intellectual demands

Language use in this category involves multidimensional and multirelational thinking. Students are required to engage in abstract reasoning, planning, analysis, and judgment, as follows.

Formatada: Espaçamento entre linhas: Múltiplo 1,1 lin

- Generalize a pattern
- Analyze or produce a deductive argument
- Describe, compare, and evaluate a situation
- Categorize, schematize
- Text hypotheses

A profile of the intellectual demands of the English and Chinese lessons will be established on the basis of this hierarchical list. Using the profiles thus established, our interpretations will be made regarding how and in which ways English lessons relate to Chinese lessons in terms of intellectual demands in the context of external prescriptions such as the linguistic requirements of the language curricula. The results are expected to illustrate to what extent the intellectual demands expected and required of students in their L1 development do or do not mirror those expected and required in their L2 development. The findings will lead to our recommendations as to how L2 teaching can exploit the cognitive maturity of the students and the intellectual demands placed on them in their L1 learning.

The proposed corpus study is expected to develop understanding of the interrelationship between language and thinking and also provide empirical evidence on how L2 teaching could take advantage of L1 competency. We hope for advice from conference attendees regarding the cross-lingual analytical framework and corpora construction for two such distantly-related languages.

Formatada: Espaço Antes: 6 pto, Espaçamento entre linhas: Múltiplo 1,1 lin

References

Bialystok, E. 2001. *Bilingualism in Development: language, literacy and cognition*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R.1998. *Corpus Linguistics: investigating language structure and use.* Cambridge: Cambridge University Press.

Biggs, J. 1996. (ed). *Testing: to educate or to select? Education in Hong Kong at the crossroads.* Hong Kong: Hong Kong Educational Publishing Co.

Bloom, B.J. 1979. *Taxonomy of Educational Objectives: handbook 1, cognitive domain*. New York: Longman.

Braine, G. 1999. (ed) *Non-Native Educators in English Language Teaching*. Mahwah, N.J.: Lawrence. Erlbaum Associates.

Braine, G. 2005. (ed) *Teaching English to The World: History, Curriculum, and Practice.* Mahwah, N.J.: Lawrence. Erlbaum Associates.

Byrnes, H. 2006. "What kind of resource is language and why does it matter for advanced language learning? " In Byrnes (ed.) *Advanced language learning*: *the contribution of Halliday and Vygotsky*. London: Continuum,

Christie, F. 2000. "The pedagogic device in the teaching of English." In *Pedagogy and the Shaping of Consciousness*, F. Christie (ed.). London: Continuum, 156-184.

Curriculum Development Council 2002. *English Language Education. Key Learning Area Curriculum Guide.* Hong Kong: Curriculum Development Council.

Durgunoglu, A. Y. 2002. "Cross-linguistic transfer in literacy development and implications for language learners." *Annals of Dyslexia*, 52: 189-204.

Geva, E. & Wang, M. 2001. "The development of basic reading skills in children: a cross-language perspective." *Annual Review of Applied Linguistics*, 21: 182-204.

Halliday, M.A.K. 1993. "Towards a language-based theory of learning." *Linguistics and Education*, 5: 93-116.

Halliday, M.A.K. 1994. *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M.A.K. 2002. *On Grammar*. London: Continuum, 369-417.

He, A.E. 2004. "Use of verbs in teacher Talk: a comparison study between LETs and NETs. " *Hong Kong Journal of Applied Linguistics*, 9,2: 38-54.

He, A.E. 2006. "Subject matter in Hong Kong primary English Classrooms: a critical analysis of teacher talk. " *Critical Inquiry in Language Studies* 3/2&3:169-188.

He, A.E. & Walker, E. 2005. "Linguistic and Discoursal Structures of Monologues in Transmission-oriented Classrooms: an exploratory study of a science lesson and a geography lesson in Hong Kong English-Medium Schools. " Paper presented at *International Society for Language Studies Conference*, Montreal, Canada.

Kirkpatrick, A. 2007. *World Englishes: Implications for International Communication and English Language Teaching.* Cambridge: Cambridge University Press.

Kramsch, C. 2003. "The privilege of the nonnative speaker." *Issue in Language Program Direction, A Series of Annual Volumes*: 251-262.

Macdonald, R. 1994. *"Why can't I understand this?" A study of some relationships between student cognitive level and the level of cognition required for understanding junior science.* Masters thesis. Townsville: James Cook university.

Nelson, K. 1996. *Language in Cognitive Development: emergence of the mediated mind*. Cambridge: Cambridge University Press.

Panizzon, D. and **Bond, T.** 2006. "Exploring Conceptual Understandings of Diffusion and Osmosis by Senior High School and Undergraduate University Science Students." In Liu X. F. (ed) *Applications of Rasch Measurement in Science Education*. Buffalo: State University of New York, 1-28.

Pennycook, A. 1994. *The Cultural Politics of English as an International Language*. London: Longman.

Tarvin, W. and **Al-Arishi, A. Y.** 1991. "Rethinking communicative language teaching: reflection and the EFL classroom." *TESOL Quarterly*, Vol. 25, No. 1: 9-27.

Walker, E., Tong, A. S. Y. and **Mok Cheung, A. H. M.** 2000. "Changes in secondary school English teaching methodologies and content (1975 to 1999)." In Y. C. Cheng, K. W. Chow, & K. T. Tsui *School Curriculum Change and Development in Hong Kong.* Hong Kong: HKIEd

Vygotsky, L.S. 1986. *Thought and Language*. Cambridge, MA: MIT Press.

A CORPUS-BASED LINGUISTIC PROFILING OF HIGH AND LOW STUDENT ENGAGEMENT CLASSROOMS IN SINGAPORE SCHOOLS

Huaqing Hong[43]

Abstract

Taking a corpus-linguistic approach to conversational analysis, this study investigates whether there is a strong correlation between teachers' talk and students' engagement levels in the classrooms of Singaporean primary and secondary schools. To identify the linguistic features of high and low engagement classrooms, we adopt two sources of authentic materials collected in the Singapore Corpus of Research in Education (SCoRE) (Hong 2005): the coding data from 455 observed authentic classroom practices as well as the 2.3 million words of transcripts of recorded classroom interactions. Linguistic and statistical analyses of the linguistic patterns used in the sample lessons of high and low student engagement classrooms revealed that the use of a list of linguistics features in the teachers' talk varies systematically across classes of different subjects, levels and streams. Thus, linguistic profiles of the teachers' talk in high and low student engagement classrooms are identified and statistically justified. These profiles can be used as an overall picture of current practice in the classes where students' engagements vary across subjects and levels, and can also be useful in future school intervention and teachers' professional development.

Keywords: Corpus, Classroom Discourse, Engagement, Linguistic Profiling

Introduction

Student engagement is undoubtedly a major factor of students' performance in schools and thus for a long time it has been a very hot topic in traditional education research and it is still the focus of many recent studies (Hake 1998, Brewster & Fager 2000, Willms 2003; Norris, Pignal & Lipps 2003; Rosario 2006; Harmin & Toth 2006; Laitsch 2007; Yeh 2007; to name just a few). Student engagement in learning has also been identified to be associated with reduced dropout rates and increased levels of student success (Blank 1997; Dev 1997; Kushman *et al* 2000; Woods 1995; to list just a few). Getting students engaged during class have always been a challenge to teachers. Many factors, such as students' age, self-efficacy, and gender, contribute to students' level of engagement in learning. For example, studies have shown that student engagement declines as students get older. Students' self-efficacy towards learning can also affect their level of engagement in learning (Anderman & Midgley 1997 & 1998). Students, who attribute poor performance to a lack of attainable skills, rather than to some innate personal deficiency, are more likely to re-engage themselves in a task and try again. Marks (2000) found that girls are more engaged in instructional activities than boys across all levels. Peer influence is another factor that contributes to student engagement in learning, especially for high school-age students. As students grow older, their motivation to engage in learning may be influenced by their peers just as much as, if not more than it is by teachers, parents, and other adults. While peer influences can be either positive or negative, it is not uncommon for older students to discourage one another from actively participating in school (MacIver & Reuman 1994).

Although teachers have no control over students' background factors as such gender, social economic status etc., research has shown that there are ways to make their classroom instructions and assigned work more engaging and more effective for students at all levels. For example, Marks (2000) found that authentic instructional work, classroom support, and parental involvement are all correlated with student engagement.

[43] Huaqing HONG is a Lead Research Associate with the Centre for Research in Pedagogy and Practice (CRPP), National Institute of Education (NIE), Nanyang Technological University (NTU), Singapore. He has been working as university lecturer, teacher trainer and education researcher both in China and Singapore since 1992. He has published in refereed journals like World Englishes, Language and Education, Language Policy, and Journal of Chinese Sociolinguistics, and presented at a dozen of international conferences in Czech, UK, France, New Zealand, China, USA, Australia, Hong Kong, Singapore, etc. His research interests include computational linguistics, corpus linguistics, discourse analysis, and cross-cultural communication and translation studies.
Contact detail:
Address: Blk5, B3-CRPP, 1 Nanyang Walk, Singapore 637616,
Telephone: +65 6216 6269,
Email: huaqing.hong@nie.edu.sg

At present, most studies on student engagement has focused on students' characteristics (such as age, attitudes), situational characteristics (such as subjects, social economic status) and teacher instruction types (such as classical, authentic). Taking a corpus linguistics approach to classroom discourse analysis, this study differs from prior studies by looking at whether there is a correlation between teachers' talk and students' engagement levels in the classrooms. It adopts some advanced statistical analyses like factor analysis and cluster analysis to classroom discourse, and tries to profile the linguistic features correlated with students' engagement levels in a large corpus of real-life classroom interactions.

Data

The Corpus

The data for the present study is a sub-corpus of the SCoRE project (see more at http://score.crpp.nie.edu.sg). The table below is the breakdown of the data used for the present study. This sub-corpus consists of a total of 121 teachers' 455 lessons recorded in 40 primary and secondary schools in Singapore. The 253.3 hours recordings of these lessons were manually transcribed to machine readable texts for the corpus compilation. These class sessions were sampled on a common design and a group of experienced researchers were assigned to observe the lessons in the classrooms to code the actual practices with a list of designated features in the Singapore Coding Sheet (Luke *et al* 2004). Engagement levels were then generated from the coding data.

To compare between subjects, levels and streams, 216 Primary Five (P5) lessons of 3 streams (EM1, EM2, and EM3) were sampled, while 239 Secondary Three (S3) lessons of 3 steams (EXP, NA, and NT) were collected. The 4 main school subjects were selected for this study and the size of this corpus for this study is over 2.3 million words (excluding event markers, comments, punctuation marks and other non-utterance tags in the transcripts). It is believed that the sampling method and the corpus size are good enough for the present study.

Subject	**Grade**	**Stream**	**Teachers**	**Lessons**	**Dur(hrs)**	**Tokens**
English	P5	EM1	5	16	14.7	145277
		EM2	7	18	12.6	111106
		EM3	1	3	21.6	178724
	S3	EXP	9	30	25.5	243064
		NA	3	11	8.1	79695
		NT	4	13	9.0	85282
Math	P5	EM1	4	11	6.3	57945
		EM2	7	29	13.6	121659
		EM3	2	9	8.8	76934
	S3	EXP	8	26	14.2	140976
		NA	4	16	9.4	93853
		NT	7	25	11.2	104427
Science	P5	EM1	6	25	9.4	83741
		EM2	4	30	11.3	101994
		EM3	3	29	1.5	12114
	S3	EXP	11	46	15.6	143124
		NA	4	14	4.8	38850
		NT	4	18	7.2	75701
Social Studies	P5	EM1	3	8	5.3	32196
		EM2	9	26	10.4	100859
		EM3	4	12	5.9	50141
	S3	EXP	8	25	17.8	148900
		NA	4	15	9.2	86095
		NT	0	0	0.0	0
TOTAL	**2 Levels**	**6 Types**	**121**	**455**	**253.3**	**2312657**

Table 1: Breakdown of the corpus data for the analysis

Student Engagement Levels

Observers scored the levels of student engagement during class on a 4-point scale. Thus each transcript has an associated score for student engagement level. Some moderation has been made when all the scores were collected, and as the distribution of score was skewed, the engagement was then categorized with statistical analysis and exporters' judgment (please refer to for more). As shown in the table below, the categorization resulted in 292 high student engagement (score=4) and 163 low student engagement (score≤3) classes/transcripts. With the distribution across social variables, such as School Subjects, Grades, Teacher Gender, Teacher Age, Teacher Qualification and Teaching Experience, the linguistic profiles of high engagement classes were then compared to those low engagement classes in next section.

Engagement	School Subject				Total
	Science	Mathematics	English	Social Studies	
Low	32	50	49	32	163
High	59	66	113	54	292
Total	91	116	162	86	455

Engagement	Grade Level		Total
	Primary 5	Secondary 3	
Low	62	101	163
High	154	138	292
Total	216	239	455

Engagement	Teacher Gender		Total
	Male	Female	
Low	47	116	163
High	80	212	292
Total	127	328	455

Engagement	Teacher Qualification				Total
	Certificate	Diploma	Bachelor	Post-graduate	
Low	15	13	128	7	163
High	23	32	228	9	292
Total	38	45	356	16	455

Engagement	Teacher Age (Year)							Total
	<25	25-30	30-35	35-40	40-45	45-50	>=50	
Low	3	66	43	17	9	3	22	163
High	7	77	86	58	27	11	26	292
Total	10	143	129	75	36	14	48	455

Engagement	Teaching Experience (Year)							Total
	<3	3-5	5-10	10-15	15-20	20-25	>=25	
Low	50	7	60	14	4	8	20	163
High	67	21	104	44	12	17	27	292
Total	117	28	164	58	16	25	47	455

Table 2: Statistics of high and low engagement lessons by school subject, grade level, teacher gender, teacher age, teacher qualification and teaching experience

Linguistic Features Selected for the Analysis

Linguistic features were automatically tagged by making use of the online version of Wmatrix corpus analysis and comparison tool (Rayson 2003 & 2008). Although the set of linguistic features identified included both grammatical and semantic features, only the grammatical features were used for the analysis in this study as the accuracy rate of the semantic tagging is not as high as expected. A total of 132 grammatical features were identified, but only 43 features were selected in the analysis. Many features were not included because they were either extremely rare in the conversations, or have little or no variation across transcripts. Still some features were not included because they had low communality estimates, suggesting that they shared very little variance with the overall factorial structure of this analysis. The set of 43 features is listed in the table below.

cardinal number, neutral for number (two, three..)
base form of lexical verb (e.g. give, work)
past tense of lexical verb (e.g. gave, worked)
plural common noun (e.g. books, girls)
was
2nd person personal pronoun (you)
1st person sing. subjective personal pronoun (I)
infinitive (e.g. to give... It will work...)
article (e.g. the, no)
do, base form (finite)
singular cardinal number (one)
have, base form (finite)
3rd person sing. subjective personal pronoun (he, she)
did
am
is
not, n't
preceding noun of title (e.g. Mr., Prof.)
singular common noun (e.g. book, girl)
modal auxiliary (can, will, would, etc.)
of (as preposition)
infinitive marker (to)
3rd person sing. objective personal pronoun (him, her)
ing participle catenative (going in be going to)
that (as conjunction)
were
interjection (e.g. oh, yes, um)
1st person sing. objective personal pronoun (me)
wh-determiner (which, what)
be, infinitive (To be or not... It will be ..)
3rd person plural subjective personal pronoun (they)
singular letter of the alphabet (e.g. A,b)
comparative after-determiner (e.g. more, less, fewer)
common noun, neutral for number (e.g. sheep, cod, headquarters)
3rd person sing. neuter personal pronoun (it)
singular proper noun (e.g. London, Jane, Frederick)
singular determiner (e.g. this, that, another)
had (past tense)
possessive pronoun, pre-nominal (e.g. my, your, our)
1st person plural subjective personal pronoun (we)
subordinating conjunction (e.g. if, because, unless, so, for)
general adverb
past participle of lexical verb (e.g. given, worked)

Table 3: List of grammatical features used in analysis

A computer tool was then used to get the frequency counts for these grammatical features selected, and the data was then further processed for some advanced statistical analyses with SPSS version 15 in next section.

Factor Analysis

Counts of each linguistic feature were computed for each of the 455 transcripts of the lessons by means of a computer program. The frequency counts were first normalized with respect to the total number of tokens in the transcripts. The standardized counts were then subjected to Factor Analysis with Principal Component extraction and Varimax with Kaiser Normalization (rotation converged in 10 iterations) using SPSS 15. Finally a five-factor solution was obtained as shown in the figure and table below.

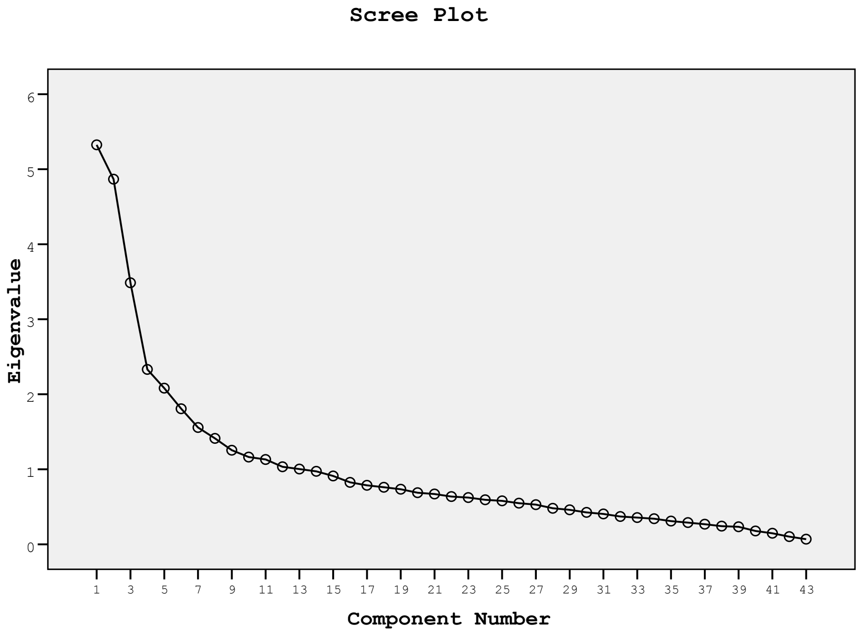


Figure 1: Scree Plot of the factor analysis result

Rotated Component Matrix

Grammatical Features	Component				
	1	2	3	4	5
past tense of lexical verb (e.g. gave, worked)	.835				
was	.765				
3rd person sing. subjective personal pronoun (he, she)	.676				
did	.638		.345		
had (past tense)	.565				
were	.506			.368	
preceding noun of title (e.g. Mr., Prof.)	.495				
singular proper noun (e.g. London, Jane, Frederick)	.494				-.410
3rd person sing. objective personal pronoun (him, her)	.473				
infinitive (e.g. to give... It will work...)		.694	.329		
infinitive marker (to)		.690			
modal auxiliary (can, will, would, etc.)		.587			
ing participle catenative (going in be going to)		.585			
am		.560			
1st person sing. subjective personal pronoun (I)		.548		-.391	
subordinating conjunction (e.g. if, because, unless, so, for)		.426			
be, infinitive (To be or not... It will be ..)		.420			
that (as conjunction)		.378	-.322		.361
do, base form (finite)			.761		
base form of lexical verb (e.g. give, work)			.673		-.363
not, n't			.638		
2nd person personal pronoun (you)		.455	.605		
general adverb			-.533		
of (as preposition)			-.484	.313	
interjection (e.g. oh, yes, um)			.470		
wh-determiner (which, what)			.429		
1st person plural subjective personal pronoun (we)			-.392		
is				-.647	.414
plural common noun (e.g. books, girls)			-.389	.637	
cardinal number, neutral for number (two, three..)		-.369		-.633	-.371
singular cardinal number (one)				-.522	
singular letter of the alphabet (e.g. A,b)				-.517	
3rd person plural subjective personal pronoun (they)				.490	
common noun, neutral for number (e.g. sheep, cod, headquarters)				.463	
singular determiner (e.g. this, that, another)				-.438	
1st person sing. objective personal pronoun (me)				-.337	
have, base form (finite)				.327	
past participle of lexical verb (e.g. given, worked)					
3rd person sing. neuter personal pronoun (it)					.577
article (e.g. the, no)		-.307			.563
singular common noun (e.g. book, girl)		-.310			.423
possessive pronoun, pre-nominal (e.g. my, your, our)					-.396
comparative after-determiner (e.g. more, less, fewer)					-.326

Table 4: List of five factors identified and the respective positive and negative loadings of the grammatical features

The number of factors extracted was based on inspection of the scree plot, eigen-values, as well as interpretability of the factors. These five factors were interpreted as: (1) *Narrative*; (2) *Simple Instructional*; (3) *First-person versus Object-oriented*; (4) *Explanatory*; and (5) *General versus Numerical*. A set of factor scores was then obtained for each of the 455 transcripts observed. Factor scores, computed based on the regression method, were then used to do a list of Logistic Regression analyses to identify the association patterns and statistical significance levels of these five factors against school subjects, grade levels, streams as well as teachers' social variables (e.g. age, gender, experience, qualification, etc.).

Results and Discussion

Comparing Across Subjects

The figure below shows the comparison of the profile of factor scores between high and low student engagement classes for different subjects. For English, it appeared that a *Narrative* and *Explanatory* style distinguished between high and low student engagement. For Math classes, the profiles did not differ much except that too much focus on what the teacher is going to do (*First-Person Oriented*) and simple instructions (*Simple Instructional*) was correlated with low student engagement. The same can be observed for Science classes. However, unlike Math, adopting a more *Explanatory* style was associated with high student engagement.

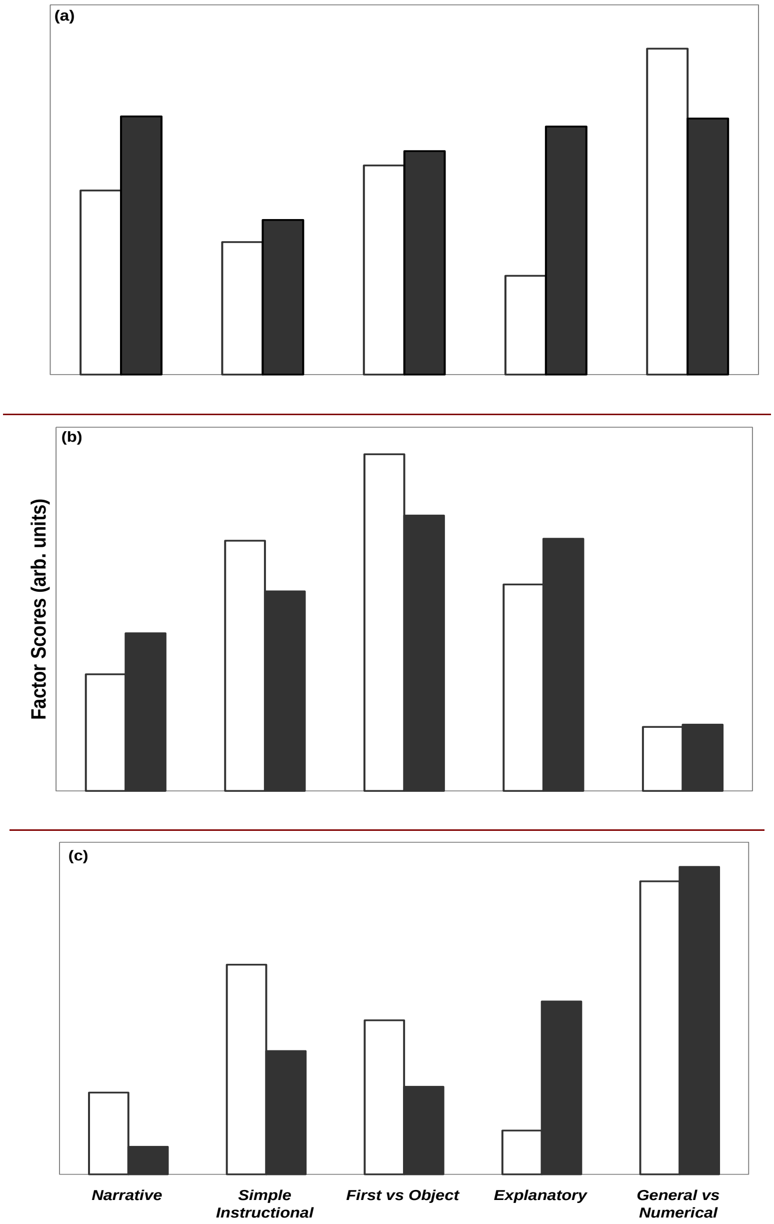


Figure 2: Linguistic profiles of high (black bars) and low (white bars) student engagement for (a) English, (b) Math, and (c) Science classes.

It appears that in Math classes, the linguistic profile of teachers' talk is less correlated with students' engagement. Similarly, with the exception of one factor, the linguistic profiles do not differ dramatically across high and low student engagement for English and Science. Notable trends are that (1) for Math and Science classes, *Simple Instructional* and *First-Person Oriented* styles are associated with low student engagement, but this trend is not observed in English classes, and (2) in general, a less *Explanatory* style is associated with low levels of student engagement, although the difference is less pronounced for Math classes. We speculate that the findings reflect the low need for simple instructions during English classes, and the low provision of elaborate explanations during Math classes. The latter is not intuitive given that there is generally a need to explain mathematical concepts during class. This could reflect a problem with the current mathematical instructions where students are spoon-fed formulas rather than encouraged to think. Further research in associating authentic instruction with linguistic profiling may reveal more.

Comparing Across Levels

The figure below shows the comparison of the profiles of factor scores between high and low student engagement classes for different levels. It can be observed that less *Explanatory* style was associated with low student engagement. However, for other factors, P5 and S3 showed very different trends. While for the first two factors – *Narrative* and *Simple Instructional* – there were no differences between high and low student engagement for P5, the S3 profiles showed that more *Narrative* style was associated with high student engagement while *Simple Instructional* style was associated with low student engagement. While focusing on what the teacher is going to do was associated (*First-Person Oriented*), and less *General*-oriented was associated with high student engagement in P5 classes, the reverse was true for S3 classes.

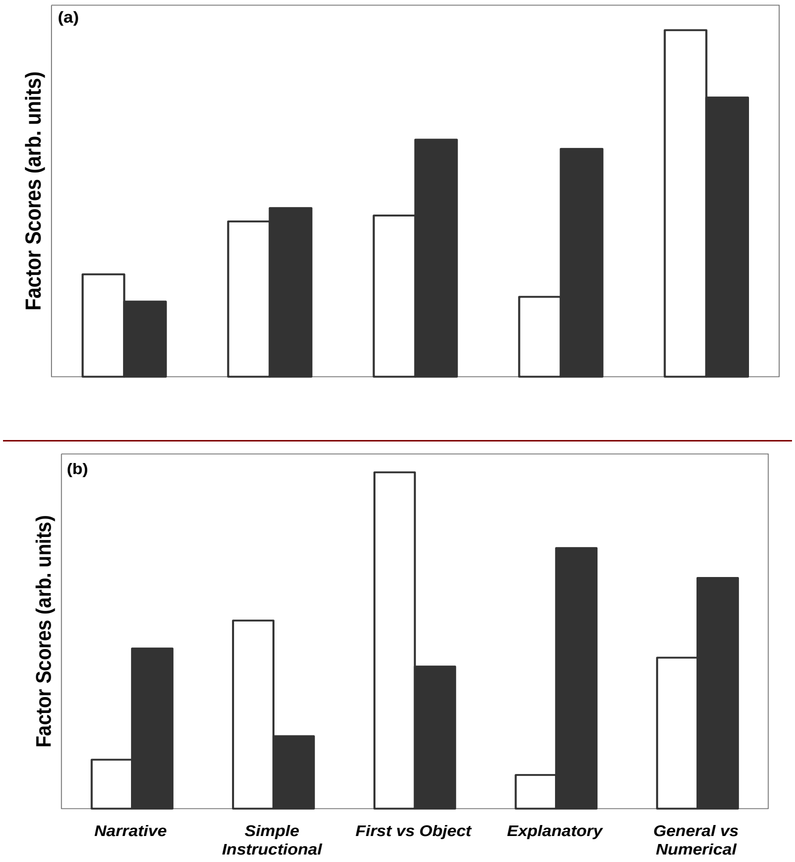


Figure 3: Linguistic profiles of high (black bars) and low (white bars)

student engagement for (a) P5, and (b) S3 classes.

The comparison across levels reveals more interesting findings. The linguistic profiles for high and low student engagement clearly differs for the two levels and in different ways. This is especially the case for S3 where the linguistic profiles for high and low student engagement are almost opposites. The findings suggest that for S3, non-narrative, non-explanatory, first-person oriented, simple instructions are associated with low engagement. This finding is intuitive as one can predict that a teacher who gives non-story-like, simple description of what she's going to do and with no explanations given is likely to bore his/her students. On the other hand, story-like elaborations of events is much more likely to engage students in class. It is notable that this applies only to high-school level students. For young students, surprisingly, such speech style is not associated with high engagement. It is the explanation of what the teacher is doing that engages young students more.

Final Remarks

Taking a corpus linguistics approach to profile the linguistic patterns of teachers' talks in terms of students' engagement levels, this study has shown that there is a strong correlation between teachers' language use in classrooms and students' engagement levels. Differing from the prior research focused only on students' variables, this study has identified a systematic variation of the correlation patterns across school subjects, grade levels, streams as well as teachers' social variables, such as age, gender, experience, qualification, etc. It is believed that the investigation of this kind is one of the first few, if not the first one, which has identified and profiled the correlation patterns between linguistic features of teachers' talks and students' engagement levels, drawn on a large corpus of authentic classroom practices and advanced statistical analyses.

A limitation of this study is that there are just too many linguistic features and social variables to select for any form of quantitative comparison. As these are correlational data, causality cannot be established. We are not certain whether the speech styles resulted in the high/low levels of student engagement are a combined effect with other factors (e.g. social variables of schools, teachers and students) or vice versa. In addition, as student engagement is rated on a relatively restricted 4-point scale, the variance across classes is admittedly limited. Operational definition of student engagement is also simplistic, as it does not capture the full range of the types of student engagement (e.g. student engagement in different types of class phases, learning tasks, and activities). However, the criticism on rating of students' engagement level in classrooms is beyond this study's scope and research focus.

Despite its limitations, profiling the teachers' talk in real classroom activities of high and low student engagement classes can reveal interesting patterns. This influence of teachers' speech on student engagement has often been anecdotally acknowledged yet often neglected in quantitative studies, probably due to the difficulty in obtaining quality data. This study hopes to illustrate the usefulness of this corpus-based approach to studying student engagement – the profiles obtained can be used as an overall picture of current practice in the classes where students' engagements vary across subjects and levels, and they can also be useful in future school intervention and teachers' professional development.

Acknowledgements

This paper makes use of the data from the research projects "Compiling a Multimodal Corpus Database of Education Discourse in Singapore Schools" (CRP7/05AL) and "Computational Language Analysis of Singapore Classroom Interactions: The CRPP computational program (CRP20/04JD)", funded by the Ministry of Education, Singapore, through the Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, Singapore. Thanks goes to Liobo Guo for his comments on the earlier draft of this paper, and I'm also grateful Chan Wai Yen and Papia Sultana, for their assistance with data analyses. The views expressed in this paper are all the author's and do not necessarily represent the views of the Centre or the Institute or the Ministry. For more about the center and the projects, please visit: http://www.crpp.nie.edu.sg.

References

Anderman, E. M., & Midgley, C. 1997. "Changes in personal achievement goals and the perceived goal structures across the transition to middle schools." Contemporary Educational Psychology, 22, 269-298.

Anderman, E. M., and **Midgley, C.** 1998. *Motivation and Middle School Students*. Champaign, IL: ERIC Clearinghouse on Elementary and Early Childhood Education.

Formatada: Inglês (Reino Unido)

Blank, W. 1997. "Authentic instruction." In W.E. Blank & S. Harwell (Eds.), *Promising practices for connecting high school to the real world* (pp. 15-21). Tampa, FL: University of South Florida.

Brewster, C., & Fager, J. 2000. *Increasing student engagement and motivation: from time-on-task to homework*. Portland, OR: Northwest Regional Educational Laboratory. http://www.nwrel.org/request/oct00/ [Access date: 15/05/2008].

Dev, P. C. 1997. "Intrinsic motivation and academic achievement: What does their relationship imply for the classroom teacher?" *Remedial and Special Education*, Vol.18 (1), 12-19.

Hake, R. 1998. "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses." *American Journal of Physics*, Vol. 66(1), 64–74.

Harmin, M., and **Toth, M.** 2006. *Inspiring Active Learning: A Complete Handbook for Today's Teachers*. Alexandria, VA: ASCD.

Hong, H. 2005. "SCoRE: A multimodal corpus database of education discourse in Singapore schools", In *Proceedings of the Corpus Linguistics Conference Series, Vol. 1, No. 1 (ISSN 1747-9398)*. Birmingham, UK, July 14-17, 2005.

Kushman, J. W., Sieber, C., & **Heariold-Kinney, P.** 2000. "This isn't the place for me: School dropout." In D. Capuzzi & D.R. Gross (Eds.), *Youth at risk: A prevention resource for counselors, teachers, and parents* (3rd ed., pp. 471-507). Alexandria, VA: American Counseling Association.

Laitsch, D. 2007. "Interactive engagement vs. traditional instruction in physics." *ResearchBrief*, Vol. 5(4), 23 April, 2007.

Luke, A., Cazden, C., Lin, A., & **Freebody, P**. 2004. A Coding Scheme for the Analysis of Classroom Discourse in Singapore Schools. Research Report, Centre for Research in Pedagogy and Practice, National Institute of Education, Singapore.

MacIver, D. J., and **Reuman, D.A.** 1994. "Giving their best: Grading and recognition practices that motivate students to work hard." *American Educator*, Vol. 17, 24-31.

Marks, H. M. 2000. "Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years." *American Educational Research Journal*, Vol. 37, 153-184.

Norris, C., Pignal, J., and **Lipps, G.** 2003. "Measuring school engagement." *Education Quarterly Review*, Vol. 9(2), 25-34.

Rayson, P. 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Unpublished PhD thesis, Lancaster University.

Rayson, P. 2008. Wmatrix*: a web-based corpus processing environment*. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/.

Rosario, J. R. 2006. *On engaging Latino students in their education: A resource guide to research and programs.* http://www.elpuenteproject.com/defiles/On%20Engaging%20Latino%20Students%20in%20their%20Education-%20%20A%20Resource%20Guide%20to%20Research%20and%20Programs.pdf [Access date: 15/05/2008]

Willms, J. D. 2003. *Student engagement at school: A sense of belonging and participation*. Paris: Organization for Economic Cooperation and Development.

Woods, E. G. 1995. "Reducing the dropout rate." In *School Improvement Research Series (SIRS): Research you can use*. Portland, OR: Northwest Regional Educational Laboratory. http://www.nwrel.org/scpd/sirs/9/c017.html [Access date: 15/05/2008].

Yeh, S. 2007. "Improving engagement and achievement through rapid assessment." *The Leader* (e-newsletter), Spring 2007. http://www.education.umn.edu/edpa/licensure/leader/2007Spring/Yeh.html [Access date: 15/05/2008]

Formatada: Inglês (Reino Unido)

"WELL I DON'T KNOW WHAT ELSE CAN I SAY" DISCOURSE MARKERS IN ENGLISH LEARNER SPEECH

Joanna Jendryczka-Wierszycka[44]

Abstract

The paper presents the findings on discourse markers (DMs) (Schiffrin, 1987) in advanced spoken English interlanguage of Polish native speakers. The hypothesis is that Polish EFL speakers underuse some functions of discourse markers and that DM clustering occurs much less frequently in Polish EFL speech than in native English speech and learner range of DMs is significantly narrower due to insufficient vocabulary skills.

The present analysis is divided into two parts. In the first part, the most frequent 3-word unit (I don't know) is extracted from the corpus of native and non-native speech. It is on the basis of this sequence that the present analysis is conducted. As a second step, the meaning and functions of I don't know are examined after Tsui (1991) and Diani (2004). In the latter part of the study, collocations of I don't know with other DMs are examined on the basis of their pragmatic meaning.

Corpora used for the comparison are the Polish LINDSEI subcorpus and LOCNEC, an English reference corpus.

*The analysis shows that both Polish EFL speakers and native English speakers employ the phrase I don't know to perform similar functions (to mark uncertainty, to avoid commitment, to express ambivalent feelings or to minimize potentially face-threatening acts and to build rapport with the interlocutor). Some functions (like avoiding explicit disagreement) are absent from those described by Diani (2004) due to the specific character of the corpus (Pulcini and Furiasi 2004). The collocates of I don't know are similar in both corpora (*well, oh, maybe, yeah*). However, some are significantly underused, others overused. The paper also provides some implications for teaching of the results reported.*

Keywords: corpus linguistics, spoken learner corpora, EFL, CIA, discourse markers

Introduction

To know

Not many would find it problematic to answer a question: "What does 'to know' mean?". In case of doubt, though, a dictionary helps. Since the paper reports on a study of learner language, let me quote from *Longman Exams Dictionary* (henceforth LED). First of all, the dictionary mentions eight basic senses of the verb in question (marked by "signposts" under a given entry). These are: "have information", "be sure", "be familiar with sb/sth", "realize", "skill/experience", "know sb's qualities", "recognize", "experience" (LED, 843-844).

Interestingly enough, at the same time the dictionary presents a range of 22 "spoken phrases" (LED, 844) with the verb "to know". They are all presented in a study note, prepared especially for a language learner. Some of them are e.g. *not that I know of* or *for all I know*. Among them, there is also a note on the expression *I don't know* which is the focus of this paper. This phrase is defined in a twofold way: "a) used to say that you do not have the answer to a question": '*When did they arrive?' 'I don't know.*' b) used when you are not sure about something: *'How old do you think he is?' 'Oh, I don't know, sixty, seventy?'*

[44] Joanna Jendryczka-Wierszycka is a PhD student at the Department of Computer Assisted English Linguistics, School of English, Adam Mickiewicz University in Poznan, Poland. As a part of her M.A. project she compiled and transcribed the Polish subcorpus of the LINDSEI project. Ever since then, she has been actively engaged in the work on the project, the effect being the inclusion of the Polish subcorpus in the LINDSEI CD-rom to be published this year. So far, she has been working on expressing vagueness in learner speech. An article reporting on this phenomenon is currently in press in a PALC 2007 post-conference proceedings volume. Apart from the PALC conference, she also took part in PLM 2007 conference and her papers were also accepted for this year's TaLC and EUROCALL conferences. She is also a beneficiary of the János Kohn Scholarship.

Study description and aims of the study

As we can see, it is only the literal meaning and a meaning expressing uncertainty that are presented in the study note. However, there are many more meanings that native speakers express while uttering the phrase *I don't know*. According to Tsui (1991), these are: avoidance of making an assessment, preface to a disagreement, avoidance of an explicit disagreement, avoidance of commitment, minimization of impolite beliefs and marker of uncertainty. These meanings are discussed below in reference to non-native and native English data. Discourse Markers (DMs) occurring in collocation with the phrase *I don't know* are also discussed as a separate issue. These follow Diani's (2004) DM choice, namely *well, oh, I mean* and *you know*.

The hypothesis is that Polish EFL speakers underuse some functions of discourse markers and that DM clustering occurs much less frequently in Polish EFL speech than in native English speech and learner range of DMs is significantly narrower due to insufficient vocabulary skills.

Data and methodology

In the first part, the investigation follows Biber et al's (1999) corpus-driven 'recurrent word combination' method and the most frequent three-word unit is extracted from the corpus of native and non-native speech. In both cases it is apparently *I don't know*. It is on the basis of this sequence that the present analysis is conducted.

The present analysis is based on Contrastive Interlanguage Analysis (Granger 1996). Corpora used for the comparison are the Polish subcorpus of the Louvain International Database of Spoken English Interlanguage (PLINDSEI) (see: Jendryczka-Wierszycka 2006) referred later to as non-native English speaker (NNS) corpus and the Louvain Corpus of Native English Conversation (LOCNEC) (see: De Cock 2004) as an English native speech (NS) reference corpus. The computer tool used for extracting most frequent DMs and their collocates is the WordSmith Tools software (Scott 1998).

I don't know – pragmatic functions

I don't know is generally taken as reply to an information question. This meaning and function is expressed in the NS and NNS corpora in literal answers - usually when the interviewee (B) is asking the interviewer (A) whether she was familiar with a particular movie. Tsui (1991) argued the conversational environment may be composed of other elements than information question only. What emerges from the present data confirms Tsui's (1991) findings. Namely, only 36% of all *I don't know* occurrences are used in the literal meaning of the phrase in the native English data, and even less, i.e. 28% in the learner data. In her research, Tsui (1991) encountered six categories of pragmatic meaning of the sequence *I don't know*. Her classification served as a springboard for learner data investigation. In line with what had been suspected in the present research, some categories are vastly underrepresented in the NNS corpus. One not projected finding is, that one category is totally absent from both corpora, most possibly due to data character (Pulcini and Furiasi 2004). Let us now proceed to the very categories. They will all be described in relation to native English data one by one.

Avoidance of making an assessment

As Tsui (1991: 610) puts it, “speakers have, or claim to have, knowledge of what they are assessing. Therefore one way of declining to make an assessment is to claim insufficient knowledge of the referent, hence denying the proper basis for its production.” The following excerpt exemplifies this phenomenon:

<B> and you know they are you know just like .. it's very it's a very beautiful city but it . definitely it is not . [Venice uh <\B>

<A> [it is famous <\A>

<A> to talk about uh: Amsterdam and also Wroclaw [as Venice <\A>

<B> [*I don't know* <\B>

<B> I've never been to Am= Amsterdam so I [*I don't know* <\B>

Avoidance of making an assessment; PLINDSEI_48

When it comes to the comparison of occurrences in NS and NNS, there is no significant difference in this category whatsoever, therefore no fear or worry for teaching implications.

Preface to a disagreement

The next use of *I don't know* brings about considerable changes in terms of the NS – NNS comparison of use. Namely, NS use a significantly greater amount of the phrase *I don't know* in order to introduce a disagreement delicately. Consider the following example:

<B> know well I I could read more or less this those Arabic scripts those Arabic letters <\B>

<A> really <\A>

<B> yeah uhm so you know reading the names of stations for instance in a subway wasn't a big problem for me but <\B>

<A> mm impressive <\A>

<B> well I don't know I had to I don't know somehow get prepared [to <\B>

<A> mhm <\A>

<B> to to this to this visit food uh:you know the Tuttenhamon's <?> curse the food <\B>

Preface to a disagreement; PLINDSEI_12

Tsui (1991: 611) believes this use of *I don't know* to be expressed for the purpose of mitigating the face-threatening effect of the disagreement. In the example below, B was describing how her landlords ate *ser smazony* (a regional Polish and German cheese) every day, which she therefore generalized to be typical of the whole region in which she lived at that time:

<A> never met any any person that would eat that every day I mean that's I don't think that's anything special for Poznan <\A>

<B> mm <clicks> well I don't know I I er didn't . notice ser smazony or or . or anything like that in Olsztyn <\B>

Mitigating face-threatening effect; PLINDSEI_20

The very use of Polish name for the kind of cheese and the speaker's reluctance to describe it in English may be a want of pointing to this invention as something extraordinary. By saying *I don't know* before claiming that nowhere else in Poland had she seen such cheese, she is delicately introducing her disagreement with A's claim of the cheese being nothing special in Poznan. Notice the use of *well* in front of *I don't know*, which will be discussed in detail below.

Avoidance of an explicit disagreement

This category is not present at all, probably due to the character of the data. When hardly any opinion questions are asked, there is little possibility for *I don't know* constituting a turn on its own, which is, according to Tsui (1991: 612), a requirement for this phrase to be carrying the meaning of explicit disagreement.

Avoidance of commitment

In this use, "a compliance or rejection is relevant" (Tsui 1991: 617). By prefacing one's opinion with *I don't know*, the speaker is also claiming insufficient knowledge of the topic, and they avoid direct, dispreferred refusal. As for the use of this sense of *I don't know*, Tsui (1991: 617) claims its frequent exploitation "in response to invitations, offers, requests for permissions, or any other speech acts which solicit commitment from the speaker." In NNS data, there are hardly any requests for permission, any invitations or offers. Still, avoidance of commitment can be observed in the data.

<A> maybe then you would have been happier more [you know <\A>

<B> [uhu <\B>

<A> but in this respect I guess for you to be happy I don't know I'm not the person [to <\A>

<B> [yeah <\B>

<A> to be sure but I guess it will be good to start anything even a two year course in psychology I don't [know <\A>

Avoidance of commitment; PLINDSEI_36

Since the speaker avoids giving a positive evaluation of herself as an expert in counseling, in the example above, the modesty maxim (Leech 1983) is in operation.

Minimization of impolite beliefs

Although there was not much floor for impolite beliefs at all in the data examined (only one occurrence in NNS data and three occurrences in NS), it did happen in both corpora. In the examples below, we can witness a negative assessment of a third party and of self. The *I don't know* prefacing the assessment minimizes the face-threatening effect of the negative assessment.

<A> but apart from the food you don't like Britain <\A>

<B> no no . it's just so miserable and . everyone's every one's miserable over here <\B>

<A> oh great <\A>

<B> no it's an exaggeration no I just don't I don't know compared to <X> countries I just don't like it here <\B>

Minimization of impolite beliefs; LOCNEC_35

The interviewee clearly withdraws from their opinion when they hear A's reaction. Therefore they utter *I don't know* immediately prefacing a negative assessment of Britain, i.e. the country where the interview took place, which is why this use may be qualified as impolite belief.

Marker of uncertainty

Seemingly, the majority of pragmatic use is occupied by the expression of uncertainty, most probably due to learner uncertainty, that is not being able to call objects precisely or not being able to express oneself precisely.

<A> but you do not feel sorry that you're not with the[i:] other guy <\A>

<B> <clicks> no . [not at <\B>

<A> [<X> <\A>

<B> all er maybe I was eh: not maybe but for sure I was er . mm <clicks> I feel erm . I feel erm . I don't know .. maybe confused and er because I I did brake up with my: er boyfriend to be with with that guy <\B>

Marker of uncertainty; PLINDSEI_20

What is very characteristic of the data researched, uncertain future plans are frequently expressed prefaced with *I don't know* as, in line with interview guidelines, the subjects were frequently asked about their plans after their graduation:

in my first year that I decided to: to major in that erm .. *I don't know* I I think I'd like to teach English abroad at some stage

Graduation plans; LOCNEC_17

Filler and hesitation marker

Sometimes *I don't know* is used even exclusively as a filler, especially by the person who mainly listens (here A), or a hesitation marker, and frequently in accompanied by other fillers.

<B> [I mean <\B>

<B> he did have one but he didn't [<XXX> <\A>

<A> [probably yeah yeah maybe <\A>

<B> [I mean <\B>

<B> he did have one but he didn't [<XXX> <\B>

<A> [probably yeah yeah maybe <\A>

<A> I mean [maybe <\A>

<B> [<X> <\B>

<A> this man was really strong [emotionally . I don't know <\A>

Filler and hesitation marker; PLINDSEI_35

It is actually the filling of hesitation moments in a learner's speech that prevails in all uses of *I don't know*. To be precise, nearly 60% of all the uses are marked by speaker hesitation in NNS while in NS it is only 40%. The statistical difference is significant with 6.81 critical value equal at p level equal 0.05 (Rayson 2008). Therefore, the suspicion that one of the DM functions of *I don't know* is underused, is not without reason. Apparently, only 13% of all *I don't know* uses in NNS, compared to 23% in NS, are used as discourse markers and not hesitations. What emerges from the data is that overall, the underuse of pragmatic meanings of *I don't know* is evident in two categories of Tsui (1991): *preface to a disagreement* and *expression of uncertainty*. Some teaching implications are pointed to at the end of this article.

***I don't know* and other Discourse Markers**

Why look at neighboring DMs at all? It is simply to investigate "how DMs (Schiffrin 1987) occurring in conjunction with *I don't know* affect its function" (Diani 2004: 158).

I have decided to look at the vicinity of DMs in terms of Jones and Sinclair's (1974) span for identifying collocation, namely “four words on either side of the node word (the node word being the word under investigation)” (Hoey 2005: 4)

It was first the DMs that Diani (2004) found most frequently (immediately) co-occurring with *I don't know* that were analyzed in the present research. These are: *well, oh, I mean* and *you know*. The table below presents the exact numbers of their occurrence in NS and NNS and Diani's corpora and also in Diani's corpora as normalized to NNS corpus in terms of its size.

	Diani's work	Diani norm	NNS[45]	left	right	NS	left	right
well	194	=36,82	9 / – 2.57[46]	5	4	16	12	4
oh	38	= 7,1	7 / + 0.61	3	4	4	3	1
I mean	85	=15,88	0 / – 7.33	---	---	5	2	3
You know	23	= 4,29	0	---	---	0	---	---
All *I don't know*	1 114	=208,22	225 / + 1,11	---	---	208	---	---
Corpus size in tokens	2 000 000[47]		115 891			118 555		

Table 1: DMs most frequently co-occurring with *I don't know* (after Diani, 2004)

The plus and minus marks next to the words in the table indicate their over- or underuse in the NNS corpus in reference to the NS corpus and where a digital value appears next to a plus or a minus, it shows the critical value pertaining to statistical significance of the difference (see footnote for details). Inscriptions saying “left” and “right” next to a corpus name serve as labels for the number of occurrences of a given word or phrase as the left or right collocate of the node phrase *I don't know.*

Well

Halliday and Hasan (1976: 269) claim that *well* “serves to indicate that what follows is in fact a response to what has preceded”. (Tsui 1991: 617) notes that if accompanied by a hesitation marker or conversation filler, e.g. *well, I don't know* conveys discomfort and reluctance. Consider the following example:

<B> know well I I could read more or less this those Arabic scripts those Arabic letters <\B>

<A> really <\A>

<B> yeah uhm so you know reading the names of stations for instance in a subway wasn't a big problem for me but <\B>

<A> mm impressive <\A>

<B> <she sighs>well I don't know I had to I don't know somehow get prepared [to <\B>

<A> mhm <\A>

<B> to to this to this visit

well; PLINDSEI_12

While the NNS use of *well* occurring in the neighborhood of *I don't know* is underused, the critical value of the underused item is less than 3.84 (Rayson, 2008) which means the difference is not statistically significant if p < 0.05 is taken as a measurement for statistical significance. Therefore, learners do not seem to have problems with this modifier.

[45] occurrence number/ critical value with over- or underuse

[46] critical value: (from http://ucrel.lancs.ac.uk/llwizard.html) :The higher the G2 value, the more significant is the difference between two frequency scores. For these tables, a G2 of 3.8 or higher is significant at the level of p < 0.05 and a G2 of 6.6 or higher is significant at p < 0.01, 95th percentile; 5% level; p < 0.05; critical value = 3.84, 99th percentile; 1% level; p < 0.01; critical value = 6.63, 99.9th percentile; 0.1% level; p < 0.001; critical value = 10.83, 99.99th percentile; 0.01% level; p < 0.0001; critical value = 15.13

[47] 133 texts with everyday casual conversations

Oh

Schiffrin (1987: 89) notes that *oh* marks unanticipated new information which is in the focus of speaker's attention, thus potentially becoming the focus of hearer's attention. Fox Tree and Schrock (1999: 281) also quote another stand, the one of Heritage (1984), where the author proposed that speakers say “*oh* to let their addressees know that the speaker’s model of the communicative exchange is undergoing a change of state”.

<A> wha= what's the name of it <A>

<B> uh: .. well ca= you can say it in Polish cause I don't know [there's a lot of <?> <\B>

<A> [yeah sure <\A>

<B> <name of a group> <\B>

<A> oh I don't know this one <\A>

<B> yeah I mean well it's not er well very well known as far as I know so <\B>

oh; PLINDSEI_47

When it comes to NNS - NS use comparison, we find learner overuse, in turn, but again below the level of statistical significance.

I mean and *you know*

Schiffrin (1987) notes that the basic meaning of *I mean*’s is to forewarn upcoming adjustments. As Diani (2004: 167) puts it, such a marker increases 'tentativeness'”.

<B> yeah so erm I didn't know anybody and .. so I was a bit worried about that <\B>

<A> [mhm <\A>

<B> [but <X> I mean I soon made friends <X> erm everybody in my corridor's been very nice <\B>

I mean; LOCNEC_28

You know, in turn, may “function as an invitation to acknowledge a new piece of information” (Diani, 2004: 169).

> <B> yeah we did like everything you know we went to like <X> theme parks and .. did everything generally touristy things [you know <\B>

you know; LOCNEC_44

While *I mean* is vastly underused (critical value equals 14.66), mainly due to the fact that it is not used at all as a neighboring DM in NNS speech, *you know* does not occur in either corpus studied. This may stem from the interview character of the data or from the relation with the speaker, whom the interviewees may not have wanted to invite to acknowledge any information they presented.

As can be observed, apart from *I mean*, Diani's (2004) collocates of *I don't know* did not occur to be the most frequent in NS corpus analyzed here, therefore those most frequent DMs searched for in both NS and NNS corpus were taken into consideration. Those most interesting (due to their statistically significant over- and underuse in NNS corpus in relation to NS corpus) are presented in the table below:

	PLINDSEI	left	right	NS	left	right
yeah	21 / – 23.64	10	11	61	47	14
er	62 / + 40.39	52	10	9	8	1
mhm	40 / + 52,37	30	10	--	--	--
like	34 / + 5,29	25	9	16	11	5
eh	30 / + 39.28	22	8	--	--	--
maybe	28 / + 11.90	13	15	7	4	3

Table 2: DMs most frequently co-occurring with *I don't know* in PLINDSEI and LOCNEC

As emerges from the table above, *yeah* as a DM is largely underrepresented in NNS data while *er, eh, mhm, like* and *maybe* are overrepresented, which is especially striking when all unlexicalized hesitation markers are counted together (critical value = + 23.87). This may stem from learner uncertainty in their language use and their inability to fill the pauses with a lexicalized word or phrase.

Implications for Teaching

Expressions characteristic of spoken language and particular spoken language uses are slowly but surely finding their way into learner dictionaries and foreign language course books. Along the lines, the hesitation function of some DMs like *you know* or *I don't know* is presented in the dictionary mentioned. (cf. also *New Opportunities Upper Intermediate* course book for this and other spoken language phenomena). Learners do not seem to have trouble using the expression *I don't know* when in lack of a word or when hesitating. Yet, they do not seem to have recognized the variety of pragmatic uses of the expression. Unfortunately, the dictionary does not help them to do so. It is therefore believed that enriching dictionary resources with the pragmatic senses of *I don't know* will help them recognize the multitude of meanings and functions and appreciate that they are getting closer to the native English model also in terms of the underestimated pragmatics.

Apart from dictionaries, Romero Trillo (2002) concludes that “there is a different rate of development for the grammatical and the pragmatic aspects of language in L2” and that the lack of pragmatic competence, rather than insufficient vocabulary skills, “leads to pragmatic fossilization and, possibly, to communicative failure in many cases.” He therefore stresses the importance of bringing “the consistent teaching of pragmatic markers to language instruction” and the fact that the teaching “has to be based on sound research studies that categorize and describe the use of these pragmatic elements on the basis of the Index of Pragmatic Use and of other mathematical and statistical methods” (Romero Trillo 2002).

References

Biber, D., Johansson, S., Leech, G., Conrad, S. and **Finnegan, E.** (eds.). 1999 *Longman Grammar of Spoken and Written English*. Harlow: Longman.

De Cock, S. 2004. “Preferred sequences of words in NS and NNS speech.” *Belgian Journal of English Language and Literatures (BELL), New Series* 2: 225-246.

Diani, G. 2004. “The discourse functions of *I don't know* in English conversation”. In *Discourse Patterns in Spoken and Written Corpora,* K. Aijmer, Stenström, A.-B. (eds.). Amsterdam : John Benjamins, 157-171.

Fox Tree, J. E. and Schrock, J. C. 1999. “Discourse Markers in Spontaneous Speech:

Oh What a Difference an Oh Makes.” *Journal of Memory and Language* 40: 280–295.

Fox Tree, J. E. and Schrock, J. C. 2002. “Basic meanings of *you know* and *I mean*.” *Journal of Pragmatics* 34: 727–747.

Granger, S. 1996. “From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora”. In *Languages in Contrast. Textbased Cross-linguistic Studies* ,K. Aijmer, Altenberg, B. and Johansson, M. (eds.). Lund: Lund University Press, 37-51.

Halliday, M.A.K. and **Hasan, R.** 1976. *Cohesion in English*. London: Longman.

Harris, M, Mower, D. and Sikorzyńska, A. 2006. *New Opportunities Upper Intermediate.* Harlow: Longman.

Hoey, M. 2005. *Lexical priming: A new theory of words and language*. London: Routledge

Jendryczka-Wierszycka, J. 2006. *Lexical bundles in Polish learner speech: a study based on the PLINDSEI corpus of spoken learner English.* (unpublished M.A. thesis)

Jones, S. and Sinclair, J. M. 1974. "'English lexical collocations'" *Cahiers de Lexicologie* 24: 15-61.

Jucker, A.H. and Smith, S.W. 1998. “And people just you know like ‘wow’: Discourse markers as negotiating strategies.” In *Discourse Markers: Descriptions and* Theory, A.H. Jucker and Ziv, Y. (eds.).. John Benjamins, Philadelphia, 171–201.

Pulcini, V., Furiassi, C. 2004. “Spoken interaction and discourse markers in a corpus of learner English”. In *Corpora and Discourse,* A. Partington, J. Morley and L. Haarman (eds.). Bern: Peter Lang, 107-123.

Rayson, P. 2008. *Log-likelihood calculator*. http://ucrel.lancs.ac.uk/llwizard.html [Access date 26/05/2008]

Romero Trillo, J. 2002, “The pragmatic fossilization of discourse markers in non-native speakers of English” *Journal of Pragmatics* 34: 769-784.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Scott, M. 1998. *WordSmith Tools Manual, version 3.0.* Oxford: Oxford University Press. http://www.lexically.net/wordsmith/ [Access date 26/05/2008]

Summers, D. (ed.) 2006. *Longman Exams Dictionary*. Harlow: Longman.

Tsui, A.B.M. 1991 “The pragmatic functions of ‘I don't know'”. *Text* 11/4: 607-622.

A GUIDED COLLABORATION TOOL FOR ONLINE CONCORDANCING WITH EFL EAP LEARNERS

Przemysław Kaszubski[48]

Abstract

IFAConc, or IFA Concordancer (http://ifa.amu.edu.pl/~ifaconc), is an online concordancing environment being developed to help students of the English department (IFA) of Adam Mickiewicz University, Poznań, to explore, gather, annotate, discuss and share facts of language use pertaining to their language needs as academic writers. Current users can access three major registers – English for General Purposes (EGP), English for General Academic Purposes (EGAP) and English for Specific Academic Purposes (ESAP), as well as learner material available for the latter two ranges. It is possible to upload personal collections. All the searches save as unique URL links in the registered user's personal History. The searches can be annotated, also collaboratively with the Teacher-Administrator of the tool, and explored independently later. To further enable what I call top-down concordancing, a range of external resources interlinked with the IFAConc searches are being built. Given the linguistics-oriented profile of the student target audience, effort is made to bind the overall system to a theoretical platform – Michael Hoey's lexical priming.

The aim of this in-progress report is to outline the foundations and current state of the tool and to summarise its first pre-experimental in-course applications.

Keywords: concordancing, data-driven-learning, corpus linguistics, academic writing, constructionism

Introduction

Concordancing remains largely a bottom-up, word-based activity, whereas – as recently reminded by Coxhead (2008) and Hyland and Tse (2007) – insofar as learners' production needs are concerned, a wordlist perspective is largely unhelpful. Instead, due attention needs to be paid to preferred patterns of co-textual co-selection, which, though, often discriminate between discoursal specialisations, re-opening the split between the so-called General and Specific Academic Purposes (cf. Hyland 2006).

Recently, DDL practitioners have been pointing out the unfulfilled promise of the 'cut-out-the-middleman' approach. Despite its potential for raising language awareness and inducing learning effects (e.g. Horst and Cobb 2001), technical and practical complications often mar regular application and testing. Probably the most difficult to surmount is the problem of time expenditure. The potential number of patterns that merit exploration is vast, while users require pre-training and, even after it, conduct analyses taking prohibitively long to complete (see a discussion in Kaszubski 2006). There thus arises an urgent need to develop tools and solutions that would aid practical application, in both teaching and self-study modes, and enable more widespread testing of DDL effectiveness, in this case in the EAP context.

My own Holy Grail, over the years of corpus work, has been to obtain a handy DDL tool seamlessly integrated with the content and methodology of my writing classes, which would be usable for both me and my students, reasonably powerful, relevant, friendly, motivating, customisable and freely available. In the paper cited above, I identified some desirable features of a pedagogic concordancer that, I argued, would suit the needs of EAP writing. Among the points raised was online access, easy parameters adjustment (e.g. selection and de-selection of corpora), the presence of a guided or 'default route' facilitating the intake of necessary technical and linguistic knowledge, and compatibility with the constructionist trends in CALL (p.171).

The IFA Concordancer project (http://ifa.amu.edu.pl/~ifaconc) is an attempt to build an environment meeting these and other demands. It is an authorial tool – so far developed without external funding support in collaboration with various student groups at both the organisational and exploratory stages (corpus compilation and pre-processing, text mark-up, MA research, programming, etc.). Equipped with monitoring facilities, IFAConc promises to be a source of knowledge for probing and optimising concordancing activities performed with and for EAP students. The first fruits of are emerging, and are briefly presented in the second part of this paper. However, most of the work is on-going, and detailed results cannot yet be reported.

[48] Przemysław Kaszubski, assistant professor at the School of English, Adam Mickiewicz University, is a teacher of academic writing at both general and specific levels. He has been exploring corpora for research and pedagogic purposes for over ten years, recently focusing on the development and testing of customised web-based concordancing tools to assist EAP writing instruction.

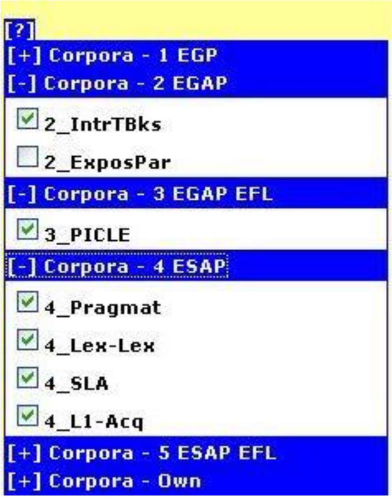
IFAConc – inspirations and response

The essential pillars of the IFAConc approach include the following:

EGAP and ESAP needs

EAP is traditionally conceptualised as a type of ESP. Aston (1998) perhaps first demonstrated how accessing properly stratified collections of texts – less and more specific – can enable teachers and students to attend to, and compare, general and specialist aspects of language use. His position is important, since, as mentioned, the world of EAP is divided over the issue of specialisation.

IFAConc aims to respond to these dilemmas by offering a range of corpora that sample various academic, quasi-academic, and 'pre-academic' (general) domains, as deemed relevant to the local, disciplinary make-up of the Poznań School of English.[49] One interface (admitting some group-level tweaking) is offered for both EGAP-level students (1st- and 2nd-year undergraduates) and ESAP-level students (3rd-year undergraduates, 1st-year graduate students). The benefit is that: 1) EGAP students are given a chance to immediately become acculturated to the linguistic aspects of disciplinary diversity; 2) ESAP students, in turn, may study specialist disciplinary patterns also in the context of the general academic and 'pre-academic' domains. The spectrum can be widened by adding corpora of one's own.



[?]

[+] Corpora - 1 EGP

[-] Corpora - 2 EGAP

☑ 2_IntrTBks

□ 2_ExposPar

[-] Corpora - 3 EGAP EFL

☑ 3_PICLE

[-] Corpora - 4 ESAP

☑ 4_Pragmat

☑ 4_Lex-Lex

☑ 4_SLA

☑ 4_L1-Acq

[+] Corpora - 5 ESAP EFL

[+] Corpora - Own

Figure 1: The IFAConc corpora selection interface

Tim Johns EAP DDL and Kibbitzer

Tim Johns' pioneering work is well-known. One resource particularly vital for the teaching and learning of academic writing with the DDL method are his Kibbitzer pages (Johns 2000), which document discoveries made during one-to-one consultations with international EAP students. A great source of contextualised knowledge, their possible weakness is that the concordance illustrations are static and not inviting a critical response from the student. IFAConc overcomes this limitation by offering the possibility of illustrating annotations and commentaries with web links that reproduce the search interface as well as the pertinent result.

Cobb – concordancing and learner writers

Tom Cobb's work has been highly influential in: 1) showing how concordances can surpass dictionaries as sources of reference; 2) pointing out that, while difficult in use, concordances enforce deeper-level processing promoting learning; and, perhaps most importantly, in 3) demonstrating that pre-cast web-links to specific KWiC

[49] The corpora are small (most within 200-300,000 words) and include excerpts of texts from the Web and from resources available to AMU under institutional subscription, thus broadly falling under 'fair use'. The corpora considered public are available in the sampler version of IFAConc.

displays can be applied as corrective feedback for learner writers (Gaskell and Cobb 2004). However, when taking a student-researcher point of view, the concordance links used are, again, static in view, mainly focused on simple collocation patterns, and generic or mono-varietal in terms of discourse. As such, I would argue, they might not be helpful enough to the academic learner writer.

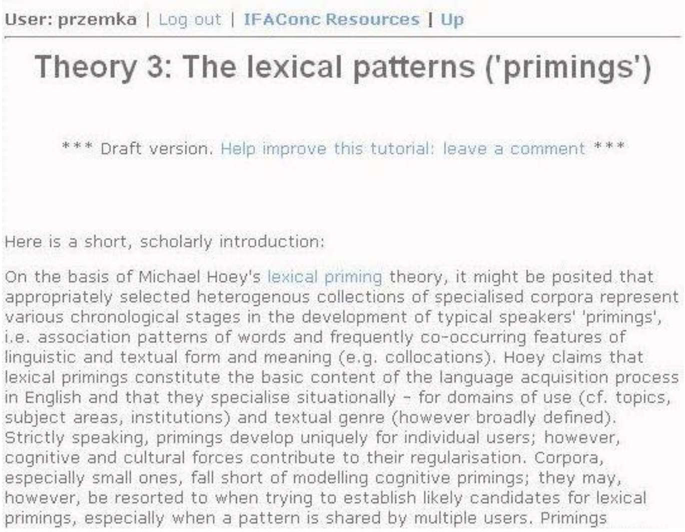
Widdowson, Aston, Gavioli – authentication, personalisation

Learners' work with corpora may pose problems with noticing and salience. Hunston (2002) warns of the decontextualisation of corpus data: in order for excerpts of others' texts to be 'meaningful' for the concordancing user, they need to be re-contextualised and 'authenticated' (e.g. Widdowson 2000). This is not necessarily a very difficult task if teachers and specialists are around, but one required for optimising acquisition. In a similar vein, Gavioli (2005) shows the distinction and transition between 'samples' of data and (typical) 'examples' of use. She also points out, in an earlier study with Aston (Gavioli and Aston 2001), the need to creatively adapt and adjust evidence from corpora to one's own productive needs.

One solution to address the questions of authentication may be the provision of rich, locally authenticated corpus annotations (e.g. Braun 2006). The IFAConc approach partly looks this way (future projects are planned for XML tagging of new information layers); however, given that corpus tagging can be labour-intensive and time-costly, other solutions have been sought. Firstly, it is the selection of the relevant corpora that facilitates meaningful and readily re-contextualised input: included need to be not only the target registers and genres, but also 'homely' (Johns 1997) and familiar ones. Thus, for example, EFL sub-parts of the EGAP and ESAP registers (i.e. students' essays and MA theses) are on offer in IFAConc. In addition (an option which has not yet been tested in practice) uploading personal collections for comparison with the standard corpora encourages 'self-concordancing' (cf. Coniam 2004), stimulating a more personal response to concordance results. Lastly, the integrated search annotation facility, used personally or collaboratively with the teacher, is also a powerful tool for authentication.

Hoey – lexical priming

Partly due to the anticipated above-average linguistic awareness of the IFA students, and partly due to the pedagogic orientation of the tool, the IFA Concordancer is being laid out with a recommended, corpus-friendly theory in mind. Probably the most complete and inclusive proposal is that of Hoey (2005), who speaks of 'primings', or associative loadings, which words (or other linguistic units) take on from their repeated contexts and to which users are naturally exposed. According to Hoey, lexical primings include both local and – as is hypothesised – textual associations: collocation, colligation, semantic association, pragmatic association, textual collocation, textual colligation and textual semantic association. Closer discussion of these concepts is beyond the scope here; also application of the whole theory is beyond IFAConc's current technological reach. Due to the small sizes of the corpora a certain amount of reinterpretation is necessary, some of which is offered on the IFAConc Resources – Tutorial pages.



Theory 3: The lexical patterns ('primings')

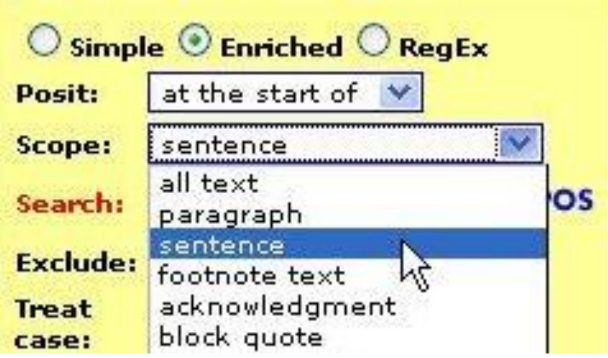
*** Draft version. Help improve this tutorial: leave a comment ***

Here is a short, scholarly introduction:

On the basis of Michael Hoey's lexical priming theory, it might be posited that appropriately selected heterogenous collections of specialised corpora represent various chronological stages in the development of typical speakers' 'primings', i.e. association patterns of words and frequently co-occurring features of linguistic and textual form and meaning (e.g. collocations). Hoey claims that lexical primings constitute the basic content of the language acquisition process in English and that they specialise situationally – for domains of use (cf. topics, subject areas, institutions) and textual genre (however broadly defined). Strictly speaking, primings develop uniquely for individual users; however, cognitive and cultural forces contribute to their regularisation. Corpora, especially small ones, fall short of modelling cognitive primings; they may, however, be resorted to when trying to establish likely candidates for lexical primings, especially when a pattern is shared by multiple users. Primings

Figure 2: IFAConc Tutorial pages – an Introduction to lexical priming

More directly, textual colligation, i.e. the tendency for words and patterns to associate with certain positions within text ‘chunks’, is featured in the corpus search mechanism:



○ Simple ● Enriched ○ RegEx

Posit: at the start of

Scope: sentence

Figure 3: Textual colligation search in IFAConc

All the corpora are marked with sentence and paragraph breaks; the ESAP corpora – thanks to the effort of MA students who undertook the XML encoding – have additionally been marked up with other academically important textual features: footnotes, section headings, extended quotations, etc.

Constructionist learning

As Tim Johns famously noted (Johns 1991: 5), students are often able to make more revealing observations than their teachers. The IFAConc search annotation system and the possibility of using it collaboratively with the teacher provide a fairly controlled but powerful constructivist environment. Owing to integration with the web, IFAConc has been used in connection with the Moodle course management system as well, and, more recently, with a blog engine and a wiki. The usefulness of the latter two is yet to be seen; so far, the blog has served as a public announcements outlet with occasional links to example findings.

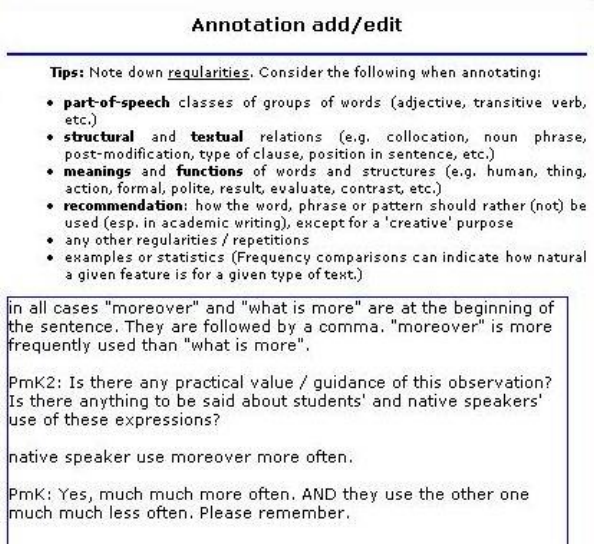
In addition, as already mentioned, seminar students are involved in the compiling, pre-processing, tagging and pre-testing stages.

Concordance monitoring

The IFAConc project also extends my earlier work in monitoring online concordancing, based on a tool called the PICLE concordancer. In a 2006 paper,[50] I provided a short analysis of the public and local search logs, and summarised questionnaire responses from a group of seminar students revising BA papers in linguistics. Among the general observations made was one that users who returned to the system were more likely to submit complex queries for collocations and patterns rather than for single words. This, it was thought, could imply these users’ higher ability of noticing, interpreting and applying corpus information, which, however, could not be studied individually due to the limited set of the logged parameters.

The decision was taken to create a friendlier tool that would encourage students to operate with patterns and hopefully steer them more swiftly towards becoming independent DDL researchers. IFAConc’s archiving and annotating system – History – was conceptualised as a tool for collaborative and teacher-monitored recording of findings, as well as – at a later stage – as a resource for organising focused and more immediately useful, concordancing sessions – ones in which the starting point would not necessarily be a form, but a meaning (or priming) category. While some amount of ‘top-down’ concordancing is enabled by powerful search syntax, the IFAConc History and other Resources containing grouped and annotated hyper-links hold a potential for being more effective concordancing starters. These hypotheses yet remain to be tested.

[50] Kaszubski (2006). An English version of this paper is available on personal contact.



Annotation add/edit

Tips: Note down regularities. Consider the following when annotating:

- **part-of-speech** classes of groups of words (adjective, transitive verb, etc.)
- **structural** and **textual** relations (e.g. collocation, noun phrase, post-modification, type of clause, position in sentence, etc.)
- **meanings** and **functions** of words and structures (e.g. human, thing, action, formal, polite, result, evaluate, contrast, etc.)
- **recommendation**: how the word, phrase or pattern should rather (not) be used (esp. in academic writing), except for a 'creative' purpose
- any other regularities / repetitions
- examples or statistics (Frequency comparisons can indicate how natural a given feature is for a given type of text.)

in all cases "moreover" and "what is more" are at the beginning of the sentence. They are followed by a comma. "moreover" is more frequently used than "what is more".

PmK2: Is there any practical value / guidance of this observation? Is there anything to be said about students' and native speakers' use of these expressions?

native speaker use moreover more often.

PmK: Yes, much much more often. AND they use the other one much much less often. Please remember.

Figure 4: The IFAConc search annotation window (a recent version) in use

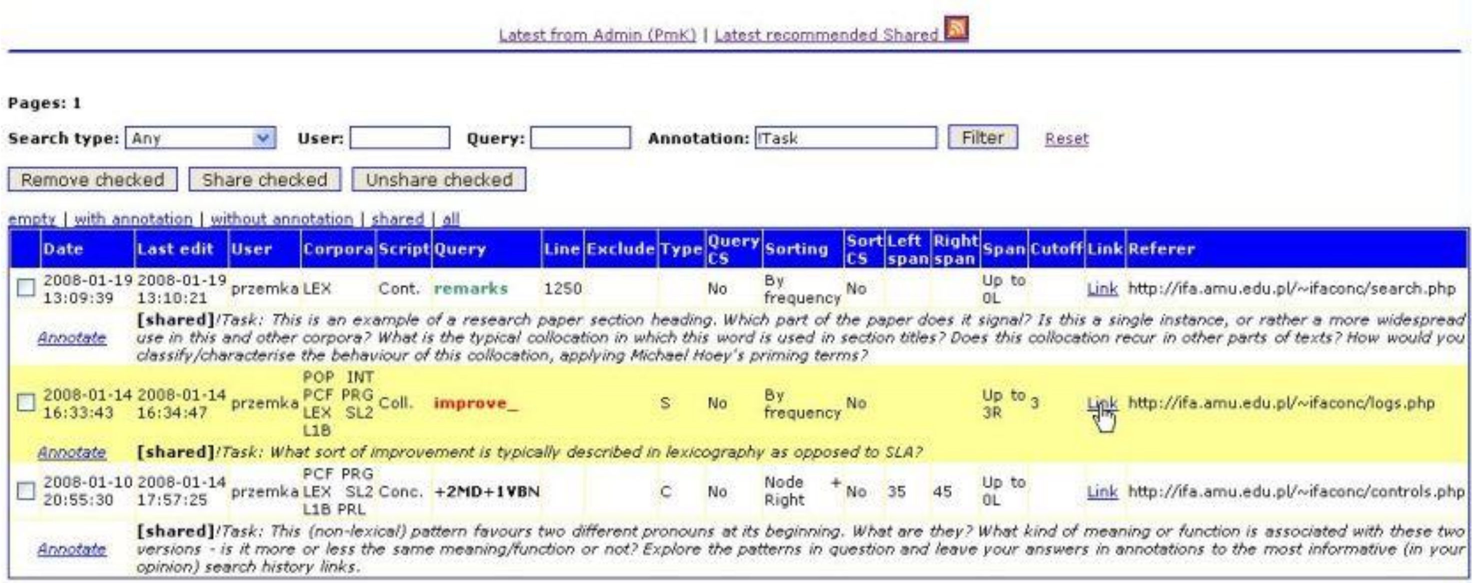
IFAConc pilot tests

In order to see to what extent IFAConc use would be practicable in a writing course situation, two pilot studies were undertaken, one performed with advanced (ESAP-level) students, the other with 1st-year EGAP students. The data are still being gathered and processed, so only a general overview can be offered below.

ESAP 'primer'

IFAConc was first tested with two small groups of participants of my one-semester seminar courses in corpus linguistics and concordance reading (Winter 2007-8, 14 students). Their higher-than-average familiarity with the corpus technology made them an unrepresentative sample of ESAP writing respondents, but a suitable team of testers at the tool launch stage. IFAConc was introduced towards the end of the course, during an approximately 30-minute computer lab activity. After that, two major tasks plus an evaluation questionnaire were administered as three successive weekly home-assignments. Some technical problems occasionally interfered (e.g. unexpected zero results revealed exceptions in the search script), and although generally eliminated quickly, may have affected students' performance and responses.

- In Task 1 the students worked with link-initiated searches provided from within the History interface. The students had to locate annotated queries containing specified key words, read instructions in the annotations, activate the search links, manipulate the concordances as needed and annotate their work with suitable answers. The search patterns were varied, illustrating disciplinary variation and textual positioning, among others (the latter not visible in the screenshot below, as the textual colligation parameters were added later to the History interface);
- In Task 2 the students were requested to run two or three of their own searches and to investigate and annotate those;
- In the Evaluation stage (in which ten students took part), ten closed (5-step Likert-scale) and two open-ended questions were used. Some of these were:



Latest from Admin (PmK) | Latest recommended Shared

Pages: 1

Search type: Any User: Query: Annotation: !Task Filter Reset

Remove checked | Share checked | Unshare checked

empty | with annotation | without annotation | shared | all

	Date	Last edit	User	Corpora	Script	Query	Line	Exclude	Type	Query CS	Sorting	Sort CS	Left span	Right span	Span	Cutoff	Link	Referer
□	2008-01-19 13:09:39	2008-01-19 13:10:21	przemka	LEX	Cont.	remarks	1250			No	By frequency	No			Up to 0L		Link	http://ifa.amu.edu.pl/~ifaconc/search.php
Annotate	[shared]!Task: This is an example of a research paper section heading. Which part of the paper does it signal? Is this a single instance, or rather a more widespread use in this and other corpora? What is the typical collocation in which this word is used in section titles? Does this collocation recur in other parts of texts? How would you classify/characterise the behaviour of this collocation, applying Michael Hoey's priming terms?																	
□	2008-01-14 16:33:43	2008-01-14 16:34:47	przemka	POP INT PCF PRG LEX SL2 L1B	Coll.	improve_			S	No	By frequency	No			Up to 3R	3	Link	http://ifa.amu.edu.pl/~ifaconc/logs.php
Annotate	[shared]!Task: What sort of improvement is typically described in lexicography as opposed to SLA?																	
□	2008-01-10 20:55:30	2008-01-14 17:57:25	przemka	PCF PRG LEX SL2 L1B PRL	Conc.	+2MD+1VBN			C	No	Node + Right	No	35	45	Up to 0L		Link	http://ifa.amu.edu.pl/~ifaconc/controls.php
Annotate	[shared]!Task: This (non-lexical) pattern favours two different pronouns at its beginning. What are they? What kind of meaning or function is associated with these two versions - is it more or less the same meaning/function or not? Explore the patterns in question and leave your answers in annotations to the most informative (in your opinion) search history links.																	

Figure 5: The ESAP pilot tasks administered as Shared entries through the IFAConc search History

3.) How useful do you find the search options implemented in IFAConc? (not useful – don't know – somewhat useful - useful – very useful)
4.) How useful did you find the link-driven tasks, i.e. the IFAConc activities initiated from annotated links? (scale as above)
9.) Which did you find *easier* to do: pursuing your own investigations or completing the link-driven investigations? (definitely former – perhaps former – no major difference – perhaps latter – definitely latter)

10.) Which did you find more *fruitful*: pursuing your own investigations or completing the link-driven investigations? (scale as above)

For the first two questions, a majority (6-7 votes) of the answers clustered in the Useful category, where they were expected. In the third question, a bimodal pattern emerged, with four students in each of the 'perhaps' categories; interestingly, the distribution correlated with group assignment, which indicates that a difference in the training procedure may have occurred (not yet identified). In the fourth question, five students were undecided, but as many as four chose the 'perhaps' and the 'definitely link-driven' options, which I found a highly encouraging result. Some students' appreciation for the link-led discoveries was also reflected in their open-ended responses.

The students' annotations were moderated towards the pedagogically useful and away from the linguistically abstract. Overall, this suggests that some students may have approached IFAConc as a serious language study tool rather than a pedagogical aid, which prompted me to add a disclaimer in one of the opening tutorials. Similarly, other students' remarks, also from the second study, have been used to post-edit training materials and resources.

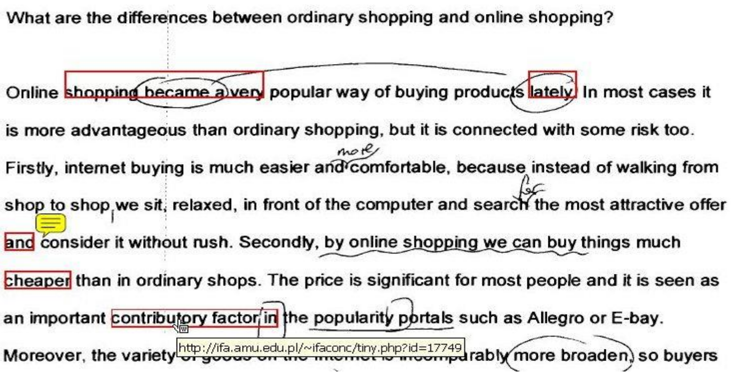
EGAP pilot - IFAConc links as feedback

The second pilot study is still ongoing, and may only be briefly recounted here. The originally conceived plan was to offer a group of 15 first-year BA-level participants of an expository writing course a series of activities gradually preparing them for the role of independent DDL researchers. Link-driven feedback was intended to be a motivating part of the 'breaking-in' process. Unfortunately, the level of the students proved considerably lower than the expected norm, forcing a re-focus on more basic types of problems, most of which were common grammatical errors rather than stylistic options (which many students had considerable problems noticing as they span across wider co-texts). In these circumstances, the major goal was transformed into one of optimising link-driven feedback. The complexity of the links offered (single words, lemmatised searches, POS searches and the like) was not controlled in any way, but spontaneously applied. A positive outcome of this is that a mini-base has now accumulated in the Administrator's History, which will be re-usable in the future, also as a repository of authentic training examples.[51]

A combination of manual and link-based feedback was established, with links occasionally balloon-tipped, as shown in the illustration below. Over the course of over two months in the Spring 2008 term, until the time of writing, students have been given links for 11 different assignments. Some global statistics are presented in the table below.

Relevant Teacher-Admin searches	c. 3000
Number of links offered as feedback	450 (incl. 27 single-word searches)
Total student searches	931
Students' feedback-link-driven searches	272
Student self-initiated searches	655 (incl. 307 single word searches)

[51] Integration with the error-tagged sample of the PICLE corpus may also be possible, as planned in the IFAConc early design stage (http://ifa.amu.edu.pl/~kprzemek/concord2adv/errors/errors.htm).



What are the differences between ordinary shopping and online shopping?

Online shopping became a very popular way of buying products lately. In most cases it is more advantageous than ordinary shopping, but it is connected with some risk too. Firstly, internet buying is much easier and more comfortable, because instead of walking from shop to shop we sit, relaxed, in front of the computer and search for the most attractive offer and consider it without rush. Secondly, by online shopping we can buy things much cheaper than in ordinary shops. The price is significant for most people and it is seen as an important contributory factor in the popularity portals such as Allegro or E-bay. Moreover, the variety of goods on the internet is incomparably more broaden, so buyers

http://ifa.amu.edu.pl/~ifaconc/tiny.php?id=17749

IFAConc-link-supported marking of students' texts

With the data still being processed, and the final questionnaire pending, only a few impressionistic observations have been made:

- The students seem to have divided into 3 groups – adopters, minimal-users, and refusers (the presence of the last category is a surprise, as some tasks were obligatory);
- Students who undertake more link-driven searches tend to develop a greater liking for self-initiated searches;
- The students are reluctant to annotate: despite continual encouragement, only one in ten searches carry annotations; most of these originate from obligatory tasks;
- The students have difficulty noticing less visible aspects of patterning (e.g. textual length, semantic characteristics, frequency differences), unless specifically guided by the co-annotating Teacher-Administrator.

An interim conclusion

Learners have more questions to ask than traditional reference tools can hold answers for (cf. Cobb 2003). The key to successful DDL lies in the effective and efficient training, as noted in a number of studies (e.g. O'Keeffe and Farr 2003). Research is needed to determine what kinds of tools, interfaces and approaches are the most amenable; it is perhaps best to test the tools and solutions in realistic quasi-experimental conditions. IFAConc is a tool developed to gradually illuminate some of these issues.

References

Aston, G. 1998. "What corpora for ESP?". In L'apprendimento linguistico all'universita: Le lingue speciali, M. Pavesi and G. Bernini (eds). Roma: Bulzoni. 205–226.

Braun, S. 2006. "ELISA: a pedagogically enriched corpus for language learning purposes." In Corpus technology and language pedagogy: new resources, new tools, new methods, S. Braun, K. Kohn and J. Mukherjee (eds). Frankfurt am Main: Peter Lang, 25-47.

Cobb, T. 2003. "Do corpus-based electronic dictionaries replace concordancers?" In Directions in CALL: Experience, experiments, evaluation, B. Morrison, G. Green and G. Motteram (eds) Polytechnic University, 179-206.

Coniam, D. 2004. "Concordancing oneself: constructing individual textual profiles." International Journal of Corpus Linguistics 9/2: 271-298.

Coxhead, A. 2008. "Phraseology and English for academic purposes: Challenges and opportunities." In Phraseology in Foreign Language Learning and Teaching, F. Meunier and S. Granger (eds). Amsterdam/Philadelphia: John Benjamins, 149-161.

Gaskell, D. and **Cobb, T.** 2004. "Can learners use concordance feedback for writing errors?" System 32/3: 301-19.

Gavioli, L. 2005. Exploring corpora for ESP learning. Amsterdam/Philadelphia: John Benjamins.

Hoey, M. 2005. Lexical priming: a new theory of words and language. London: Routledge.

Horst, M. and T. Cobb. 2001. "Growing academic vocabulary with a collaborative on-line database." Paper presented at IT-MELT'01, Polytechnic University of Hong Kong, June 2001.

Hunston, S. 2002. Corpora in applied linguistics. Cambridge: Cambridge University Press.

Hyland, K. 2006. English for Academic Purposes: An Advanced Resource Book. London: Routledge.

Hyland, K. and Tse, P. 2007. "Is there an academic vocabulary?" TESOL Quarterly 41/ 2: 235-253.

Johns, A. M. 1997. Text, role, and context: Developing academic literacies. Cambridge: Cambridge University Press.

Johns, T. 1991. "Should you be persuaded - two samples of data-driven learning materials", in Classroom concordancing, : T. Johns and P. King (eds.), Birmingham : Birmingham University, 1-13.

Johns, T. 2000. Tim Johns EAP Page. http://www.eisu2.bham.ac.uk/johnstf/timeap3.htm [Access date 15/02/2008].

Kaszubski, P. 2006. "Konkordancer internetowy w nauce języka: w stronę optymalizacji", in Korpusy w angielsko-polskim językoznawstwie kontrastywnym: teoria i praktyka, A. Duszak, E. Gajek, and U. Okulska (eds.), Kraków: Universitas, 329-359.

Kaszubski, P. 2006. "Web-based concordancing and ESAP writing", Poznań Studies in Contemporary Linguistics 41: 161-193.

O'Keeffe, A. and Farr, F. 2003. "Using language corpora in initial teacher education: Pedagogic issues and practical applications". TESOL Quarterly 37/3: 389-418.

Widdowson, H. G. "On the limitations of linguistics applied". Applied Linguistics 21/1: 3-25.

TRACING THE EMO SIDE OF LIFE. USING A CORPUS OF AN ALTERNATIVE YOUTH CULTURE DISCOURSE TO TEACH CULTURAL STUDIES

Bernhard Kettemann[52]

Abstract

This paper presents the plan for an investigation into the usefulness of corpus analysis, both as a method of investigation and a source of data for the teaching and learning of Cultural Studies. Four sets of data, a semantically categorized list of verbs following first person singular I, a keyword list comparison between the Emo corpus and the Bergen Corpus of London Teenage Language (COLT), a list of the most frequent content words and a concordances of alone, lonely, and on my own will be presented to the students and a questionnaire will check their reactions to this material. The research question is whether inductive autonomous learning with the help of corpus data can be argued to have any influence in the learning of Cultural Studies. The four levels of processing are supposed to show whether students can learn to interpret data culturally within a short time span if they have been exposed to a certain approach beforehand.

Keywords: Cultural Studies, Emo, corpus analysis, autonomous learning

Corpus analysis, both as a method of investigation and a source of data, has something to offer to the teaching and learning of Cultural Studies (= CS), as has already been demonstrated – albeit from very different angles and not always explicitly – by, among others, Eppler, Crawshaw and Clapham (2000), Minugh (2007), Kettemann and Marko (forthcoming). The present paper intends to explore this potential further.

Teaching Cultural Studies

The past two decades have seen the rise of CS not only as a subject in its own right, but also as a major component of philological disciplines. This development has gone hand in hand with a pedagogical reconceptualization of these disciplines. The major objective is no longer the mere top-down transfer of knowledge from the teacher to the students, but the focus has shifted towards awareness-based competencies. This necessitates pedagogical procedures no longer relying on a strict separation of teacher and student roles, reducing teachers' authority as a reservoir of explicit knowledge and at the same time strengthening students' confidence in their ability to find their own paths of understanding.

On the level of culture, this means moving away from factual knowledge of historical, geographic, political, economic and social details of different countries towards a socio-cultural awareness enabling students to become socio-cultural mediators and disseminators, roles that are likely to be at the core of what they will be doing professionally in the future. It seems plausible to assume that Cultural Studies as a field focusing on ideologies and narratives, i.e. systems of beliefs and attitudes, and their impact on a society's interpretation of the world seems better able to meet such needs than traditional 'facts & figures'-courses, which usually present snippets of information in a decontextualized form.

Cultural Studies courses, however, pose some serious pedagogical challenges due to the abstractness of the ideas presented at first sight seem further removed from students' lifeworlds than the said facts and figures.

[52] Bernhard Kettemann, Professor of English Linguistics at Graz and an early TaLCer, editor of Arbeiten aus Anglistik und Amerikanistik, co-editor of Moderne Sprachen, president or member of governing boards of VERBAL (the Austrian Association of Applied Linguistics), of VÖN (the Austrian Modern Languages Association), of AAUTE (the Austrian Association of University Teachers of English). Research interests in Corpus Linguistics, in Critical Discourse Analysis and Cultural Studies.

The role of corpus analysis in the teaching of Cultural Studies

What is the proposed role of corpus analysis? Cultural Studies emphasize the importance of language as means by which ideologies and narratives are disseminated across a society. Examining language in use thus means trying to understand how we make sense of the world and ourselves. Word lists, cluster lists, keyword lists, concordances, collocations, and data derived from these, e.g. on semantic domains and semantic prosodies, provide valuable tools for raising students' awareness, enabling them to explore cultural meanings and eventually to 'teach' cultural competence – whether as teachers or in another function – themselves.

A CS course integrating a corpus linguistic module should involve a two-step procedure:

Step I: Teacher input: Students work with corpus material and data specially prepared by the teacher.

Step II: Student-centred exploratory research: Students work with corpora provided by the teacher and later with their own corpora in an exploratory, data-driven fashion.

In my paper, I will concentrate on the first step. As an example of this awareness raising approach I will work with data from a corpus of English texts by Emos.

Emo is the label of an alternative youth culture characterised by introversion and withdrawal from an outside (adult) world perceived as unsympathetic, misunderstanding and demanding and the concomitant emphasis on negative and depressive moods and suicidal ideas. Externally, Emos show a preference for dark colours in clothes, hairstyle and make-up and for androgynous styles (cf. Kelley/Leslie 2007).

Emo discourse seems an appropriate topic for CS because it offers an ideal starting point for discussions about such important cultural concepts as lifestyles, identities, values, and the opposition alternative vs. mainstream. I also assume that students can be motivated to deal with this topic as Emo ideas will not be completely alien to a many of them.

The corpus

The corpus I will be using was compiled by Kerstin Florian for a paper in the seminar "The language of alternative lifestyles". She included a heterogeneous set of genres written and made publicly available by young people identifying as Emos, such as blogs, fashion and lifestyle articles, poems and song lyrics. Corpus linguistically speaking, the collection is thus very 'dirty'. But it can be argued that the corpus in its inconsistent composition represents the chaotic textual universe in which young people shape and enact their Emo identities.

The corpus comprises 141,614 word tokens. For the analysis, Wordsmith Tools 4.0 by Mike Scott (2006) was used.

The study

In the long version of this paper, to be presented at TaLC 2008 in Lisbon, I will report and discuss the results of a study to be carried out in June 2008 at the English Department of Karl-Franzens-University Graz. Its objective is to find out about students' interpretations of corpus data presented to them in a questionnaire.

The students are all participants in a course called "British and American Cultural Studies Foundation Course," which I have been teaching for several years. This is a first-year course of the English and American Studies programme at our university, which normally constitutes students' first contact with Cultural Studies.

The idea is to present the participants in the study with four sets of data and some interpretatory questions concerned with the construction of an Emo identity and an Emo perspective on life. The four sets of data represent declining degrees of processing, from highly processed to raw. The processing primarily involves selection of particular items (e.g. by using a stoplist), comparisons (e.g. by using keyword lists), and semantic categorization of words and phrases. A final evaluative part should provide me with feedback about the participants' perception of the tasks.

The main question behind the whole study is whether inductive autonomous learning with the help of corpus data can be argued to have any influence in the learning of Cultural Studies. The four levels of processing are supposed to show me whether students can learn to interpret data culturally within a short time span if they have been exposed to a certain approach beforehand.

The sets of data to be drawn upon are:

1. A semantically categorized list of verbs following first person singular *I*.
2. A keyword list comparison between the Emo corpus and the Bergen Corpus of London Teenage Language (COLT).
3. A list of the most frequent content words.
4. Concordances of *alone*, *lonely*, and *on my own*.

Method

All students in the course will get a questionnaire with the datasets given below, together with questions to be answered in writing and to be handed in as compulsory course work within a week. In order to see the full effect of the corpus data, I will not provide any additional information on Emos.

As a last part of the questionnaire, the participants in the study will be asked about general conclusions concerning the project, especially focusing on the role that language played in these. I am particularly interested in whether they have noticed a change in their approach to the use of the data from the first to the fourth set and in what ways the processing included on earlier stages has helped them in coping with input in later stages.

The questionnaires will be evaluated qualitatively and – the responses permitting – quantitatively.

Verbs following I

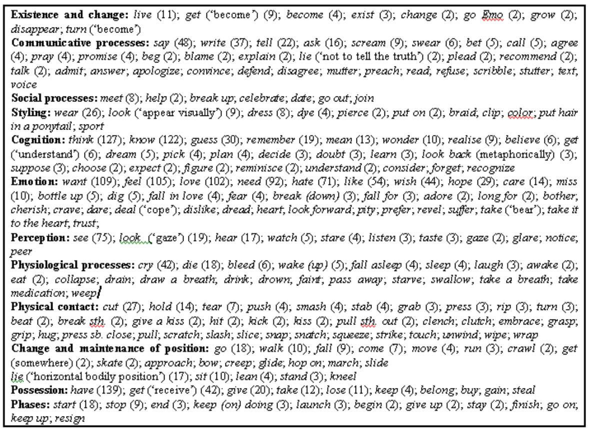
For the first set of data, I have chosen to provide students with quantitatively and qualitatively processed data on verbs which take the first person singular pronoun *I* as the subject (I limited the search to the first word directly following the latter) because what Emos say they are doing will be pivotal in their construction of their own identities.

Qualitative processing here means categorizing the verbs found according to the following classes.

- **Existence and change:** e.g. *exist*, *become*.
- **Communicative processes:** *say, promise, convince*.
- **Social processes:** e.g. *date*, *meet*.
- **Style:** processes to do with dressing, make-up and hairstyle, e.g. *wear*, *dye*.
- **Cognition:** processes concerned with thinking, e.g. *know*, *remember*.
- **Emotion:** affective and evaluative processes, e.g. *fear*, *enjoy*, *need*.
- **Perception:** processes of sensual perception, e.g. *see*, *hear*, *look at*.
- **Physiology:** active and reactive processes of the body, e.g. *eat*, *sleep*, *die*.
- **Physical contact:** processes of physical contact with people or objects, e.g. *touch*, *cut*, *kiss*.
- **Movement and static position:** processes of changing– e.g. *walk*, *come* – or keeping one's position – e.g. *lie*, *sit*.
- **Possession:** processes of having, getting or losing, e.g. *own*, *obtain*, *lose*.
- **Phases:** processes of starting, finishing or continuing, e.g. *stop*, *begin*, *keep on*.

Verbs that could not easily be assigned to any of the categories and verbs of minor semantic classes were not included.

Students will get these definitions and the following set of data.



Existence and change: *live* (11); *get* ('become') (9); *become* (4); *exist* (3); *change* (2); *go Emo* (2); *grow* (2); *disappear*; *turn* ('become')

Communicative processes: *say* (48); *write* (37); *tell* (22); *ask* (16); *scream* (9); *swear* (6); *bet* (5); *call* (5); *agree* (4); *pray* (4); *promise* (4); *beg* (2); *blame* (2); *explain* (2); *lie* ('not to tell the truth') (2); *plead* (2); *recommend* (2); *talk* (2); *admit*; *answer*; *apologize*; *convince*; *defend*; *disagree*; *mutter*; *preach*; *read*; *refuse*; *scribble*; *stutter*; *text*; *voice*

Social processes: *meet* (8); *help* (2); *break up*; *celebrate*; *date*; *go out*; *join*

Styling: *wear* (26); *look* ('appear visually') (9); *dress* (8); *dye* (4); *pierce* (2); *put on* (2); *braid*; *clip*; *color*; *put hair in a ponytail*; *sport*

Cognition: *think* (127); *know* (122); *guess* (30); *remember* (19); *mean* (13); *wonder* (10); *realise* (9); *believe* (6); *get* ('understand') (6); *dream* (5); *pick* (4); *plan* (4); *decide* (3); *doubt* (3); *learn* (3); *look back* (metaphorically) (3); *suppose* (3); *choose* (2); *expect* (2); *figure* (2); *reminisce* (2); *understand* (2); *consider*; *forget*; *recognize*

Emotion: *want* (109); *feel* (105); *love* (102); *need* (92); *hate* (71); *like* (54); *wish* (44); *hope* (29); *care* (14); *miss* (10); *bottle up* (5); *dig* (5); *fall in love* (4); *fear* (4); *break (down)* (3); *fall for* (3); *adore* (2); *long for* (2); *bother*; *cherish*; *crave*; *dare*; *deal* ('cope'); *dislike*; *dread*; *heart*; *look forward*; *pity*; *prefer*; *revel*; *suffer*; *take* ('bear'); *take it to the heart*; *trust*;

Perception: *see* (75); *look* ('gaze') (19); *hear* (17); *watch* (5); *stare* (4); *listen* (3); *taste* (3); *gaze* (2); *glare*; *notice*; *peer*

Physiological processes: *cry* (42); *die* (18); *bleed* (6); *wake (up)* (5); *fall asleep* (4); *sleep* (4); *laugh* (3); *awake* (2); *eat* (2); *collapse*; *drain*; *draw a breath*; *drink*; *drown*; *faint*; *pass away*; *starve*; *swallow*; *take a breath*; *take medication*; *weep*

Physical contact: *cut* (27); *hold* (14); *tear* (7); *push* (4); *smash* (4); *stab* (4); *grab* (3); *press* (3); *rip* (3); *turn* (3); *beat* (2); *break sth.* (2); *give a kiss* (2); *hit* (2); *kick* (2); *kiss* (2); *pull sth. out* (2); *clench*; *clutch*; *embrace*; *grasp*; *grip*; *hug*; *press sb. close*; *pull*; *scratch*; *slash*; *slice*; *snap*; *snatch*; *squeeze*; *strike*; *touch*; *unwind*; *wipe*; *wrap*

Change and maintenance of position: *go* (18); *walk* (10); *fall* (9); *come* (7); *move* (4); *run* (3); *crawl* (2); *get* (somewhere) (2); *skate* (2); *approach*; *bow*; *creep*; *glide*; *hop on*; *march*; *slide*
lie ('horizontal bodily position') (17); *sit* (10); *lean* (4); *stand* (3); *kneel*

Possession: *have* (139); *get* ('receive') (42); *give* (20); *take* (12); *lose* (11); *keep* (4); *belong*; *buy*; *gain*; *steal*

Phases: *start* (18); *stop* (9); *end* (3); *keep (on) doing* (3); *launch* (3); *begin* (2); *give up* (2); *stay* (2); *finish*; *go on*; *keep up*; *resign*

Table 1: Major semantic classes of verbs following first person singular *I* in the Emo corpus.

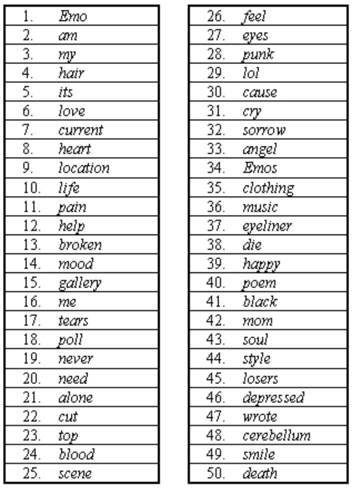
The questions will focus on general conclusions concerning Emo identity, especially in comparison to other youth culture identities. I will additionally ask students about details revealed by the data not easy to explain, e.g. the proliferation of rather violent terms of physical contact and the mix of positive and negative emotions in the respective category (which stands in opposition to the stereotypes of Emos as concerned with dark moods only).

In addition, students are also supposed to indicate – as in the ensuing tasks, too – whether there was any item that surprised them and which they would pursue further, e.g. by looking at concrete examples.

Keyword comparison between two teenage corpora

The second dataset presents a keyword comparison between the Emo corpus and COLT, a 500,000-word component of the BNC, containing spoken language of teenagers (from 1993) (cf. http://torvald.aksis.uib.no/colt). Students are provided with the top 50 words that occur significantly more often in the Emo corpus.

Even though the comparison between the two corpora is problematic – especially the spoken vs. written difference is responsible for some differences – it still can highlight certain aspects peculiar to the Emo perspective of the world.



1.	*Emo*	26.	*feel*
2.	*am*	27.	*eyes*
3.	*my*	28.	*punk*
4.	*hair*	29.	*lol*
5.	*its*	30.	*cause*
6.	*love*	31.	*cry*
7.	*current*	32.	*sorrow*
8.	*heart*	33.	*angel*
9.	*location*	34.	*Emos*
10.	*life*	35.	*clothing*
11.	*pain*	36.	*music*
12.	*help*	37.	*eyeliner*
13.	*broken*	38.	*die*
14.	*mood*	39.	*happy*
15.	*gallery*	40.	*poem*
16.	*me*	41.	*black*
17.	*tears*	42.	*mom*
18.	*poll*	43.	*soul*
19.	*never*	44.	*style*
20.	*need*	45.	*losers*
21.	*alone*	46.	*depressed*
22.	*cut*	47.	*wrote*
23.	*top*	48.	*cerebellum*
24.	*blood*	49.	*smile*
25.	*scene*	50.	*death*

Table 2: The 50 most significant keywords of the Emo corpus as compared to COLT.

The questions asked will resemble those for the first task, with a slightly stronger emphasis on linguistic peculiarities, e.g. the prominence of certain words likely to occur in fixed expressions (e.g. *broken* as in *broken hearted*) or common nouns probably used as nicks in forums (e.g. *cerebellum*).

List of content words

The third task involves the presentation of the top 50 content words in the Emo corpus (partly lemmatized as indicated below), produced with Wordsmith's Wordlister function and the help of a stoplist.

hair	530
love	472
life/live	416
think	400
make/makes/made	382
need	355
say/said	339
know	338
good	331
help	294
go/going	280
joined	275
heart	268
people	263
friend/s	261
dead/death/die	256
want	247
see	242
really	239
top	235
current	234
feel	234
pain	222
black	220
look	218
music	212
location	211
tell	211
thing	204
time	204
leave/left	199
cut	192
way	192
find/found	181
cry/crying	179
hate	177
mood	170
alone	162
day	155
world/s	154
eyes	153
let	147
right	135
gallery	134
guy	134
try	134
take	133
new	131
blood	130
girl	130

Table 3: Top 50 content words in the Emo corpus.

The questions asked will practically be the same as in task 2, with a slightly stronger focus on the difference between thematically motivated lexemes and those whose inclusion could be explained with reference to the nature of the corpus (e.g. structure of blog or forum entries).

Concordances of alone, lonely, on my own

The last part of the questionnaire just contains an extract of a concordance – not sorted in any particular content-related way – of three expressions of loneliness, which is supposed to be the dominant emotional condition of Emos.

etimes I do feel rlly depressed cuz im so **lonely** and I just dont feel many ppl understand
and felt like trying it. I cannot survive **alone**, it feels like I lost everything I've kno
the broken hearts, for the people who are **lonely** and lost, the ones who run out of hope -
ybe living isn't for everyone." Ever felt **alone**, unwanted, deserted, unloved, close to de
The world is silent, should forever feel **alone** - NO! Cause you are gone and I will never
sh which will be my last I cannot survive **alone**, it feels like I lost everything I've kno
2007 11:41 pm Post subject: ur so not **alone** . i get called emo ...mostly cause i am b
7:35 pm Post subject: Well,i'M really **aloNe** in faCt i don haVe any close friend thou
t friend that pays attention to me too im **alone** ....alot well im here if u wanna talk
ted suicide and now i'm left cursed to be **alone** if there is a god he truly hates me but i
have about a zillion piercings in my face **alone** rite now...lol?•??my <3 whispers for the
d listenin to her...i started doin things **on my own**....the only thing she didn't agree to
i dont feel that im ready to take that on **on my own** rite now... just respect them as best
... so im just waiting to do that till im **on my own**... and i dont feel that im ready to t
l. I hate it , i wish they would leave me **alone**. Im at high school now , ive lost to much
Trust me, it comes in handy on those cold **lonely** nights. Back to top Deathonabun User's
ted suicide and now i'm left cursed to be **alone** if there is a god he truly hates me but i
mom more gentle wid u or is ur dad ..talk **alone** id one of them n if it goes well take his
ackdoor? 4. do you leave your boyfriends **alone** when they are trying to ignore you? if yo
I look back up and realize I'm once again **alone**. The lamp post slowly flickers on and off
er801.html About Me: I miss walking under **lonely** lamp posts, looking down at me as if I'v
wreck i cant stop shaking i sit at lunch **alone** listening to my ipod they ignore me and t
ck to say you're sorry Why'd you leave me **alone**~? And now my mama's depressing smile hurt
, From pre-k to present, It's all sad and **lonely**, And I'm still a peasant. My life has su
art, dont attack loving you makes me hack **alone** now, and black A lonley sleep (plz commen
ow? you left me here to feel like a queer **alone** i hate being together isnt my fate i feel
remember Make me cry and wish I wasn't so **alone** Make me wish it wasn't my fault I confuse
my smile and hope its good enough.... <3 **ALONE** Empty room and empty heart But why does m
resort ive tried hiding smiling and being **alone** it just leaves time for thoughts to creep
own Live with no regrets, you may end up **alone**… what am i?? 12-17-2007 1:47 pm What am I
own Take a chance in life, you may end up **alone** I can still smell his Axe, feel his arm '
where in the dark, my thoughts are flying **alone**. Now is my turn, to realise what is next.
old and unloved Hated and pained Hurt and **alone** Nothing is gained Helpless and sad Anxiou
e And without an answer he thought he was **alone** He walked the many stairs Up to the highe
the dice I waited and sat in the darkness **alone** It happened you left now I'm chilled to t
3:54 pm i'm sat here waiting, in my room **alone**. i'm gripping and clenching, onto my mobi
k alone so0o0 alone 12-30-2007 3:09 pm So **alone** in my bed Alone listening to nightly whi
rs are show up Here I am Sitting here all **alone** Waiting for something I lay down here Wai
l me what you guys think!!! Current Mood: **lonely** (2 broken hearted losers | you'll never
thought of your goodbye. How you leave me **alone**. Crying till the mildew comes. Dreaming o
love of my life and my best friend I sit **alone** in the dark, I look at the view of the pa
knew now you would be fine. Now I feel so **alone**, I cant face the people at home, I dont w
h for them I'm just a mistake I wish I we **alone** with no one here just me, myself, and I b
te my life theres no doubt about that, my **lonely** shame; for that I'll always have I'm nev
3 pm - my emo poem HIDDEN FEELINGS **On my own** feeling lonely, Tension building up i
never need) x_vampire_kid_x 10:30 pm I'm **lonely** and depressed I'm broken and I'm bare Bl
r need) x_vampire_kid_x 10:18 pm Standing **alone** on a beach Cold wind bites my skin A bla
took your life away, You thought you were **alone** But bady i was by your side NOW IM SLEEP
opes and dreams you came to me so sad and **lonely** you told what he did to you I began to h
it was great but now we dont date for im **alone** and she is gone but the memories they shi
It doesn't make me happy It makes me feel **alone** I'm not good enough for anyone And I shou
ld heart still searching for something My **lonely** soul is lost in the darkness Try to find
I can find my true love? Sitting here all **alone** Watching the stars Hoping for the sign Th
cribed in her scars She's now broken Left **alone** in the dark Never to be found As she sits

Concordance 1: Search words/phrase alone, lonely, on my own.

The questions will deal with the conceptions of loneliness revealed by the language found in these examples, also touching upon peculiar linguistic features, e.g. deviant orthography.

Preview

As the questionnaire is to be distributed in early June, no results have been gained yet. I can therefore only talk about my expectations here.

My basic assumption is that even within the short time span of a questionnaire, the participants in the study will go from reading off the obvious from the data provided to coming up with quite sophisticated interpretations on their own, learning in the process how language contributes to a cultural phenomenon such as the identity construction of an alternative youth movement.

This growing awareness should motivate them to explore the topic of Emos further, whether by looking for information more concretely (probably with the help of Internet sources) or by their own exploratory corpus work. With a short introduction to the technical side of corpus analysis, this first phase of corpus-based Cultural Studies teaching should easily lead to a stage where students can work with the corpus alone, probably with a higher chance of success than if you leave them to their own corpus analytical devices right from the beginning, as I did in another study (Kettemann/Marko forthcoming).

References

Eppler, Eva/Robert Crawshaw/Caroline Clapham (2000). “The Interculture Project corpus: data classification, access and the development of intercultural competence.” In: Lou Burnard/Tony McEnery (eds.). *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt a.M. et al.: Lang. 155-164.

Kelley, Trevor/Simon Leslie (2007). *Everybody Hurts – An Essential Guide to Emo Culture*. New York: Collins Harper Publishers.

Kettemann, Bernhard/Georg Marko (forthcoming). “Data-driving Critical Discourse Analysis. Learning about language and ideology by autonomously exploring a corpus (of creationist literature).”

Minugh, David (2007). “George Bush and the Last Crusade or the fight for truth, justice and the American way.” In: Encarnación Hidalgo/Luis Quereda/Juan Santana (eds.). *Corpora in the Foreight Language Classroom.* Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6). University of Granada, Spain, 4-7 July, 2004. Amsterdam & New York: Rodopi. 191-205.

Scott, Mike (2006). *WordSmith Tools 4.0*. Oxford: Oxford University Press.

The Bergen Corpus of London Teenage Language (n.d.). University of Bergen, Norway. [Online]. http://torvald.aksis.uib.no/colt/ (accessed May 28, 2008).

USER-FRIENDLY CORPUS TOOLS FOR LANGUAGE TEACHING AND LEARNING

Iztok Kosem[53]

Abstract

The use of corpora in language learning and teaching is still not as widespread as some think it should be. The explanations for this often mention teachers and their lack of motivation, or lack of computer/corpus skills, or both. Rarely is the blame attributed to corpus tools, the first point of contact between teacher/learner and corpus data. This important role of corpus tools could mean that teachers may sometimes not be put off by corpus methodology, but by the medium used.

This paper evaluates some of the existing corpus tools and interfaces in terms of their user-friendliness to language teachers/learners. The categories tested are ease of use, ease of navigation through menus, speed of output production, help/tips provided, and customizability of font and font size.

The paper concludes that there is considerable room for improvement in corpus tools/interfaces for the purposes of language teaching/learning. Most existing tools were initially designed for researchers and not teachers, so it would be much better to create pedagogically-oriented corpus tools from scratch, and involve teachers in the creation of these tools from the very beginning. In addition, these new tools could be part of a new online corpus resource, resulting in a complete education corpus package.

Keywords: user-friendly, corpus tools, language teaching, language learning

Introduction

The use of corpora in language learning and teaching has grown since the pioneering work of Johns (1986, 1988), however it has been largely limited to researchers and a minority of teachers keen on exploring new approaches to language teaching. Teachers' lack of enthusiasm for the use of corpora in the classroom has often been attributed to the time-consuming nature of training in corpus use, and of preparing corpus-based exercises (Swales, 2000; Chambers, 2005). According to the preliminary results of a recent survey on the use of corpora in language education, conducted by Chris Tribble (personal communication), lack of computer skills is among the most often mentioned reasons why corpora are not used in learning and teaching.

One aspect of the use of corpora in language education that has so far received little attention has been corpus tools. Corpus tools are merely a medium – the user (in this case, teacher or learner) should focus on corpus data displayed by the tools, and not on the tools themselves. And some of the issues mentioned in the previous paragraph indicate that this is not the case.

It seems that instead of promoting corpus methodology to language teachers, corpus linguists should perhaps focus more on improving the user-friendliness of corpus tools. The next section evaluates some of the corpus tools currently available, considering the needs of language teachers and learners.

Existing corpus tools – suitable for language education or not?

Six corpus tools were evaluated, three of them computer programs (WordSmith Tools, AntConc, and MultiConcord), and three web-based corpus interfaces (Sketch Engine, MICASE and PhraseBox).

Requirements and cost

[53] Iztok Kosem is a Ph.D. student in Corpus Linguistics at Aston University. His Ph.D. topic is "Designing a model for a corpus-driven EAP dictionary". He is also Assistant Director of the Aston Corpus Network (ACORN) project. He has worked on a recently published English–Slovene Dictionary; and has an MA in Language and Lexicography (Birmingham University). He recently co-authored an article with Ramesh Krishnamurthy in the Journal of English for Academic Purposes, titled "Issues in creating a corpus for EAP pedagogy and research".

Teachers interested in using corpus tools in the classroom or to produce materials will often consider the budgetary constraints of their department. If a corpus-tool licence is expensive, and if a major investment in computer equipment is needed, the plan to use corpora for education purposes may be abandoned before the methodology is even properly tested.

The two tables below show the requirements and costs of computer-based programs and web-based tools respectively.

	WordSmith Tools	**AntConc**	**MultiConcord**
Supported platforms	Windows Apple Mac (*Crossover* software required; cost: £39.99)	Windows Apple Mac Linux	Windows
Cost	Single user: £50 Organization (up to 10 users): £250 Network (up to 50 users): £500 Network (up to 200 users): £1000	Free	Single user: £50 Organization (up to 10 users): £350 Organization (up to 25 users): £500
Demo version available	YES	N/A	YES

Table 1: Computer-based corpus tools

	Sketch Engine	**MICASE**	**PhraseBox**
Works on:			
-Internet Explorer	YES	YES	YES
-Firefox	YES	YES	YES
-Apple Mac (Safari)	YES	YES	YES
Cost	Single user: €55.25 + VAT Site licence: €1.080 + VAT (Licences are for non-profit use only, and are valid for one year) 30-day free trial available	Free	Free (limited functionality)
Additional requirements			Adobe/Macromedia Flash 9 (free software)

Table 2: Web-based corpus tools

Out of six corpus tools evaluated in this paper, three are freely available. The free online version of PhraseBox, a tool designed by John Sinclair for schools in Scotland, is the one that is used by pupils, while full functionality is only available to teachers who were given a password. The three corpus tools that are not free have broadly similar licence fees, but they can be tested for free.

Features

Corpus tools come with many features but only some of them are likely to be used by language teachers and learners. The list of features in the table below was designed according to the practice and comments of teachers and students at Aston University, and is believed to be a reasonable representation of the features useful for language education.

	WordSmith Tools	**AntConc**	**MultiConcord**	**Sketch Engine**	**MICASE**	**Phrasebox**
ready-made corpora provided	NO	NO	YES	YES	YES	YES
can load your own corpora	YES	YES	YES	YES	NO	NO
single-word search	YES	YES	YES	YES	YES	YES
multi-word search	YES	YES	YES	YES	YES	YES
concordances	YES	YES	YES	YES	YES	YES
word frequency list	YES	YES	NO	YES	NO	NO
n-grams (or clusters) frequency list	YES	YES	NO	NO	NO	NO
collocates	YES	YES	NO	YES	NO	YES
keywords	YES	YES	NO	NO	NO	NO
automatic creation of exercises	NO	NO	YES	NO	NO	NO

Table 3: A selection of features in six corpus tools

User-friendliness

To get an idea of the user-friendliness of the six corpus tools, the features mentioned in the previous section, where available, were put to the test, using the following steps:

a) loading/selecting a corpus
b) creating concordances: single-word search: *could*
- sorting concordances: first word to the right, then first word to the left
- creating a 4-gram frequency list
- creating list of collocates

c) creating a word frequency list
d) creating concordances: multi-word search: *could be*
e) automatically creating exercises

Steps b-d included producing a Word version of the output, aligned by the search word. The qualities evaluated during testing were ease of use, ease of navigation through menus, speed of output production, help/tips provided, and customizability of font and font size. The categories were rated on a 4-point ranking, using a slightly adapted version of Tribble's (2003) scale:

● = unacceptably difficult

●● = difficult

●●● = easy

●●●● = encouragingly easy

Speed of output production used a different form of evaluation:

○ = notoriously slow (enough time to open a web browser window, do a search in Google and open one of the links)

○○ = slow (enough time to open a web browser window and do a search)

○○○ = fast (slight delay, barely enough time to open a web browser window)

○○○○ = really fast (no delay).

It was thought that measuring the speed with actions rather than time was a better reflection of real-life teacher/learner practice.

A 3.8-million-word corpus of literary texts from the Gutenberg Project website was used in testing WordSmith Tools and AntConc, the 100-million-word British National Corpus was selected in Sketch Engine and PhraseBox, and the corpora were already predetermined in MultiConcord (small English-French parallel corpora offered by the demo version; size not provided) and MICASE (1.8-million-word MICASE corpus).

Each corpus tool was tested separately, and the computer was restarted between individual tests. The computer specifications were: 1.6GHz Intel Pentium M processor, 512MB RAM, 40GB, running under Windows XP Professional and using 2MB broadband internet connection. The web-based corpus tools were tested using only Firefox, version 2.0.0.14.

Ease of use	●
Ease of navigation through menus	●
Speed of output production	○○
Saving outputs into a Word document	●
Help/tips provided	YES
Customizability of font and font size	YES

Table 4: WordSmith Tools

WordSmith Tools is a program for corpus enthusiasts, i.e. researchers and academics, rather than language teachers and learners. It is likely to take a long time, and continuous use, for any user to be able to use the program quickly and efficiently. There are simply too many decisions to be made at each step; worse still, default settings (at least in this version of the program) seem to force the user into more decisions. For example, after using the “Choose texts” option in the “File” menu of the main window and selecting the texts, one would think that the corpus is now loaded. However, this is only true for concordance search; when making a word list, the texts need to be selected again.

Concordances and word frequency list took a long time to produce, while sorted results, collocates and 4-grams were produced instantly. The speed of producing concordances is affected by the fact that collocates, plot, patterns, and 3-word clusters (default setting) are calculated at the same time; which can be useful but also unnecessary if a teacher does not need those functions, or needs settings other than the default ones.

Saving results into a Word document proved to be very problematic. Even if you manage to get through the difficulty of saving a file in TXT format and opening it in Microsoft Word, you will find that the results are not perfectly aligned by the searched word. On the positive side, WordSmith Tools has a very useful “Print” function that prints the data as seen on the screen.

Ease of use	●●●
Ease of navigation through menus	●●●
Speed of output production	○
Saving outputs into a Word document	●●
Help/tips provided	NO
Customizability of font and font size	YES

Table 5: AntConc

AntConc is a user-friendly program with a simple one-window interface with tabs for different functions. It is easy to use and offers functionality similar to WordSmith Tools. A big advantage of AntConc is that it does not require installation, so you can simply download it from the internet and start using it. Menus are quite easy to navigate, especially as functions for manipulating text (e.g. sorting) are presented under the results and not in a separate window.

Speed results were very disappointing. It took a long time to produce concordances and a word list, whereas 4-grams and collocates were not even produced as the program was shut down by the computer after 3-5 minutes, seemingly due to using too much internal memory. The performance of AntConc may improve on better computers, however there is a possibility that large data sets and searches of frequent words may still cause problems.

Saving results into a Word document is easy and it produces concordances aligned by the search word. There is one big problem, though: only all the results can be saved (in this case, over 7000), which means the Word document takes a long time to open and any manipulation of text is time-consuming. The user can, of course, select the text in AntConc window and copy-and-paste it, but right-clicking the mouse button does not work – the user needs to be familiar with CTRL-C keyboard command (for copying text).

Ease of use	●●
Ease of navigation through menus	●●●
Speed of output production (in demo version)	○○○○*
Saving outputs into a Word document	●●
Help/tips provided	YES
Customizability of font and font size	NO

Table 6: MultiConcord

MultiConcord is a robust parallel concordancer, lacking many analysis functions such as collocations and N-grams. The interface, written in the mid-1990s, is rather crude for the modern user. Menus can be confusing and the small fonts are not helpful either; and settings cannot be customized. Results were produced very quickly, although it should be pointed out that the test was done on a demo version which uses a small parallel dataset.

The main feature of the program is its Testing function. This allows teachers to produce exercises without needing to extensively manipulate the data. The user has the option to create exercises with source language texts, target language texts, or parallel texts. The scope of exercises is limited as only different types of gap-filling exercises are available.

Saving outputs into a Word document can only be done in the “Testing” option. Easy-to-follow instructions are provided on how to retain parallel alignment of concordances. The program does not allow copying-and-pasting of text on the screen.

Ease of use	●●●
Ease of navigation through menus	●●●
Speed of output production	○○○
Saving outputs into a Word document	●●
Help/tips provided	NO
Customizability of font and font size	NO

Table 7: Sketch Engine

The selection of ready-made corpora in Sketch Engine is impressive; both in terms of corpus size and languages. The interface is quite easy to use, although there are sometimes still too many parameters to choose. Sorting of concordances of more frequent words can be problematic as the program does not ignore characters such as punctuation; for example, when sorting by first word to the right of *could*, the words beginning with the letter "A" were found around page 200 of the concordances.

Speed is one of the strong points of the Sketch Engine. Any searches, with the exception of word lists, are produced almost instantly.

The Sketch Engine comes with a "Save" function that allows any output to be saved. There is a problem with saving concordances in a Word document, as they are not aligned according to the search word. Interestingly enough, copying and pasting the text produces better results (Note: when pasting the text into Word, select "Paste Special" and then the "Unformatted text" option).

Ease of use	●●●
Ease of navigation through menus	●●●
Speed of output production	○○
Saving outputs into a Word document	●
Help/tips provided	NO
Customizability of font and font size	NO

Table 8: MICASE

The MICASE interface is easy to use, but has a very limited set of functions; only concordance searches and sorting concordances. And even concordance output is not without problems: text on the right or on the left, or on both sides of the search word can be presented in two lines which makes concordances more difficult to read, and to analyze.

Speed was not overly impressive, considering that this is only a 1.8-million corpus. Saving the output into a Word document proved problematic as there is no option to save the concordances in .txt format. Copying and saving the concordances on the screen worked, but with some glitches (not all the concordances copied were aligned by the searched word). Finally, the link to help files did not work.

Ease of use	●●●●
Ease of navigation through menus	●●●●
Speed of output production	○○○
Saving outputs into a Word document	●
Help/tips provided	NO
Customizability of font and font size	NO

Table 9: PhraseBox

As a corpus tool developed for (L1) language teachers and learners, PhraseBox comes with many features that demonstrate its pedagogic focus. The interface looks rather modern, which should help attract pupils' interest. As the name suggests, PhraseBox focuses on phraseology, so by default, collocates are displayed first, but the user can switch to concordance view. Functionality is limited, but that is understandable, considering that the tool is meant for L1 language classrooms. Advanced access is available to teachers who are given a password – another very useful pedagogic feature of the program.

Output is produced quickly, and results can be saved to the program's own clipboard. The data from the clipboard can then be saved in a Word document, but the results are not very useful; the concordances are displayed with HTML codes.

There is no help provided; a guide is available as a separate PDF document. PhraseBox offers no customizability; in fact, it seems to override the settings of the web browser – even font cannot be increased or decreased. This is probably a part of pedagogic orientation of PhraseBox – the support in showing pupils how to use the tool is left to the teacher, rather than being provided by the tool itself.

The six corpus tools clearly have room for improvement, especially in terms of performance speeds and saving results into a Word document. PhraseBox seems to be the best one for teachers and learners, but that is to be expected as it was designed for educational purposes. Other tools, however, have several useful features not found in PhraseBox, for example a wide selection of ready-made corpora (Sketch Engine) and the automatic creation of exercises (MultiConcord). Nonetheless, many of the tools try to cater for both researchers and educators. The result is a complex tool that many teachers, and (even more so) their students will find difficult to use.

Creating a user-friendly corpus tool for language teaching and learning

The previous section exposed the shortcomings of some of the existing corpus tools. One solution would be to improve their functionality and user-friendliness, but it may prove difficult to strip the tool of its research-oriented character. Therefore, it is argued that a corpus resource created solely for the purposes of language teaching and learning is a much better alternative. This section looks at some of the features such a corpus tool should have. Proposed solutions are based on the experience obtained from the ACORN (Aston Corpus Network) project at Aston University, which offers corpora in four different languages to be used by teachers and students at the university, with the primary focus being on pedagogical applications. Although ACORN focuses on language teachers/learners, it is also used by teachers and students of other subjects.

Input of teachers and learners

It is vital that a corpus tool for language teaching and learning is developed according to the requirements and feedback of teachers and learners. The ACORN project is doing this in several ways: corpora are compiled according to the needs of teachers, feedback on the interface is regularly sought (e.g. there is a "Feedback" link available in ACORN for users to leave feedback on any aspect of the interface), and reports on any classroom use of ACORN are closely scrutinised.

Web-based tool

It is more efficient to have the corpus tool available online. This normally requires that the user has a fairly quick internet connection, but there is no longer any dependency on each user having access to a computer "with plenty of memory and hard disk space" (Scott, 2008). Web-access allows regular updates of the interface and adding new corpora. At Aston, we also found it useful to monitor usage; for example, the words or phrases being looked up more frequently can be used in tutorials or ready-made exercises.

Ready-made corpora

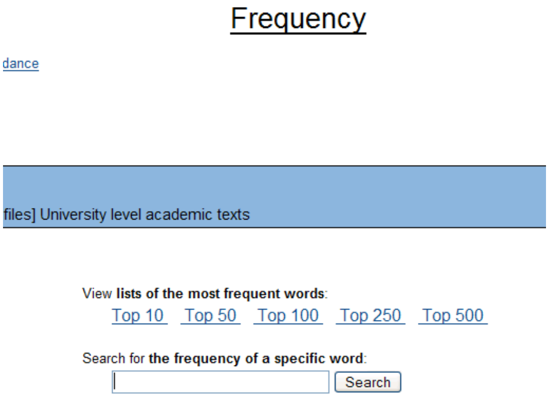
Teachers should not be expected to compile the corpora themselves, especially when first attempting to use corpus methodology. They should have a selection of ready-made corpora available, but the corpora should reflect their requirements. The majority of ACORN corpora were compiled by consulting teachers at the School of Languages and Social Sciences on the type of data they use for their courses.

User-friendly and fast interface

There are two features that significantly contribute to the user-friendliness of the interface: simplicity and speed. Functions on each page should be self-explanatory, but helpful tips should be provided just in case. Quick production of outputs is also very important, so a corpus tool and corpora need to be hosted on a fast server, preferably dedicated only to corpus activity.

Functions

Functionality very much depends on the needs of teachers and learners. At Aston, word frequency lists, n-gram frequency lists (for 2-, 3-, 4- and 5-grams), concordances, and collocations were identified as the functions most useful for the teachers. Word frequency lists and n-gram frequency lists have been calculated in advance, and divided into the following options: Top 10, Top 50, Top 100, Top 250, and Top 500 (see figure below). Alternatively, the user can search for the frequency of a specific word in a frequency list, or compare the frequency of a specific word in several corpora.



Frequency

dance

files] University level academic texts

View **lists of the most frequent words**:

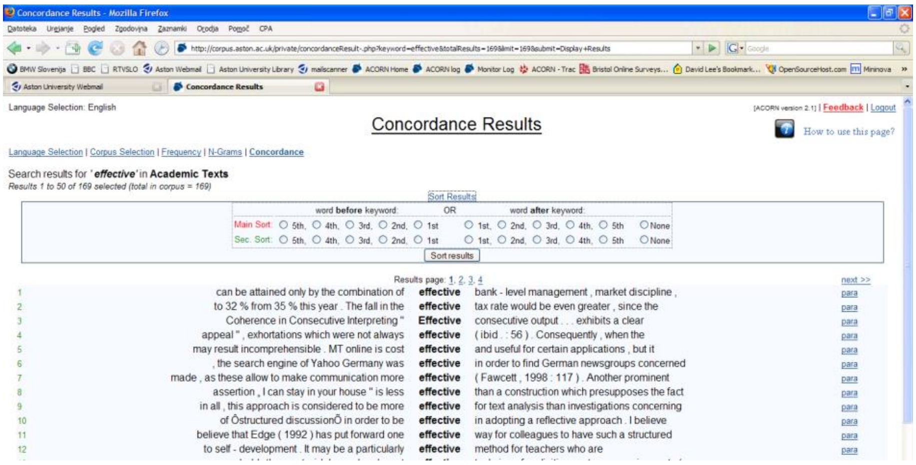
Top 10 Top 50 Top 100 Top 250 Top 500

Search for **the frequency of a specific word**:

Figure 1: Partial screenshot of the frequency feature in ACORN

The Concordance function in ACORN allows the user to search for words or phrases. Before displaying the concordances, the user selects the number of concordances to be displayed. Concordances can be sorted and the sorting options are presented above the concordances (see figure below), in a similar way to the MICASE interface. As with many other features in ACORN, the aim is to avoid the user being required to open menus in separate windows in order to manipulate the data.

There are many translation students at Aston, therefore the Parallel Texts function was also required. At the moment, parallel data for language pairings of English, French, German, or Spanish are available.


Concordance Results - Mozilla Firefox
Datoteka Urejanje Pogled Zgodovina Zaznamki Orodja Pomoč CPA
http://corpus.aston.ac.uk/private/concordanceResult-.php?keyword=effective&totalResults=169&limit=169&submit=Display+Results
BMW Slovenija BBC RTVSLO Aston Webmail Aston University Library mailscanner ACORN Home ACORN log Monitor Log ACORN - Trac Bristol Online Surveys... David Lee's Bookmark... OpenSourceHost.com Mininova
Aston University Webmail
Concordance Results

Language Selection: English

[ACORN version 2.1] | Feedback | Logout

Concordance Results

How to use this page?

Language Selection | Corpus Selection | Frequency | N-Grams | Concordance

Search results for '***effective***' in **Academic Texts**
Results 1 to 50 of 169 selected (total in corpus = 169)

Sort Results

	word **before** keyword:					OR	word **after** keyword:					
Main Sort:	○ 5th,	○ 4th,	○ 3rd,	○ 2nd,	○ 1st		○ 1st,	○ 2nd,	○ 3rd,	○ 4th,	○ 5th	○ None
Sec. Sort:	○ 5th,	○ 4th,	○ 3rd,	○ 2nd,	○ 1st		○ 1st,	○ 2nd,	○ 3rd,	○ 4th,	○ 5th	○ None

Sort results

Results page: 1, 2, 3, 4 next >>

1	can be attained only by the combination of	**effective**	bank - level management , market discipline ,	para
2	to 32 % from 35 % this year . The fall in the	**effective**	tax rate would be even greater , since the	para
3	Coherence in Consecutive Interpreting "	**Effective**	consecutive output . . . exhibits a clear	para
4	appeal " , exhortations which were not always	**effective**	(ibid . : 56) . Consequently , when the	para
5	may result incomprehensible . MT online is cost	**effective**	and useful for certain applications , but it	para
6	, the search engine of Yahoo Germany was	**effective**	in order to find German newsgroups concerned	para
7	made , as these allow to make communication more	**effective**	(Fawcett , 1998 : 117) . Another prominent	para
8	assertion , I can stay in your house " is less	**effective**	than a construction which presupposes the fact	para
9	in all , this approach is considered to be more	**effective**	for text analysis than investigations concerning	para
10	of Ŏstructured discussionŎ in order to be	**effective**	in adopting a reflective approach . I believe	para
11	believe that Edge (1992) has put forward one	**effective**	way for colleagues to have such a structured	para
12	to self - development . It may be a particularly	**effective**	method for teachers who are	para

Figure 2: Screenshot of a sample of concordances for *effective* with sorting feature activated

Simple export of outputs into Microsoft Word

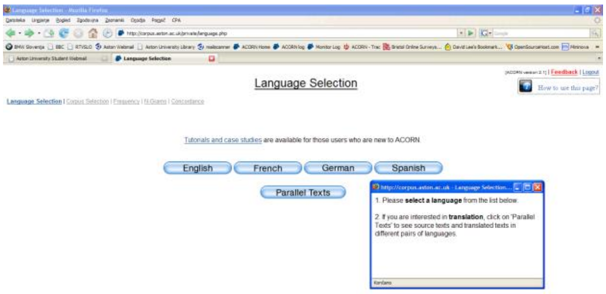
Most teachers prepare their classroom material in Microsoft Word, so it is important that the corpus outputs can be easily exported into a Word document. And this seems to be one of the issues yet to be properly addressed by developers of corpus tools. One important fact any teacher using corpora should be made aware of is that the concordances need to be displayed in a fixed-character-width font (such as Courier) in order to be aligned by the search word. Some tweaks in Microsoft Word are still required, such as setting the layout to Landscape, and reducing the font size so that each concordance line fits in one line.

Automatic production of exercises and tests

The automatic production of corpus-based material can be regarded as the ultimate teacher-friendly function of a corpus tool. The "Testing" function in MultiConcord is an example of good practice, and ACORN plans to offer something similar in the future.

Help and support

It is important that a corpus tool offers constant help on functions being used. In ACORN, links to helpful tips are available in the top right corner of the screen, and open in a new window (see screenshot below). The feedback form can be used to report any problems or errors encountered while using the interface.


Language Selection - Mozilla Firefox
Datoteka Urejanje Pogled Zgodovina Zaznamki Orodja Pomoč CPA
http://corpus.aston.ac.uk/private/language.php
BMW Slovenija BBC RTVSLO Aston Webmail Aston University Library mailscanner ACORN Home ACORN log Monitor Log ACORN - Trac Bristol Online Surveys... David Lee's Bookmark... OpenSourceHost.com Mininova
Aston University Student Webmail
Language Selection
English
French
German
Spanish
Parallel Texts
http://corpus.aston.ac.uk - Language Selection...
Končano

[ACORN version 2.1] | Feedback | Logout

Language Selection

How to use this page?

Language Selection | Corpus Selection | Frequency | N-Grams | Concordance

Tutorials and case studies are available for those users who are new to ACORN.

1. Please **select a language** from the list below.

2. If you are interested in **translation**, click on 'Parallel Texts' to see source texts and translated texts in different pairs of languages.

Figure 3: Screenshot of the main ACORN window (help tips open in a separate window)

Training needs to be provided for teachers, but not only when they start using corpora. Refresher training should be offered at the beginning of each academic/school year (especially if there have been any changes in data, functions, or display formats), or more often if there is a demand for it. Experience at Aston also points to the need for a corpus-proficient person who can liaise with teachers and even help them in administering lessons initially.

Conclusion

There is obviously considerable room for improvement in corpus tools/interfaces for the purposes of language teaching/learning. We must bear in mind, though, that most existing tools were initially designed for researchers and not teachers. And while researchers often compile their own corpora and have time to train themselves in using corpus tools, teachers need ready-made corpora and do not want to spend a lot of time learning how to use corpus tools. This becomes even more problematic when language learners access corpus data themselves; then, the teacher needs to know how to use a corpus tool well enough in order to be able to train learners in its use.

It is therefore much better to create completely new, pedagogically-oriented corpus tools, rather than trying to tweak the existing ones to the needs of language teachers. This is likely to make teachers more open to the use of corpora, as they are more involved in the creation of these tools from the very beginning. Furthermore, corpus developers can spend less time guessing what the needs of teachers are, and are able to focus more on developing suitable corpus resources.

References

ACORN (Aston Corpus Network). http://acorn.aston.ac.uk. [Access date 17/05/2008]

Cánan Ltd. 2008. *PhraseBox, version 2.* http://www.phrasebox.com/client.html [Access date 08/05/2008]

Chambers, A. 2005. "Integrating corpus consultation in language studies." *Language Learning & Technology*, 9/2: 111-125.

Johns, T. 1986**. "**Micro-concord: A language learner's research tool." *System,* 14/2: 151-162.

Johns, T. F. 1988. "Whence and whither classroom concordancing?" In *Computer applications in language learning,* T. Bongaerts, P. de Haan, S. Lobbe, & H. Wekker (eds.). Dordrecht, The Netherlands: Foris, 9–33.

Swales, J. M. 2000. "Integrated and fragmented worlds: EAP materials and corpus linguistics." In *Academic discourse,* J. Flowerdew (ed.). Harlow: Pearson, 150–164.

Anthony, L. 2008. *AntConc, version 3.2.1w.* http://www.antlab.sci.waseda.ac.jp/software/antconc3.2.1w.exe. [Access date 03/02/2008]

Lexical Computing Ltd. 2008. *Sketch Engine.* http://www.sketchengine.co.uk/ [Access date 14/05/2008]

MICASE (The Michigan corpus of academic spoken English). http://quod.lib.umich.edu/m/micase/ [Access date 03/02/2008]

Scott, M. 2008. *WordSmith Tools, version 5*. http://lexically.co.uk/wordsmith/version5/index.html. [Access date 17/05/2008]

Scott, M. 2004. *WordSmith tools, version 4*. http://www.lexically.net/wordsmith/ Oxford: Oxford University Press.

Tribble, C. 2003. "Five electronic learners' dictionaries." *ELT Journal* 57: 182-197.

Woolls, D. 2008. *MultiConcord, version 1.53.* http://www.copycatchgold.com/ [Access date 10/05/2008]

RHETORICAL TEXT STRUCTURE IN ACQUIRING READING SKILLS IN L3

***Svitlana Kurella*[54]**
***Serge Sharoff*[55]**
***Antony Hartley*[56]**

Abstract

Our ultimate aim is to develop a methodology for English (L1) speakers to acquire reading competence in a third language (L3, here Ukrainian), based on their prior knowledge of a second, cognate language (L2, here Russian). The research is based on the automatic collection of small, dynamic corpora from the Internet and their automatic annotation and classification .The results are to be integrated into the development of supportive learning methods. The focus is on uncovering rhetorical text structure and genre-specific organization patterns in order to promote successful reading strategies. This paper focuses more specifically on the application of text analysis methods from corpus linguistics to select reading materials by topic and genre in order to meet the needs and interests of specific groups of language learners.

Keywords: reading skills, cognate language, rhetorical structure, connectors, dynamic corpus

Teaching reading in a third language

The growing concern for multilingualism over recent years was acknowledged in January 2007 by its recognition as a policy area in its own right and the appointment of a responsible Commissioner. EU policy recommends that EU citizens acquire one language with an international status plus a language of a neighbouring community. This has had an impact on research in third language acquisition (TLA), which is no longer considered as just another instance of second language acquisition (SLA) but as a distinct field developing its own theories and applications (Jessner, 2008). From the psycholinguistic perspective we draw inspiration from two models of TLA. The first (Hufeisen, 2001) underlines the advantages of L3 learners over L2 learners, described in terms of sets of factors which influence the language learning process, such as neurophysiological, socio-cultural, emotional, cognitive and linguistic. This model holds that L3 learners have specific foreign language knowledge and competences (knowledge about the foreign language learning process, individual learning strategies, an individual learner type, etc.) which L2 learners do not have. Therefore, it can be assumed that the L2 can take the role of a supporting language – become a "bridge" to TLA.

The second, multilingual processing model (Meißner, 2004) is more narrowly relevant to our own research goals, as its concentrates on processes which enable development of receptive skills. In particular, it focuses, like us, on reading comprehension in a new foreign language/L3 on the basis of a previously learned etymologically close L2. Again, L2 (in which learners should be proficient) takes the role of a bridge language and serves as a matrix against which new lexical and grammatical structures are compared (Jessner, 2008:24). The process of acquiring reading skills takes students' learned transfer strategies in stages towards systematic multilingual knowledge which can be used in interaction with texts.

We suggest that teaching reading in L3 should pay more attention to the knowledge of text type and its conventions, especially to the similarity and differences between previously learnt languages and the target language. Researchers agree that learning to recognise and navigate different text types is of great importance for the whole process of foreign language acquisition and not only for the reading process as such. Students benefit

[54] Svitlana Kurella is a PhD student in the Centre for Translation Studies (CTS), University of Leeds. Her project is devoted to teaching reading skills in cognate languages (for instance, teaching Ukrainian to students who know Russian). S he is also involved in teaching Russian at Leeds.
[55] Serge Sharoff is a lecturer in CTS. He is involved in several projects related to corpus collection and corpus-based technologies for language learning and translation.
[56] Anthony Hartley is the Director of CTS. His research interests are in Machine Translation, controlled languages and quality of translation and interpreting.

generally from knowledge of a given text type because it provides a basis for development of the knowledge of other text patterns in a foreign language (Feld-Knapp, 2005).

Text selection

We assume that an awareness of text types and the ability to deal with a range of genres in the classroom is a requirement for specialized, vocational learners such as trainee translators. In addition, specific linguistic usage and cultural references can be important for such a group, which usually deals with particular text types. Therefore, we set out to explore the relation between text structure and text types in a specific subject domain. To do this we conducted an experiment in the automatic collection and classification of texts according to genre, with students of translation as our target group.

A further consideration is that students' motivation is enhanced if they work with texts reflecting their own interests. Thus we concentrated on texts published on the internet in the domain of social, political and business affairs. Such internet media texts provide a wide range of topics to suit the interests of any student, and they reflect the most recent trends and changes in language. We deliberately favoured the collection of a small corpus over a large one for two reasons. Firstly, in this domain a corpus becomes outdated almost immediately after it has been collected and so a large but static corpus is less useful than a small but 'dynamic' one that can be regularly updated with recent authentic texts related to the specific interests of language learners. The text selection ranges from short news reports on current affairs to be used at the initial stages of learning to commentaries and analysis for use at an advanced level. A wide variety of short texts ensures successful reading right from the beginning. Secondly, this dynamic corpus is collected from controlled sources, so it is possible to control the quality of the texts from the outset instead of filtering the corpus after collection (Fairon 2006).

We used RSS technology to collect texts in English (L1), Russian (L2) and Ukrainian (L3): the subscribed texts were automatically downloaded, filtered for relevance and classified according to their topic and their rhetorical organisation characteristic of different text types. We based our identification of rhetorical organisation on the types and distribution of connectors in the texts. For instance, reports about official meetings and visits usually contain connectors of *time* ('while', 'during') or *consequence* ('as a result') with a low relative frequency in text. These features place these texts within the text type of *chronicle of current events.*

Text classification

Corpus work in the Data Driven Learning (DDL) and Languages for Special Purposes (LSP) framework involves both research and observation skills: a student should be able to recognise and utilise the typical ways of organising language within the particular genres (Bernardini 2004). Accordingly, contrastive rhetoric has recently become a focus of corpus-based research (Lenko-Szymanska 2007). We identify two principal high-level genres, which have different uses in the classroom as they cover reading for different purposes: reading for information or reflective/critical reading (Alderson, 2000). These genres are constituted by descriptive/narrative text forms on the one hand ('information') and expository and argumentative text forms ('opinion') on the other. The main problem for classifying individual texts in our collection of media texts is that news items are not easy to classify into one or the other category at first glance: they can provide information (newswires, reports) as well as present the author's or an expert's opinion (commentary, column, article). In our experiment we paid attention to sampling a range of different text types and taking into account the writer's dominant point of view – i.e., whether it tends towards 'fact' (which places it in text class 'information') or towards opinion, attitude, mood or wish (which places it in the text class 'opinion').

Connectors

For the task of text selection and text classification for the classroom we assumed that *conjunction* as a type of text cohesion (Halliday & Matthiessen, 2004) is most appropriate marker of text structure and genre. From among units of conjunction we selected connectors[57] as a primary textual cohesive device: functioning to mark semantic

[57] Although Halliday (1976) defines the "units" of conjunction as *conjunctives, conjunctive adjuncts,* or *discourse adjuncts,* and later (2004) even as *conjunctions,* we use the term ***connectors,*** as it is closer to the terminology known by language teachers and learners across languages. Moreover, that the term *conjunction* is easy to associate with a part of speech, which can cause confusion.

relations between parts of the text, they signal the logico-rhetorical text structure. Thus, we started from the hypothesis that it is possible to classify texts using textual connectors.

Connectors have become a focus of interest of Russian linguists over the past decades. Although current research in Russian text linguistics has paid attention to some aspects of the issue, such as analysing the discourse structure of academic texts (Bol'shakova & Baeva, 2004) or determining the category of textual ties (Prijatkina, 2002), an analysis of this type of cohesion is still lacking for East Slavonic language, including Ukrainian.

Lists of connectors were compiled for each of the three languages following the functional classification from (Halliday & Matthiessen, 2004). Since no classification of conjunctive relations was available for Russian and Ukrainian, we developed our lists by adopting the English model of classification according to type of semantic function (Halliday & Matthiessen, 2004), by collecting connectors from academic grammars of Russian and Ukrainian, and later by extracting them from our corpus. Text connectors ('текстовые скрепы') in Russian and Ukrainian can be considered as a special type of function word which can be realised formally by single-word units or phrases, including adverbial, adjective, substantive and predicative word forms, phraseological units, fixed expressions with conjunctions and particles. Their main distinctive feature is syntagmatic isolation from the left and right components of the text structure, marked by punctuation marks in written text. One of the special features which characterise Russian and Ukrainian connectors is their position: they often appear at the beginning of the sentence, whereas such positioning in English texts may be considered as bad style (e.g., 'But…'). Taking this observation into account, we avoid confusion with conjunctions by considering only sentence-initial units for Russian and Ukrainian. These connectors are more likely to function as marker of semantic relations at a text level.

Our aim was to detect which connectors are the most significant and characteristic for marking certain semantico-logical relations, rather than to undertake a full, detailed categorisation of connectors. Currently we identify 14 categories of connectors listed in Table 1 and exemplified in Table 2. These relations formed the basis for defining a set of text organization patterns. The constellation of such relations in a text reveals its logico-rhetorical structure.

ADDITION, ADVERSARIAL, CONDITION, ARGUMENTAL,
TIME REFERENCE, SEQUENCE / CONCLUSION,
RESULT, REASON, PURPOSE, CONCESSION,
COMPARISON, EXEMPLIFICATION / CLARIFICATION,
OPINION

Table 1: Classification of connectors

	REASON Причины
Ukr	тому що, через те що, у звязку з тим що, тим що, з нагоди, з приводу, бо, оскільки
Rus	потому что, так как, ибо, оттого что, ввиду того что, благодаря тому что, вследствие того что, в связи с тем что, в силу того что, затем что
Eng	in account of this, for that reason, in connection with, therefore, for this reason, because of that
	PURPOSE Цели
Ukr	щоб(и), з тим щоб, аби, для того щоб, затим щоб, задля, заради, в імя, в інтересах, з метою, на користь
Rus	чтобы, чтоб, для того чтобы, затем чтобы, с тем чтобы, дабы, в имя, в интересах, с целью, в пользу
Eng	for that purpose, with this in view, in order that, so that
	COMPARISON Сравнение (classification, analogy)
Ukr	як, (не)мов, (не)наче(б/бто), що, ніби(то), (не)мовби, ніби, буцім би/то/би то, подібно до того як, інакше, інак
Rus	как, как бы, будто, будто бы, будто б, как будто, как будто бы, как будто б, словно, словно как, подобно тому как, точно, чем, чем...тем, равно как
Eng	likewise, in the same way, similarly, in a different way, as if, as it were, as though
	OPINION/EVALUATION Авторская оценка (expressing opinion and evaluation; recommendations, desire)
Ukr	на … думку, я думаю, (не) думаю, (не) вважаю, наголошу, як відомо, звісно
Rus	я думаю, я надеюсь, без сомнения, как видно, как надо думать, очевидно ....
Eng	in…opinion, personally, obviously, without doubts

Table 2: Example connectors

Experiment

In order to test our hypothesis that it is possible to classify texts using textual connectors, we compiled small training corpora of news texts, which comprised 48 texts in Ukrainian and 47 texts in Russian. We counted for each text the number of connectors in each of the categories in Table 1, and used these counts as features for predicting whether the text belongs to the 'information' or 'opinion' category. Independently the texts were classified as such by human adjudicators. To estimate the accuracy of the automatic classification, we used the freely available machine-learning tool Weka, in particular the SVM classifier (Witten & Frank, 2005) with 10-fold cross-validation. Table 3 presents the results of this evaluation.

```
Ukrainian:                                      Russian:

  Precision  Recall  F-Measure  Class             Precision  Recall  F-Measure  Class
  0.912      0.969   0.939      information       0.903      1       0.949      information
  0.929      0.813   0.867      opinion           1          0.842   0.914      opinion

=== Confusion Matrix ===                        === Confusion Matrix ===

  a   b   <-- classified as                       a   b   <-- classified as
 31   1 | a = information                        28   0 | a = information
  3  13 | b = opinion                             3  16 | b = opinion
```

Table 3: Classification accuracy

Results

The classification is reliable for texts in both languages, achieving an F-measure of 0.867 and 0.939 for the two classes in Ukrainian and 0.949 and 0.914 in Russian. This is achieved despite the fact that we used a relatively small dataset for training the classifier. Predictably, the most important class of connectors for differentiating between two text classes is the OPINION category of connectors, (e. g. *in my opinion, as well known*) with a weight of 1.4088 in Ukrainian and 1.8233 in Russian, followed by the classes of PURPOSE (1.391) and COMPARISON (1.0125) in Ukrainian and CONDITION (1.1688) and ADVERSARIAL (1.4324) in Russian. The results suggest the need for further investigation of the category OPINION.

Future work

We are developing a CALL environment which allows students to highlight the constellation of connectors revealing the particular logico-rhetorical structure of the text they are reading. This is intended to enable students to visualise rhetorical relations in more complex texts, such as *analysis* and *commentaries* that typically have more than two classes of connectors, including those of causal-effect relation and of modal assessment. Students will also be able to compare rhetorical conventions across the three languages and automatically find topic-related texts with a similar structure.

References

Alderson, J. C. 2000. *Assessing Reading*. Cambridge: Cambridge University Press.

Bernardini, S. 2004. Corpora in the classroom. An overview and some reflections on future developments. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (Vol. 12, pp. 15-36). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bolshakova, E. I., & N. V. Baeva. 2004. *Avtomaticheskij analiz diskursivnoj struktury nauchnogo teksta - Automatic analysis of academic text discourse structure*. Available: http://www.dialog-21.ru/Archive/2004/Bolshakova.htm [2007, 08/01].

Fairon, C. 2006. Corporator: A tool for creating RSS-based specialized corpora. Paper presented at the *2nd International Workshop on Web as Corpus*, EACL-2006, Trento.

Feld-Knapp, I. 2005. *Textsorten und Spracherwerb. Eine Untersuchung zur Relevanztextsortenspezifischer Merkmale für den "Deutsch als Fremdsprache"-Unterricht - Text genres and language acquisition. A study of the relevance of textual genre-specific features for teaching German as a foreign language.* Hamburg: Verlag Dr. Kovac.

Halliday, M. A. K., & C. Matthiessen. 2004. *An introduction to functional grammar* (3 ed.). London: Arnold.

Hufeisen, B. 2001. Deutsch als Tertiärsprache - German as a third language. In G. Helbig & L. Götze & G. Henrici & H.-J. Krumm (Eds.), *Deutsch als Fremdsprache. Ein internationales Handbuch* (Vol. 1, pp. 648-653). Berlin, New York: Walter de Gruyter.

Jessner, U. 2008. Teaching third languages:Findings, trends and challenges. *Language teaching*, *41*(1), 15-56.

Lenko-Szymanska, A. 2007. Different cultures or different skills? Cohesive devices in native and foreign language learners' texts. Paper presented at the *Language learning and teaching in multilingual and multicultural contexts*, Paris.

Meißner, F.-J. 2004. Transfer und Transferieren: Anleitungen zum Interkomprehensionsunterricht. In H. G. Klein & D. Rutke (Eds.), *Neue Forschungen zur Eropäischen Intercomprehension.* (pp. 39-66). Aachen: Shaker.

Priyatkina, A. 2002. Tekstovye skrepy i skrepy-frazy (o rasshyrenii kategorii sluzhebnykh edinic russkogo yazyka) - Textual "ties" and "Tie-phrases" (about extending a category of function words in Russian), *Predlozhenie. Tekst. Rechevoe funkcyonirovanie yazykovykh edinic. Mezhvuzovskij sbornik nauchnykh trudov*. Elets: Bunin Elets State University.

SMALL WORDS, BIG DEAL: TEACHING THE USE OF FUNCTION WORDS AND OTHER KEY ITEMS IN RESEARCH WRITING

David Y.W. Lee[58]

Sylvia Xiao Chen[59]

Abstract

In many mainland Chinese universities, undergraduate students specializing in English language and applied linguistics are required to write a dissertation, in English, of about 7,000 words exploring some aspect of (original) research. This is a task which is of considerable difficulty even for native speakers of English, not only at the genre or discourse level, but also at the lexico-grammatical level. The teaching of academic writing in Chinese universities tends to focus on general discourse-level features such as "move" structures, while the more micro, form-focused knowledge and skills are comparatively underexplored. This is reflected in the literature, with previous studies focused on the discourse functions of linguistic items such as connectors and hedges (e.g., Feng & Zhou, 2007; Liu, 2005; Mo, 2005; Pan, 2007; Wang, 2007; Zhang, 2006). The choice of items in these studies was usually based on intuition or an arbitrary selection of what was felt to be important.

In this paper, we present a data-driven, pedagogically oriented analysis of a corpus of 78 Chinese undergraduate dissertations, focusing on characteristically problematic areas, as revealed through keywords analyses (Scott, 2000, 2001; Tribble, 2000) and complementary qualitative investigations of collocations and word clusters. Most of the underuse and overuse of words and phrases turn out to involve function words and high-frequency "common" words which are typically not the focus of academic writing instruction. These usages are highly patterned rather than random, thus being amenable to remedial teaching using a data-driven pedagogical approach. As illustration, we present a set of ready-to-use classroom pedagogical materials that show how the corpus-based approach can scaffold learners in learning the lexico-grammar of academic writing, and thus constitute a system of writing apprenticeship that can be characterised as partly self-learning and partly exemplar- and instructor-led learning..

Keywords: EAP, writing, keywords, collocations, pedagogy

Introduction

Findings from research on (native-speaker) English language corpora have resulted in many changes in learner dictionaries (almost all of them, in fact) and generated some new reference books (e.g. the Longman Grammar of Spoken and Written English). Dictionaries now have definitions that are based on the careful analysis of huge text databases, and examples that are drawn from them. New general reference grammars are similarly nowadays corpus-based, and organized around the notions of frequency in corpora, dispersion across texts and genres, and descriptive adequacy with respect to the authentic data represented by the corpora used by the publishers.

[58] David Y.W. Lee is an Assistant Professor at City University of Hong Kong, working on the application of corpora to teaching and researching English. He is currently compiling a corpus of academic speech by native and non-native speakers of English at the university, and several written learner corpora by Hong Kong and mainland Chinese students. He previously taught English communication, applied linguistics and cross-cultural communication at universities in Japan and Thailand, and also worked as a post-doctoral research fellow at the English Language Institute, University of Michigan, as part of the Michigan Corpus of Academic Spoken English (MICASE) project. His doctoral research at Lancaster University (UK) was on modelling variation in spoken and written English using the British National Corpus, for which he did the genre categorisation of texts. He recently completed co-authoring a book on corpus-based language study based on BNCweb—a user-friendly Web interface to the BNC.

[59] Sylvia is a lecturer in the School of Foreign Studies, South China Normal University, and also concurrently a PhD student at City University of Hong Kong. Her research interests include corpus linguistics, discourse analysis, ESP, systemic functional grammar and EFL pedagogy. She has published in Chinese journals and co-edited books on curriculum development and teacher training as well as on corpus linguistics in EFL education and research. She has also been involved in a number of corpus projects, including LINDSEI (Louvain International Database of Spoken English Interlanguage).

Such a corpus-based paradigm has yet, however, to make a serious mark in the world of university EAP and ESP pedagogy. One reason for this is that research on learner corpora (texts produced by learners of the language rather than native speakers) is still relatively in its infancy compared to research on native-speaker corpora, and it is precisely such research on the characteristics of learner writing that is most immediately useful for pedagogical purposes. A second reason is that currently available learner corpora are still quite limited in their generic scope and variety and do not contain the specific academic genres that non-native speakers of English are expected to write at the university level. For example, the ICLE project corpora (International Corpus of Learner English; Granger 1998) and the Chinese Learner English Corpus (CLEC; Gui and Yang 2002) all contain only argumentative or expository essays on topics such as "Crime does not pay" or "Pollution: a silent conspiracy" or "Health gains in developing countries", "My view on fake commodities", and, by design, explicitly exclude other genres such as essays of a descriptive, narrative or technical focus.

Against this backdrop, students of English language and linguistics at many universities in China, Hong Kong and elsewhere are expected to write into many academic genres that are currently not well represented in learner corpora: literature reviews, research reports, conference abstracts, bachelor's and master's dissertations, etc. The ICLE Web site, in fact, gives the following topic title as an example of what should be excluded from ICLE, but which linguistics students might certainly be expected to write about: "The position of the adverb in journalistic English". In many mainland Chinese universities, undergraduate students specializing in English language and applied linguistics are required to write a dissertation, in English, of about 7,000 words exploring some aspect of (original) research. This is a task which is of considerable difficulty even for native speakers of English, not only at the genre or discourse level, but also at the lexico-grammatical level. Genre analytic studies have provided valuable insights into the generic structure of many academic genres (including dissertations), and previous studies on Chinese EFL learners' thesis writing have tended to focus on either move structures (Swales 1990) or the discourse functions of linguistic items such as connectors and hedges (e.g., Feng and Zhou 2007; Gui and Yang 2002; Liu 2005; Mo 2005; Pan 2007; Wang 2007; Zhang 2006). The more micro-level lexicogrammatical aspects of the academic writing of both native speakers and learners in specific disciplinary fields are yet to be fully explored, however, and it is therefore not surprising that corpus-based insights are not making systematic inroads into EAP/ESP pedagogy and materials development at many universities.

In this paper, we present a data-driven, pedagogically oriented analysis of a corpus of Chinese undergraduate dissertations on linguistics and applied linguistics, focusing on characteristically problematic areas, as revealed through keywords analyses (Scott, 2000, 2001; Tribble, 2000) derived through comparisons with two comparable corpora of academic writing. The analyses are complemented by qualitative investigations of collocations and n-grams (unrestricted word clusters). Our study was prompted by the fact that such research has not (to our knowledge) been carried out before, and will meet a need in the field. Our survey of previous studies on Chinese EFL writing (such as those on discourse markers) revealed that the choice of items was usually based on intuition or an arbitrary selection of what was felt to be important. And in the case of studies that were corpus-based or corpus-driven to some extent, the data examined consisted of argumentative and expository essays by Chinese university (or pre-university) students from a mix of disciplines, or timed examination essays (Milton & Hyland 1999). For example, Chuang and Nesi (2006) investigated an error-tagged corpus of 50 argumentative essays by Chinese business studies foundation-year undergraduate students (on topics such as the ethics of genetic engineering, the European Monetary Union, methods of restricting car use, and the advantages and disadvantages of identity cards). Findings on article usage errors in argumentative writing by Chinese university students also were found to be common in corpus-based studies by Milton (2001) and Papp (2004). On the grounds that a focused analysis of just one type of text in just one specific discipline can be pedagogically more fruitful (Hyland 2007, 2008), the present study aims to fill a gap in the literature on the problems faced by Chinese students in writing an extended piece of research writing in linguistics or applied linguistics. As a corollary to our research, we also present a sample set of classroom pedagogical materials that show how the corpus-based approach can scaffold learners in acquiring the lexico-grammar of academic writing in linguistics, and thus constitute a system of writing apprenticeship that can be characterised as partly self-learning and partly exemplar- and instructor-led learning.

The Data

In order to get a fuller picture of the research writing of Chinese undergraduates in linguistics/applied linguistics, a multiple-comparison approach was adopted. Three different types of corpora were compiled for the purpose of our analysis: two were compiled from scratch, while the third was a subset of the recently released British Academic Written English (BAWE) corpus. These three corpora will now be described in detail.

The first corpus, which will be called the CAWE (Chinese Academic Written English) corpus, constitutes the main area of focus, and consists of 78 dissertations in linguistics or applied linguistics (totalling 410,388 words) written by Chinese undergraduates at a mainland Chinese university (South China Normal University)[60]. The second corpus, called the EXJA (Expert Journal Articles) corpus, consists of 56 journal articles (376,455 words in total) from a variety of high-ranking linguistics and applied linguistics journals (e.g. *TESOL Quarterly, Applied Linguistics*), with most of the articles selected to roughly match the topics of the Chinese learners' dissertations. We call the authors of these texts "experts" rather than native speakers because we deliberately did not check their native-speaker status. Indeed, many of them are probably non-native speakers, judging from their names and personal knowledge of some of the authors. On the grounds that these papers were published in the top journals in the field, and that they had all been through a proof-reading, peer-reviewing, and editorial process, there was no justifiable grounds on which to discriminate between native and non-native. Instead, they are assumed to be good models of writing to which learners can aspire. The third corpus is a subset of the BAWE (British Academic Written English)[61] corpus, which is a collection of student assignments collected from undergraduate and masters students at three British universities across thirty five disciplines. For the present study, only the linguistics and applied linguistics assignments (undergraduate and masters) written by students whose first language was English were selected, and so we will call this corpus "BAWE-L". This corpus consists of 76 files totalling 174,908 words. The BAWE texts[62] are, on average, shorter than both the CAWE and EXJA texts, because they were focused assignments rather than a major, extended piece of writing. Nevertheless, for the purposes of the present study, the texts were considered suitable for comparison with the CAWE and EXJA corpora. The BAWE assignments have been described by the BAWE project team using genre labels such as 'case study', 'critique', 'exercise', 'explanation', 'literature survey', 'methodology recount', 'narrative recount', 'problem question', 'proposal', 'research report'. The following table gives some basic statistics on the three corpora used in this study.

Name of Corpus	**No. of Texts**	**No. of Words**	**Average Length per File**	**Standardized Type/Token Ratio (%)**
CAWE Corpus: Chinese Academic Written English (undergraduate linguistics/ applied linguistics dissertations)	78	410, 373	5,261	34.69
EXJA Corpus: Expert Journal Articles corpus (linguistics/applied linguistics journal articles)	56	376, 455	6,722	38.78
BAWE-L Corpus: British Academic Written English (written assignments by British students of linguistics/ applied linguistics; 14 of the texts (22% of the word tokens) are by postgraduate students)	76	174,908	2,301	38.19

Table 1: Descriptive statistics for the research corpora

These three corpora were carefully chosen to be as comparable as possible: they are all research-oriented

[60] The authors would like to thank Anping He of South China Normal University for kindly providing us with the raw data.

[61] The British Academic Written English (BAWE) corpus was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for English Language Teacher Education, Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800)

[62] A few of the BAWE-L files were "compound files" in the sense that several similar short assignments were aggregated together into one file. (For more details on this point, please refer to the BAWE manual.) This fact does not have any real implications for the purposes of the present research.

academic writing (not argumentative, narrative or descriptive), and all in the disciplinary fields of linguistics or applied linguistics. While CAWE consists of Chinese learner texts, BAWE-L may be considered to represent the writing of native-speaker "apprentices"—writers not quite as polished as the professional linguists, and closer to the Chinese undergraduates in terms of age and amount of exposure to academic writing. We hypothesized that BAWE-L may perhaps represent something of a more realistic goal for the learners, in terms of the lexico-grammatical features of academic writing. As Milton and Hyland (1999: 149) write:

> Nor should we make the assumption, not uncommon among EFL teachers of writing, that NS students automatically and easily produce error-free and effective academic arguments. It is often forgotten that there is a huge gulf between NS student papers and the professionally edited articles of experienced academics, and that it takes many years of professional apprenticeship before NSs adopt the norms of their discourse community.

The use of BAWEL-L in the present study is, in part, meant to address the above comment.

Method

The raw data for CAWE and EXJA had to be considerably "processed" and "cleaned up" and turned into proper text files for use with WordSmith (version 3), the concordancing program chosen for our analysis. The journal articles were the hardest to deal with, as the text extracted from the PDF-format originals had to be manually (and laboriously) processed to get rid of the following: running headers and footers (e.g. the name of the journal and/or the author and the title of the paper), tables, charts, any large chunks of foreign language material, references and appendices. We also had to manually stitch back sentences that were split across pages or across intervening tables or charts, and replace "smart quotes" (curly quotes, single and double) with normal quote marks, and en- and em-dashes with hyphens, in order to make the text files more "compatible" with the WordSmith concordancing program. In addition, we also had to manually fill in quite a lot of missing characters that resulted from the use by some publishers of single-character ligatures instead of the letter sequences "ff", "fi" and "fl". In the case of the Chinese learners' dissertations, the texts came as Microsoft Word files, and so were considerably easier to process: as with the journal papers, tables, charts and other non-body-text material were removed, curly quote marks converted to normal quote marks, and the text files saved as plain text (UTF-8) files. The BAWE-L files were already in plain text format, and so required no further processing.

For the analyses in this paper, the following WordSmith settings were used: a "word" or token was defined as anything containing alphanumeric characters, and, in addition, hyphens and apostrophes were treated as word-internal characters (i.e. "classroom-based" was treated as a single word (and did not add to the counts of "classroom" or "based"), and "John's" was treated as a different word from "John"). For the keywords, the chi-square method was used for the keyness calculations and the p value cutoff was set at 0.000001.

Results

In this short paper, there is only space to present a small selection of our findings. A fuller report will follow. Our investigation of the frequency lists, keyword (and key n-gram) lists and key-keyword lists revealed that many of the underused and overused words and phrases are closed-class/function words (e.g. *can, the, some*), which are typically not present in keyword lists, and high-frequency common words (e.g. *make, besides, get, help*), which are typically not thought to be interesting because they are "non-academic". Their presence in our lists is therefore a red flag, given that the focus of a lot of academic writing instruction is often on what are considered more "difficult" or "academic" vocabulary items. In our analyses, we therefore concentrated on function words and "simple" words and found that their usage in our learner data is highly patterned rather than random, thus being amenable to remedial teaching using a data-driven pedagogical approach.

The following table presents, in descending order of keyness, a selection of items which are among the top 150 significantly "overused" and "underused" items (words, 2-, 3-, 4- and 5-grams), compared against our two reference corpora. (We are using the terms "overused" and "underused" here as purely descriptive terms—there are no assumptions that all such overused or underused items necessarily represent bad writing practices, or that NNSs necessarily have to match native or near-native speakers in frequency of usage of all items.) As mentioned above, we have concentrated on function words and seemingly "simple" words because we feel that these point to the more basic and common constructional patterns that have been not adequately addressed in current pedagogical materials. In the table, key-keywords (i.e. the key words that are dispersed across many texts) are indicated by an asterisk, while items that are key in both reference corpora (shared across the two columns) are in bold. Negative keywords are in italics. As can be seen from the table, most of the top 150 key items are not only shared across the two reference corpora, but are also key-keywords (i.e. they are frequent across most of the texts rather than being idiosyncratic).

	Keywords vis-à-vis EXJA	**Keywords vis- à -vis BAWE-L**
Keywords	*are, ***can**, *is, ***the**, ***should**, ***according,** ***make**, ***besides,** ***some**, ***out**, ***get**, ***them**, ***help**, *it, *so, *there, **good**, **know**, *use, ***while**, ***find**, ***kinds, *part,** *causes, **pay**, *two, ***great**, *they, *therefore ----- *set, indeed, me, seemed, than,* ***felt***, ***such***, ***cases***, *through,* ***despite***, *simply, appear,* ***suggest***, ***been***, *across, between, being, thus, able*	***the**, ***should**, ***some**, ***according,** ***help**, ***them**, ***while**, ***besides, *get**, ***find**, ***kinds,** *among, ***out**, *since, **know**, *kind, *new, *four, *better, ***part**, ***make**, *seldom, *most, *their, *two, *second, ***can,** ***pay**, ***great,** *of, ***good** ----- *her,* ***suggested***, ***despite***, *often,* ***been***, ***felt***, *any, these, whereby, had, allows, look, looked, throughout, does,* ***such***, ***case***, *to, fact, also, could, perhaps, yet*
Key 2-grams	***the author**, ***according to,** ***there are**, ***we can**, ***of them**, ***find out**, ***teachers should**, ***out the**, *can see, *it is, *are the, ***is to**, *they can, to make, ***most of**, ***this paper**, ***can not,** ***should be**, *how to, ***for it,** ***so the**, *it can, ***get the**, ***all the**, ***teacher should,** ***kinds of,** ***what's more**, *the above, <beyond keyword rank 150: > ***as for**	***the author**, ***according to,** ***of them**, *the subjects, *kind of, ***most of**, ***kinds of,** ***as for**, ***this paper**, ***find out**, ***teachers should,** ***out the**, ***of the**, ***all the,** ***we can**, ***should be**, **for it,** ***teacher should, *is to,** <beyond keyword rank 150: > ***get the**, ***what's more,** ***so the**, ***can not**, ***there are** ----- *it would, within a, way that, context of, relation to, led to, therefore the, to determine, as being, I would, be said, to refer, be able, important to, seems to, not necessarily*
Key 3-grams	***according to the**, ***we can see**, ***there are #**, *can see that, ***find out the**, ***to find out,** ***most of the,** ***the help of,** *it can be, ***with the help,** *the usage of, ***the data of,** *there are some, *this kind of, *from the above, *the aspect of, *so as to, ***in the process**, ***the process of,** *tend to use, **as for the,** it is necessary, *the subjects of ------ *the effects of,* ***the fact that***, ***the context of***, *in figure #,* ***in relation to***, ***appears to be,*** *with respect to, there was a, as opposed to*	***according to the**, *the present study, ***to find out**, ***most of the,** ***find out the**, ***as for the**, in relation to, ***the help of,** ***there are #,** ***with the help,** ***the data of,** ***the process of,** ***in the process**, *and so on, *a kind of, ***this kind of**, *the teacher should, *of the study, *from the data, *from table #, *different kinds of, at the same, in other words, ***we can see**, *the most important, *in this paper ----- *et al #, to look at,* ***the fact that***, *the majority of, an example of, way in which, the idea that, is important to, be seen as, found to be, the way in, to suggest that, the idea of, to refer to, the case of,* ***appears to be***, *in the case,* ***the context of***, *ways in which,* ***in relation to***, *look at the, it is important*
Key 4-grams	***we can see that**, ***to find out the**, ***is one of the,** ***with the help of**, ***there are # of,** *make good use of, ***in the process of**, *it can be seen, that is to say, *as a matter of, *a matter of fact, *it seems that the, *the number of the	***to find out the**, ***with the help of, *in the process of**, **is one of the**, at the same time, ***there are # of**, the context of the, in the case of, the way in which, it is important to
Key 5-grams	*there are # of students, *as a matter of fact, *we can see that the	----

Table 2: Preliminary Results: Selection of Key n-grams (within the top 150 positive & negative key items) for CAWE vis-à-vis EXJA and BAWE-L

As can be seen from the table, there is a wealth of interesting observations. In this short paper, only a few key items (overused in CAWE) will be touched on briefly: *the, make,* and *besides*.

Overuse of "the"

The definite article constitutes 6.3% of the tokens in EXJA but 7.4% CAWE. This represents a 1.2 times higher frequency of use by the Chinese students (the difference is significant at the p <0.001 level, using the log-likelihood test). This fact has often been remarked upon, and is often attributed to the fact that Chinese does not have an article system, leading to learners overuse it in whenever they are not sure of its necessity. Our investigation of our data, however, showed that another reason could be because of the overuse of "the author" (a key-keyword) by Chinese writers to refer to themselves (probably as a result of being explicitly told by teachers not to use the personal pronoun "I" in formal academic writing), as in the following example: "In this research, <u>the author</u> aims to solve these problems:…". In our analysis, we also found a key keyword 5-gram, "the author of this paper", that was used by four different Chinese authors to refer to themselves.

Overuse of "make"

The common verb *make* (and its inflected forms) occurs significantly more frequently in CAWE than in EXJA and BAWE-L, and is key-keyword (although not among the top 150). One reason seems to be that MAKE is used by Chinese learners in a lot more light verb constructions than native speakers (i.e. in V+NP constructions where *make* is semantically very light and contributes very little to the complex predication). The R1 collocates and 3-word clusters of the different forms of MAKE in our data reveal the following (mostly odd) collocations: MAKE + (a/an) {*analysis, conclusion, survey, adjustments, study, communication, investigation, research, judgment*…}. Expert writers, in comparison, use the above nouns with verbs other different collocations. For instance, among the first fifty collocates in the L1-L3 positions of CONCLUSION in CAWE (frequency: 122, excluding its use in section headings) and EXJA (frequency: 30), the possible verb collocates include:

CAWE: *draw/draws* (6/37), *make* (11), *come* (12), *arrive/arrived* (34/35)

EXJA: *reached* (11), *arriving* (16), *be* (20), *cites* (25), *constructed* (23), *draw* (40)

(Note: The numbers show the ranking of the word in the list of collocates.)

Another reason for the overuse of MAKE is that it is often used as a general substitute for other causative verbs. Human objects, such as *learners, one, students* and *them* collocate more often with MAKE in CAWE than in EXJA. They occur frequently in causative constructions. Here are some examples taken from CAWE.

rategy to remember new words. The goal is to **make the students use** the strategy consciously and become profic

of teamwork and cooperation, then in the end **make them master** the language in a relaxing and encouraging

which is one of the most important causes to **make them lose** interest and confidence in English. 2) The

se is not to let the students finish tasks, but to **make them be** familiar with the language forms. Two examples

can attract students' attention easily, as well as **make them relax**. Besides, after each unit, there is a funny story

ranslation. By using the method, teachers can **make the learners see** the differences between Chinese and English

ing. Besides, the teachers have to find way to **make the students open** their mouths while learning English so as to

auses: 1) The different teaching methods **make the students of Junior One unadapted** to the change from

s with their synonyms or antonyms in order to **make their students have** a better understanding of words. As time

Examples of MAKE in the CAWE corpus, showing causative constructions

Many of these examples are problematic or unnatural and can be improved by substituting MAKE with *help* (e.g. *make them master), *cause* (e.g. *make them lose interest), or *give* (e.g. *make their students have a better understanding), or by using the verb form of the adjective in the object complement, such as *familiarize them (with)* instead of **make them be familiar*. Chinese learners seem not to be aware that in English the construction "MAKE (someone) VERB" often has the meaning "force someone to do something (unpleasant/something against their will)", as attested in the following examples from the British National Corpus:

New entrants put pressure on existing companies and **make them change** their existing practices (A2H)

She must never ever do anything that might **make them fight**. (A6J)

I'm in no real position to offer the kind of money to **make them change** their mind (AAW)

I saw state policemen drag strikers across the road and **make them kneel** in the ditch there while they held shotguns in their backs. (AAX)

New books and dictionaries are expected to contain the new spellings, while Proust, Racine and the rest will gradually be re-edited so as to **make them conform**. (ABD)

I think the best way to teach hairdressers might be to **make them become** clients. (A7N)

One of the factors contributing to the overuse of the causative MAKE is that the learners may have associated this verb with令 (ling4) or 使 (shi3) in Chinese which are neutral in meaning and used more liberally and productively in Chinese causative constructions than MAKE in English. In fact, all the above examples from CAWE can be translated into 令/使 constructions in Chinese. Based on this research finding, we have developed a set of corpus-driven exercises designed to help raise the consciousness of learners concerning the use of MAKE in causative constructions.

Overuse of "Besides"

"Besides" is one of the positive key-keywords in CAWE (relative to both BAWE and EXJA). In English, "Besides, …" has the meaning "Oh, and by the way, here's some less important information (subsidiary detail)…". "Besides" can function as an adverb, meaning 'as well', but it is often used to introduce an afterthought (e.g. "It's too late to start another round of tennis now. We'll never finish before dark. Besides, it's starting to rain."). The information that comes after "Besides" is usually somewhat less important—something less crucial to the argument. "Besides" is therefore not used by native speakers to add an important new point/argument, and is particularly unsuitable for starting a new paragraph. The word also has a colloquial flavour, and is thus used more often in speech. The usages of "Besides" in CAWE below are examples of the word being used inappropriately as a straightforward alternative to "In addition" or "Moreover":

the teacher should choose the materials that are more representative and general. **Besides,** the students can be organized to watch English films, dramas, soap operas….

ocess as the learning result. As a result, students' confidence might be increased. **Besides,** their interests might be stimulated and their learning motivation might be…

Go for It is bigger than that of *JET*, i.e. *JET* contains more exercises than *Go for It*. **Besides,** it can be seen that the exercises in *Go for It* are mostly about speaking…

hen challenging one's opinions, students should give evidence to contradict them. **Besides,** students should be taught how to invite someone else to speak, to take….

Concordance of "Besides," in CAWE

The "afterthought" connotation of "Besides" therefore does not seem to have been acquired by the Chinese learners, and they overuse the expression as a consequence, thinking it means the same as another other connector. This could partly be blamed on Chinese-English dictionaries that give the following (inaccurate) translations: 此外 / 况且 (conj.) = *besides; in addition; moreover*.

Conclusion

Through his recent corpus-based analyses of academic texts, Hyland (2007, 2008) has shown that EAP practitioners really ought to be taking a more discipline-specific approach to corpus-based research and pedagogy rather than focusing on general, all-purpose vocabulary and lexical bundles, simply because the same word or phrase can often have very different meanings and functions in texts from different disciplines. He issues the following plea: "Corpus-informed lists and concordances can be used to help establish frequently occurring and otherwise productive bundles for EAP courses and the design of relevant teaching materials. It is important, however, that these lists and concordances are derived from the genres students will need to write and read" (Hyland 2008: 20). We hope that the present paper has taken a small step in this direction, in uncovering some common words and phrases that are specific to linguistics and applied linguistics, based on multiple analyses of genre-relevant learner, native-speaker apprentice and expert corpora.

References

Chuang, F-Y. and **Nesi, H.** 2006. "An Analysis of Formal Errors in a Corpus of L2 English produced by Chinese Students." *Corpora* 1/2: 251-271.

Feng Y. and **Zhou, R.** 2007. "Using Hedges in Academic Writings: A Comparative Study". *Foreign Language and Literature Studies, 2.*

Gilquin, G. and **Paquot, M.** 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1/1: 41–61.

Gui S-C. and **Yang H-Z.** 2002. *Chinese Learner English Corpus*. Shanghai: Shanghai Foreign Language Education Press.

Hyland, K. 2008. "As can be seen: Lexical bundles and disciplinary variation." *English for Specific Purposes 27/1: 4-21* .

Hyland, K. and **Tse, P.** 2007. "Is there an ‘academic Vocabulary’?" *TESOL Quarterly* 41/2: 235-253.

Liu, J. 2005. "Contrastive Study of Discourse Connectives in English Academic Papers.". *Journal of Wuhan University of Science & Technology (Social Science Edition)*, 7/4.

Milton, J. 2001. "Elements of a Written Interlanguage: A Computational and Corpus-Based Study of Institutional Influences on the Acquisition of English by Hong Kong Chinese Students". *Research Reports*, Volume 2. Hong Kong: Language Centre, the Hong Kong University of Science and Technology.

Milton, J. and **Hyland, K.** 1999. "Assertions in students’ academic essays: a comparison of English NS and NNS student writers". In Language analysis, description and pedagogy, Proceedings of an international conference organized by Language Centre, HKUST (1996), HKUST, 1999. p. 147-161.

Mo, J-H. 2005. "A Corpus-based Study of the Use of Causal Connectives in Chinese EFL Learners' Argumentative Writings". *Foreign Language Education* 26/5.

Pan, C-K. 2007. Hedges and the Instruction on Thesis Writing of English Majors, Journal Beijing University of Chemical Technology (Social Science Edition) (2).

Papp, S. 2004. "The use of learner and reference corpora to foster inductive learning and self-correction in Chinese learners of English". Paper presented at the "Meeting the Needs of the Chinese Learner in Higher Education" Conference, University of Portsmouth, July 17–18, 2004.

Tribble, C. 2000. "Genres, keywords, teaching: towards a pedagogic account of the language of project proposals." In *Rethinking Language Pedagogy from a Corpus Perspective*, L. Burnard and T. McEnery (eds.). Oxford: Peter Lang, 75-90.

Scott, M. 2000. "Focusing on the Text and Its Key Words." In *Rethinking Language Pedagogy from a Corpus Perspective*, L. Burnard and T. McEnery (eds.). Oxford: Peter Lang, 103-122.

Scott, M. 2001. "Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs." In *Small Corpus Studies and ELT: Theory and Practice*. M. Ghadessy, A. Henry & R.L. Roseberry (eds.). Amsterdam: Benjamins, 47- 67.

Swales, J. M. 1990. *Genre Analysis*. Cambridge: Cambridge University Press.

Wang, X-Y. 2007. "The Contrastive Discourse Markers in English Writing of Chinese College Students." *Journal of Guangxi Radio and TV University*, 18/4.

Zhang, B-Y. 2006. "Corpus-based Analysis of Discourse Connectors in Advanced Chinese EFL Students' Writing." Unpublished MA thesis, Beijing Normal University.

FORMULAIC SEQUENCES IN APPRENTICE WRITING – DOES MORE MEAN BETTER?

Agnieszka Leńko-Szymańska[63]

Abstract

The aim of this study is to compare the use of formulaic language by native and non-native novice writers as well as professional writers. It is also to investigate whether, and to what extent, the use of formulaic sequences affects the perceptions of human raters concerning the quality (of the lexical aspects of) the learner's written production.

The data used in this study were drawn from the PELCRA learner corpus. Three batches of 50 argumentative essays each were used in the study. They were produced by two groups of Polish learners, and by a group of American students. In addition, sections B containing press editorials were extracted from the FLOB and FROWN corpora.

Each batch of learner essays was randomly divided into two equal sections to be used for two different rounds of assessment by four human raters. In the first round the focus was on global assessment, in the second -- on vocabulary use. Subsequently, the batches of novice and professional writing were analysed in terms of recurring multi-word chunks. Next, the lists were compared with the aim of locating similarities in the use of formulaic sequences by the native and non-native apprentice writers as well as experts. Finally, the frequency of multi-word items was calculated for each essay and correlated with the raters' marks.

The results showed a fairly complex picture of the use of formulaic sequences by the three groups of novice writers. In addition, no relationship was observed between the raters' scores and learners' use of formulaic language. The interviews with the raters turned out to be very revealing in this respect. They reported that idiomaticity and the accuracy of word grammar, of which formulaic sequences are believed to be a good indicator, constituted only one of many criteria in the assessment of the essays.

Keywords: phraseology, writing, assessment, learner corpora, second language acquisition

Introduction

New approaches to second language acquisition underline the importance of formulaic sequences in L2 learning. It is frequently maintained that that phraseology is one of the hardest aspects of foreign language learning and it poses problems even to advanced L2 users. It is even claimed that apart from a faulty pronunciation, incorrect use of collocations is the most salient marker of non-nativeness of advanced learners' output.

The aim of this study is to compare the use of formulaic language by native and non-native apprentice writers as well as native professional writers. The additional goal is to investigate whether, and to what extent, the use of formulaic sequences affects the perceptions of human raters concerning the quality (of the lexical aspects of) the learner's written production.

Data

The data used in this study were drawn from the PELCRA learner corpus and two reference corpora of written English: FLOB and FROWN. The following section contains a short description of each corpus and of the sampling criteria employed for the purpose of this study.

PELCRA learner corpus contains argumentative essays produced by Polish upper-intermediate and advanced students of English at the university level. One section of the corpus contains 288 essays written on the same topic and in identical conditions (timed in-class writing). In addition, the collection comprises 81 essays also written on the same topic and in similar conditions by American native-speaking students at the university level. To ensure a close comparability of the data three batches of 50 argumentative essays were drawn from this section of the corpus. They were produced by two groups of Polish learners at the upper-intermediate and advanced levels (Year

[63] Agnieszka Leńko-Szymańska is currently on three-years' leave from Warsaw University where she holds a position of a lecturer at the Institute of Applied Linguistics. Her research interests are primarily in psycholinguistics, second language acquisition and corpus linguistics, especially in lexical issues in those fields. She has published a number of papers on the acquisition of second language vocabulary and explorations of learner corpora. She is currently working on a book on assessing vocabulary knowledge of second language learners. She teaches applied linguistics, foreign language teaching methodology and SLA.

1 and 4 respectively), and by a group of American students. Subsequently, each batch of essays was randomly divided into two equal sets to be used for two different rounds of assessment by human raters. The composition of the non-native and native student data samples used in the study (hence referred to as corpora) is presented in Table 1.

Set 1	**Set 2**
75 essays: 25 – Year 1 25 – Year 4 25 – Native	75 essays: 25 – Year 1 25 – Year 4 25 – Native

Table 1. The composition of the student corpora

Most studies involving learner corpora compare the production of EFL learners with essays written by British and American students who match the learners in age and educational background. The choice of such a benchmark over standard reference corpora containing published texts has been postulated by Granger (1998). She claims that EFL essays are equivalent to native students' compositions in terms of authors' expertise in writing, thus the observed differences will only reflect the disparities in linguistic systems and will not be a result of discrepancies in the level of writing skills. However, as it has already been pointed out by Leńko-Szymańska (2006b, 2007) a choice of native student data as a base for comparison can lead to exactly the same problem. In the process of second language learning students of English, particularly at advanced levels, usually receive a thorough training in writing. They are often exposed to more writing instruction in English than in their mother tongues. The model presented in such training is expert writing rather than native student production. In consequence, the writing of learners of English can often be closer to professional texts than to native students' essays. Therefore, in order to gain a better insight into the factors influencing the development of interlanguage it seems desirable to compare EFL learners' output not only with the production of equivalent native students but also of expert writers.

For this reason data drawn form FLOB (Freiburg-London-Oslo-Bergen) Corpus and FROWN (Freiburg Brown) Corpus were also used in the study. Both are commercially available reference corpora of British and American English respectively which match each other in size and composition. Both equal to 1 million running words and contain published texts form beginning of 1990s. Each corpus consists of 15 sections (marked with letters from A to R) which correspond to different genres of written language. Section B of the two corpora containing press editorials was drawn from the corpus for the study since it was assumed that this genre is the closest equivalent to students' argumentative essays. The size of each section is 55,458 and 55,837 running words respectively.

The choice of the two varieties of English as a benchmark for learners' production was motivated by the fact that both varieties of English are used as a linguistic model in language instruction in Poland, with the British standard being more dominant.

Procedure

Four experienced raters (two native speakers of English and two Polish teachers of English) assessed the essays. In the first round of assessment, they evaluated 75 essays (25 from each batch) on the basis of their effectiveness as written discourse. In the next round, the same raters marked the other 75 essays focusing only on vocabulary use. No information on the authorship of the essays, including the authors' level or native command of English was available to the raters. Furthermore, no additional marking guidelines were provided for two reasons. First, all the raters were experienced teachers of writing working at the same institution, thus it was assumed that they shared al least some general expectations concerning writing standards. Second, the raters have undoubtedly developed their own criteria that they normally use in their routine of marking essays and, it was exactly these personal criteria that were targeted in the study. In other words, providing the raters with a set of specified guidelines for assessment might have influenced their marks by forcing them to give more importance to certain criteria of assessment (phraseology included) that they are normally not much concerned with. The trade-off of such a solution was that the correlation between the raters was expected to be rather weak.

Subsequently, the batches of novice and professional writing were analysed in terms of recurring multi-word chunks. The multi-word chunks targeted in the study were lexical bundles or n-grams which were defined by Biber *et al.* as a recurring sequences of three or more words (1999: 990). In most cases lexical bundles cut across phrasal and clausal boundaries and they are often not recognized as fixed phrases by native speakers. The can be composed of the beginning of a main clause followed by the beginning of an embedded clause (e.g. *I don't know why*)of a noun phrase followed by the preposition typically introducing its complement (e.g. *a reason for*). Lexical bundles are relatively frequent and they reoccur in language produced by different speakers and in different situations. They are also assumed to be the basic building blocks of accurate, natural and idiomatic language (Bibet et al. 1990:990-991).

The extraction of lexical bundles from the four corpora was done with the help of Collocate (Barlow 2004), a software tool that enables users to extract from a corpus lists of n-grams. The program was set to pull out n-grams ranging from 3 to 6 words. The Mutual Information was chosen as a statistical test for the retrieval of the items and since the corpora used in the study were very small, the values for cut-off points were set very low (0,1 for the initial round of extracting bi-grams and 1,0 for the subsequent rounds). 2 occurrences were assumed to be sufficient to qualify an item for the final list. The low settings for the criteria for generating a list n-grams allowed for the extraction of all lexical bundles occurring in the four corpora. Unfortunately, at the same time it resulted in the inclusion of a lot of 'noise' in the lists, that is all the lists contained not only proper lexical bundles but also a lot of recurrent items characteristic only for a one text (e.g. *dr owen's sudden surge in popularity*) or one topic (e.g. *space nuclear power systems*). These items do not match the definition of a lexical bundle discussed above and they were named by Biber *et al.* 'local repeated combinations' (1999: 991). An attempt was made to delete such items from the lists manually, yet this process led to many arbitrary decisions, and thus it was abandoned.

Next, the lists of n-grams extracted from each corpus were compared with the aim of identifying the overlapping items.

Finally, each student essay was analysed in terms of the frequency of n-grams extracted from the reference corpora. This frequency was correlated with the raters' marks.

Results

Table 2 below presents the totals of n-gram types extracted from each of the four corpora.

corpus	**n-grams**	**corpus size**	**n-grams per 10K**
Expert	5375	111295	482,95
Year 1	1710	18652	916,79
Year 4	1864	21793	855,32
Native	2222	21230	1046,63

Table 2. Number of n-gram types extracted from the four corpora

A straightforward comparison of the totals of n-gram types in each corpus normalised per 10.000 words may suggest that native professional writers used much fewer lexical bundles than apprentice writers, both native and non-native. However, it has to be noted that the professional corpus is over five times bigger than the student corpora and thus the same effect applies here as in the case of the type/token ratio, which drops gradually with the growing size of a corpus. Furthermore, due to the fact that the lists contained a lot of 'noise' or local repeated combinations, no comparisons could be made and no conclusions could be drawn regarding the length of the lists.

More plausible seems an analysis of those n-grams that overlap between the corpora. Such a solution allows to exclude many of the local repeated combinations from the analysis. The overlap between the corpora is presented in Table 3.

corpus	**n-gram totals**	**overlap**		
		Expert	**Year 1**	**Year 4**
Expert.	5375			
Year 1	1710	183		
Year 4	1864	235	447	
Native	2222	179	233	261

Table 3. The overlap between the lists of n-grams

The results show that Year 4 students use the largest number of n-grams (235) that can also be found in the professional texts and that the native students use the smallest number of these items (179). The results of a contingency table *chi-square* statistical test relating the numbers of overlapping items to the corpus sizes demonstrated that this difference between the three corpora is statistically significant (X^2= 6,13, p<0,5). The results confirm the initial assumption that advanced learners' writing can be closer to professional norms than native apprentice writing. Such a conclusion calls once more for reconsidering the benchmark for comparison of learner data in learner corpus studies.

The overlap between the n-gram types occurring in two non-native student corpora was quite large (447). In addition to a large number of sentence and phrase fragments, the overlap lists contained several classic collocations such as *advantages and disadvantages* as well as local recurring combinations such as *a mobile phone is*. This is understandable given the fact that the essays in both corpora were written on the same topic. What seems more surprising is a much smaller overlap between the two non-native student corpora and the native student corpus, considering the fact that the American students also wrote on the on the same topic. This can only be explained by the fact that even though the topic of the Polish and American essays was identical, their content was very different and reflected the differences in rhetorical conventions existing in the cultures. These differences in the content between the Polish and American` student essays were already thoroughly explored in an earlier study (Leńko-Szymańska 2006a)

The statistical test Spearman's *rho* was used to assess the inter-rater reliability in the task of marking essays. Tables 4 and 5 present the correlations between the raters' marks in both rounds of assessment.

Round 1	**N1**	**N2**	**P1**
N2	r = 0,40* p < 0,01		
P1	r = 0,42* p < 0,01	r = 0,27 p < 0,05	
P2	r = 0,36* p < 0,01	r = 0,38* p < 0,01	r = 0,54** p < 0,01

Table 4. Inter-rater reliability: Global assessment

** – moderate correlation, * – weak correlation

Round 2	**N1**	**N2**	**P1**
N2	r = 0,49* p < 0,01		
P1	r = 0,48* p < 0,01	r = 0,50** p < 0,05	
P2	r = 0,22 p > 0,05	r = 0,40* p < 0,01	r = 0,37* p < 0,01

Table 5. Inter-rater reliability: Assessment of vocabulary

As expected at the outset of the study, the correlations between the raters were moderate or weak. This is understandable given the fact that the teachers were not given specific marking criteria for the assessment of the essays and no training session was offered to them prior to the study. The lack of correlation between the raters means that the marks given by each teacher should be treated separately in the correlational analysis. This analysis may reveal that the frequency of n-gram use influenced in varying degrees the marks assigned by different raters.

Tables 6 and 7 present the correlations between the frequency of n-grams and the marks attributed by the four raters. All the values are either statistically insignificant or close to 0. Thus, no relationship can be observed between the raters' scores and learners' use of formulaic language.

Round 1	**N1**	**N2**	**P1**	**P2**
n-grams	r = 0,37 p < 0,01	r = 0,29 p < 0,05	r = 0,09 p >0,05	r = 0,20 p > 0,05

Table 6. Global assessment :Correlations

Round 2	**N1**	**N2**	**P1**	**P2**
n-grams	r = -0,12 p > 0,05	r = 0,05 p > 0,05	r = -0,02 p > 0,05	r = 0,02 p > 0,05

Table 7. Assessment of vocabulary: Correlations

The interviews with the raters helped to explain why the frequency of lexical bundles in students' essays did not correlate even weakly with the marks in both rounds of assessment. All of the raters reported that in global assessment of essays they used several marking criteria such as content, organization of ideas, coherence and layout. The range and accuracy of linguistic units (both grammar and vocabulary) constituted only one of the factors contributing to the mark. In fact, two raters (N1 and P2) stated that they paid more attention to the communicative (discoursal) aspects of an essay that to its linguistic features. The other two raters (N2 and P1) claimed to have paid equal attention to the discoursal and linguistic features of an essay. In both cases the assessment of language formed only part of the final mark. Moreover, in the assessment of vocabulary several criteria were taken into consideration by the raters. These included collocation and idiomaticity, accuracy (semantics and word grammar), appropriate register and style, as well as lexical sophistication and variation in essays. Lexical bundles can be assumed to be a good indicator of the accuracy of word grammar and of idiomaticity, in a limited extent also collocation (see the discussion above); however they tell nothing about the semantic accuracy of individual word use, lexical sophistication or appropriate register and style. Thus, the frequency of n-grams could have a very limited influence on the final mark.

It also should also be pointed out that the above analysis did not consider the accuracy of n-gram use by native and non-native students. It is plausible that some lexical bundles were used erroneously by students which could have also influenced the mark.

Conclusion

The study demonstrated that the frequency of lexical bundles does not contribute, at least directly, to the perceptions of good writing by human learners. Yet, it has to be remembered that lexical bundles do not tell everything about the phraseology of the text. The study of the frequency of traditional collocations in the corpora can shed additional light on the factors influencing the perception of good writing.

References

Barlow, M. 2004. *Collocate 1.0: Locating collocations and terminology*. Houston, TX: Athelstan.

Biber, D., Johansson S., Leech G., Conrad S. and **Finegan E.** 1999 *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Granger, S. 1998. "The computerized learner corpus: a versatile new source of data for SLA research." In *Learner Language on Computer,* S. Granger (ed.). London-New York: Longman, 3-18.

Leńko-Szymańska A. 2006a. "The curse and the blessing of mobile phones – a corpus-based study into American and Polish rhetorical conventions." In *Corpus Linguistics Around the World,* A. Wilson, D. Archer and P. Rayson (eds). Amsterdam-New York, NY: Rodopi, 141-154.

Leńko-Szymańska A.. 2006b. "Problemy metodologiczne Kontrastywnej Analizy Interjęzyka." In *Korpusy w angielsko-polskim językoznawstwie kontrastywnym: teoria i praktyka,* A. Duszak, E. Gajek, U. Okulska (eds). Kraków: Universitas, 115-142**.**

Leńko-Szymańska A. 2007. "Past Progressive or Simple Past? The acquisition of progressive aspect by Polish advanced learners of English**.**" In *Corpora in the Foreign Language Classroom,* E. Hidalgo, L. Quereda and J. Santana (eds). Amsterdam-New York, NY: Rodopi, 253-266

WORD TYPE GROUPING IN SECONDARY SCHOOL TEXTBOOKS

Inger Lindberg[127]

Sofie Johansson Kokkinakis[128]

Abstract

According to second language research (Saville-Troike 1984), vocabulary size is the single most determinant factor for second language students in order to be successful in a school setting. This has to do with the close relationship between reading comprehension and vocabulary knowledge (Read 2000). The persistent gap between reading performance of first and second language students observed in many studies (Taube 2002) is thus intimately related to low vocabulary among second language students. According to some estimates, differences in vocabulary size between first and second language students at school start may amount to several thousand words and tend to increase over the school years (Verhoeven & Vermeer 1985). According to some researchers there is a yearly increase of approximately 3000 words in the vocabulary size of school children in general (Viberg 1993). This means that many second language students face the task of trying to close a gap in vocabulary size of thousands of words while at the same time trying to keep up with the extensive vocabulary growth of first language students.

But much could be done to make vocabulary instruction more systematic and efficient if we knew more about the vocabulary needs for successful learning in different subjects at school. Schoolbook texts constitute important data for finding answers to questions like What characterizes the vocabulary of schoolbook texts in general and in different subjects at different levels? and Which words present particular problems for students studying in their second language? To answer such and other questions related to school related vocabulary and second language learning we have compiled and analyzed a corpus of secondary school textbooks of one million words (OrdiL) with texts from eight different school subjects (Lindberg & Johansson Kokkinakis 2007). To identify and categorize various types of words in textbooks from a second language perspective, we propose a model based on earlier research by Coxhead & Nation (2001) and Hyland & Tse (2007) modified to account for all the word types of potential difficulty for second language secondary school students.

Keywords: Second Language Learning, Language Awareness, Academic Language, School books, Word classification, Semantic meaning

Background

The OrdiL corpus consists of texts contributed by three different publishers in eight different subjects: mathematics, physics, chemistry, biology, social sciences, history, religion and geography. In order to perform comparative studies and analyses on these texts we have performed computer-based disambiguation and semantic analyses at a lemma- and lexeme-level. Moreover, we have carried out word frequency analyses, including comparative analyses of relative frequency and dispersion between different subjects. To be able to distinguish characteristic features of school book vocabulary we use an equally sized reference corpus of easy-to-read texts as a point of comparison. The vocabulary of this corpus is representative of the kind of written language that children at this age can be exposed to out of school.

[127] Inger Lindberg is Professor of Swedish as a Second Language at Gothenburg University, Sweden, where she teaches in the MA and doctoral programs in Swedish as a second language. Her research interests include the study of classroom discourse from a sociocultural perspective, the role of collaborative dialogue for promoting focus on form and language awareness, and second language perspectives on language and learning.

[128] Sofie Johansson Kokkinakis is a Ph. D. in Computational Linguistics at the department of Swedish at Gothenburg University, Sweden. Her research interests are natural language processing, readability and in particular lexical analysis. She has been working with computer based models of lexica, developing a part-of-speech tagger and a chunk parser for Swedish and various other tools for lexical and semantic analysis of Swedish. The last years she has been focusing on L2-corpora and other corpora with a second language perspective when analyzing and making corpora accessible for research.

As for the classification of word types we refer to previous research by Coxhead & Nation (2001) on academic texts in which text words are grouped into word families and classified into three groups identified in relation to frequency, register and range:

(1) *high frequency words* (belonging to the 2000 most frequent and widely used word families, covering about 80 % of most texts),

(2) *academic vocabulary* (words which are reasonably frequent in academic writing and corresponding to some 8-10 % of the running words) and

(3) *technical vocabulary* (which differs by subject area and covers up to 5 % of texts) Coxhead & Nation (2001).

Coxhead and Nation's classification rests on base words with its inflected forms and transparent derivation gathered in *word families* (Bauer & Nation 1993). The most recent version of the Academic Word List (AWL) contains 570 such word families considered essential for higher education irrespective of discipline (Coxhead 2000). The assumption behind compiling words in word families is that knowledge of a base word makes understanding of derived words easier. Since words are listed without the separation of lemmas such as "fire (noun)" and "fire (verb)" and lemmas with unrelated meanings belonging to different lemmas such lemmas are grouped in the same word families. As pointed out by Hyland & Tse (2007), Hyland (2008) and Eldridge (2008), the problem of polysemic words and homopgraphs is thus overlooked by Coxhead and Nation. Since different disciplines tend to show preferences for particular uses, meanings and collocations of words, substantial disciplinary variability may be concealed in the urge of defining a general academic variety such as the AWL. Gardner (2007) has explored the construct of "word" in corpus-based vocabulary research. When counting and analyzing words, she finds three problematic areas a) morphological relationships between words, b) homonymy and polysemy, and c) multiword items. She points out that there will be lexical distortions if the concepts are not defined and taken into consideration.

The issue of homonymy and polysemic words is often highlighted in relation to second language readers who tend to have greater problems disambiguating such words than first language readers (cf. Verhallen & Schoonen 1993, Vermeer 2001, Golden 2005, Lindberg 2007). This is partly due to the fact that the vocabulary in a second language is often not as well established as in a first language and that second language learners do not have access to as many different meanings of the different words in their vocabulary as first language users. Moreover, second language learners do not to the same extent as first language users take advantage of contextual clues for the disambiguation of homonyms and polysemic words.

Method

In our study we have chosen to disambiguate words at a semantic level to capture both homographic and polysemic words. Based on analyses of frequency and range we propose two main groups of words divided into at least two subgroups. The main groups are cross-disciplinary and disciplinary specific words.

Cross-disciplinary words:

(1) the 1000 most frequent words (lemmas) hereby counting only nouns, verbs, adjectives and adverbs. These are words that would be found among the most frequent ones in most texts. Examples of these words are: och 'and', att 'to', i 'in', en 'a', ha 'to have'.

(2) School-related words often associated with formal written language corresponding to the so called 'academic words' referred to in earlier research. Examples of these words are: motsvara 'correspond to', utbredning 'extension', föremål 'object (noun)', avta 'decline'.

Disciplinary specific words:

(3) domain-related words also appearing in every day language and linked to a specific school subjects. Examples are: blandning 'mixture', klimat 'climate', muskel 'muscle', helig 'holy'.

(4) technical terms which are often unique for a particular subject. Examples of such words are: produktionsfaktor ‘factor of production’, kopplingsschema ‘wiring (connection) diagram’, kromosom ‘chromosome’.

To identify these groups of words automatically with computer-based methods we have used the following strategies.

Group 1) The thousand most frequent words are extracted by using a statistically generated list of the OrdiL-corpus consisting of the thousand most frequent lemmas. These lemmas are also among the most frequent ones in the reference corpus.

Group 2) School-related words are identified automatically by selecting words that occur in more than four subjects (with a dispersion value >= 0.4) and do not appear in the reference corpus.

Group 3) are words used in everyday language, but are not among the thousand most frequent words and occur at least twice as many times in one subject than in any other subject. They sometimes appear with several lexemes and/or lemmas and rarely occur in the reference corpus.

Group 4) are words often unique within a subject which never occur in the reference corpus. These words would typically be longer than average and more than 7 characters long.

Exploring potential and actual vocabulary difficulties

To identify words that are potentially problematic for second language students in school books it is essential to perform a deeper semantic analysis on texts and rest on a classification of word groups which is relevant in relation to findings in research on second language vocabulary development. Since second language learners often master the most frequent vocabulary in the second language words belonging to Group (1) do not generally present problems to second language students. Words in Group (2), on the other hand, consisting of abstract words characteristic of formal written language may often be unfamiliar to these students since their experience of formal registers of the language of instruction may be very limited. Words belonging to Group (3) can also be problematic since they are often domain specific and therefore might be known only in the students mother tongue in domains related to the more private sphere. This group also includes homographs and polysemic words which might appear familiar but due to a lack of vocabulary depth are not known in their full semantic potential. They may therefore cause confusion when they appear in unfamiliar meanings in the school books. The more technical words in Group (4) do not necessarily offer any particular challenge to second language learners since they represent words and phrases unknown to all students and will therefore normally be explained by the content teacher.

For answering the question “What characterizes the vocabulary of schoolbook texts in general and in different subjects at different levels?” the classification of word types applied in the project is of considerable value. To explore the question “Which words present particular problems for students studying in their second language?” we need empirical data on students.’ actual vocabulary difficulties. For this purpose we will construct a number of computerized tests focusing various types of lexical knowledge in relation to the different word types. These tests will be administered to students with different background characteristics to make comparisons of quantitative as well as qualitative aspects of vocabulary knowledge between different groups of students possible.

The need for evidence-based assessment tools focusing the academic language proficiency of second language students is becoming increasingly obvious in many schools in multilingual settings where decisions whether second language students are ready to meet the language demands of content area instruction and assessment have to be made.. The research carried out within he OrdiL project will therefore have important pedagogical applications.

References

Bauer, L. & Nation I. S. P. 1993. ‘Word families’, International Journal of Lexicography, 6, 253-279.

Coxhead, A. & Nation I. S. P. 2001. “The specialized vocabulary of English for academic purposes”. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes*:252-267. Cambridge: Cambridge University Press.

Eldrige, J. 2008. "No, There Isn't an 'Academic Vocabulary,' But…": A Reader Responds to K. Hyland and P. Tse's "Is There an 'Academic Vocabulary'?" In *TESOL Quarterly*, Volume 42, Number 1, March 2008: 109-113(5).

Gardner D. 2007. Validating the Construct of Word in Applied Corpus-based Vocabulary Research: A Critical Survey. Applied Linguistics 2007 28(2):241-265. Oxford University Press.

Golden, A. 2005. ”Å gripe poenget. Forståelse av metaforiske uttrykk fra lærebøker i samfunnskunnskap hos minoritetselever i ungdomsskolen”.In *Acta Humaniora* 227. Oslo: UniPub Forlag.

Hyland, K., Tse P. 2007. “Is There an “Academic Vocabulary?”. In *TESOL Quarterly*, Volume 41, Number 2, June 2007: 235-253(19).

Hyland, K. 2008. “The Author Replies”. In *TESOL Quarterly*, Volume 42, Number 1, March 2008: 113-114(2).

Lindberg I., Johansson Kokkinakis S. 2007. (Eds.), *OrdiL - en korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år. ROSA-rapport 8.* Institute for Swedish as a Second Language, Gothenburg University.

Lindberg, I. 2007. ”Forskning om läromedelsspråk och ordförrådsutveckling.” In I. Lindberg & S. Johansson Kokkinakis, (Eds.), *OrdiL - en korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år. ROSA-rapport 8.* Institute for Swedish as a Second Language, Gothenburg University.

Mühlenbock K. 2008. "Legible, readable or plain words" - presentation of an easy-to-read Swedish corpus. Readability and Multilingualism, workshop at 23rd Scandinavian Conference of Linguistics.

Nation, P. 2001. “Learning vocabulary in another language.” Cambridge: Cambridge University Press.

Read, J. 2000. Assessing vocabulary. Cambridge: Cambridge University Press.

Saville-Troike, M. 1984. What really matters in second language learning for academic achievement? TESOL Quarterly 18:2.

Taube, K. 2002. Reading in Sweden. In Papanastasiou Constantinos. Froese, Victor. (Eds.) *Reading Literacy in 14 countries*. Cyprus: University of Cyprus & IEA.

Verhallen, M. & Schoonen, R. 1993. “Lexical knowledge of monolingual and bilingual children”. In *Applied Linguistics*, 13.

Vermeer, A. 2001. “Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input.” In *Applied Psycholinguistics*, 22:2.

Viberg, Å. 1993. Andraspråksinlärning i olika åldrar. I E. Cerú (red.) Svenska som andraspråk. Lärarbok 2. Stockholm: Natur & Kultur.

A CORPUS-BASED APPROACH TO AUTOMATIC FEEDBACK FOR LEARNERS' MISCOLLOCATIONS

Anne Li-E Liu[129]
David Wible[130]
Nai-Lung Tsao[131]

Abstract

This paper reports a novel statistical method applied to corpora to achieve automatic correction of miscollocations. Our approach borrows the notions of collocation cluster and intercollocability from Cowie & Howarth (1995). A collocation cluster is a set of semantically similar collocations in which the various collocates and focal words in the cluster exhibit a limited degree of substitutability. Limits on this intercollocability within a cluster can lead to language learners' miscollocations. For example, the members of a cluster convey a point, communicate a point, and convey condolences can yield the overextension *communicate condolences. Instead of offering explanation of miscollocations, we use the ideas of collocation clusters and intercollocability to support automatic correction of miscollocations. A miscollocation found from English Taiwan Leaner Corpus is improve problems as in She tries to improve her students' problems. We hypothesize that the miscollocation arises from overextending the intercollocability in the collocation cluster with the core sense 'leading to a usually successful conclusion.' Possible verbs in this cluster are resolve, reduce and improve whereas the noun collocates include situation, matter, problem, quality and efficiency, etc. The intercollocability, however, is limited because it is not correct to say reduce the matter nor resolve the quality. With our approach, we successfully find reduce and resolve as corrections to the miscollocation '*improve problems.' Our approach is illustrated using 84 attested learner miscollocations. Comparing our results to those of using MI as the criterion for finding corrections, our approach shows superior performance in k-best evaluation. It also significantly improves results yielded by MI when the two are combined. Some applications to pedagogy are suggested for further research.

Keywords: automatic feedback, learner corpus, collocation, intercollocability, miscollocation

Introduction

Collocation has been widely considered as one of the criteria in gauging the language proficiency of L2 learners. Studies that compare the language of L1 speakers and L2 learners have suggested that the amount of collocations learners are capable of employing are not only fewer than L1 speakers but also are limited to a small set (Granger 1998, Nesselhauf 2004, Kaszubski 2000, Howarth, 1998). The underlying causes of learners' miscollocations are believed to be eclectic, ranging from L1 interference, the use of a near-synonym to learners' creative use. An

[129] Anne Li-E Liu received her M.A. degree in TESOL from Tamkang University. Her MA thesis was a corpus-based lexical semantic investigation in which she examined the verb-noun miscollocations produced by Taiwan learners. Her research interests are in L2 lexical acquisition, computational linguistics and lexical semantics. After receiving her MA degree, she worked as the director of the English Teaching Center at Huaxing Elementary School for two years. Other than doing research as a research assistant at the moment, she is also an adjunct lecturer at National Central University. She plans to undertake her Ph.D. studies in UK in the near future. anneliu0407@gmail.com

[130] David Wible received his Ph.D. in Linguistics from the University of Illinois at Urbana-Champaign. He has taught theoretical and applied linguistics to graduate and undergraduate students at University of Florida at Gainesville and at Penn State University in the USA. He has also taught full-time as associate professor in the Department of Foreign Languages and Literature at National Taiwan University and in the English Department at Tamkang University in Taiwan. Currently, he is a professor in the Graduate Institute of Learning and Instruction at National Central University in Taoywan Taiwan. He has twice been a visiting scholar at the Institute of Information Science at the Academia Sinica. His research interests include comparative Chinese and English syntax, lexical semantics, lexical representation, and the application of computational tools to research in second language acquisition and web-based language learning environments. wible45@yahoo.com

[131] Nai-Lung Tsao received his Ph.D. in Computer Science and Information Engineering from Tamkang University at Taiwan. He had undertaken his postdoctoral research at the Academia Sinica. He is currently a post-doc in Graduate Institute of Learning and Instruction at National Central University in Taiwan. He is also an adjunct assistant professor of the Department of Computer Science and Information Engineering at Tamkang University, Taipei Taiwan. His research interests include natural language processing, information retrieval and data mining. beaktsao@mail2000.com.tw

issue arises in a language classroom is that, unlike idioms that are introduced to learners as a whole and are frequently highlighted, collocation, though significant, is commonly ignored or seen as marginal in pedagogy; yet when assessing proficiency, particularly in writing, collocations/miscollocations use are considered a deciding factor. For instance, Howarth (1998) compares the academic writing of advanced L2 learners and native speakers by exploring the collocational density and use of each. Although there is no correlation between the general proficiency of a learner and the number of collocations used as Howarth points out, the fact that collocation remains an unresolved issue even for advanced learners is worthy noticing. In an investigation of a 4-million word corpus of English writing produced by learners in Taiwan (English Taiwan Learner Corpus), Liu (2002) reveals that verb-noun miscollocations make up the bulk of the lexical collocation errors in learners' compositions.

To specifically deal with learners' miscollocations, researchers have applied statistical techniques to corpora and have created a range of resources for teaching and learning collocations (Aston 1997, Chambers 2005, Curado 2001, Shei and Pain 2000, Kita and Ogata 1997, Horst et al. 2005, Sinclair 1991, Stubbs 2002.) Such resources, however, mainly are used in providing authentic input in general but are not developed for the specific need of assisting learners in remedying their miscollocation output. We therefore propose our method in that not only lexical collocations from BNC are retrieved, such information are further applied in offering suggestions for learner miscollocations.

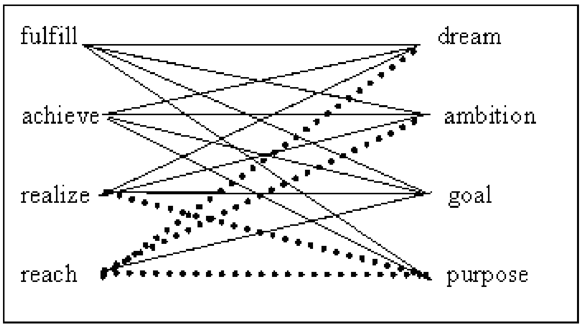
The Study

This study aims to explore an approach to finding correct collocation suggestions for verb-noun miscollocations via the notion of intercollocability. Collocation formation involves the knowledge of word meaning as well as that of the possible neighboring words. That is, dealing with miscollocation, instinctively, will require the knowledge of semantics. Our approach differs from others in that the approach relies solely on the concepts of collocation cluster and intercollocability without reference to semantic knowledge sources (such as WordNet). How do we deal with lexical semantic problems without the inference to semantic knowledge source is explained in what follows.

Intercollocability, Collocation Cluster and Miscollocation

In examining non-native writer's academic writing, Cowie and Howarth (1995) propose that certain collocations form clusters on the basis of the shared meaning they denote and this collocation cluster has in turn caused the difficulty that non-native writers face in writing. Collocations in a collocation cluster exhibit a certain degree of intercollocability, the collocates of some of these collocations being substitutable. For example, a collocation cluster sharing the sense 'pass on message from this person to another' includes related expressions *convey a point*, *get across the message* and *express opinion*. The shared intercollocability implies that collocations such as *convey a point, convey a message; get across a point, get across a message* are all acceptable. Cowie and Howarth propose that one cause of collocation errors occurs when a learner overgeneralizes the intercollocability from one collocation cluster to an adjacent cluster or within a cluster; that is, learners might overlook the fact that some verbs collocate with some but not all of the nouns in a cluster or only appear in one cluster whereas others might occur more than one cluster and thus co-occur with two different groups of noun collocates. For example, they suggest this is the underlying cause of the miscollocation **communicate condolences* found in the writing of non-native speakers. Specifically, it is seen as an overextension of the similarity in collocability that the verbs *communicate* and *convey* exhibit: *convey a point/idea*, *communicate a point/idea*, *convey condolences*, **communicate condolences*.

Another miscollocation that is produced due to the intercollocability confusion is '**reach* their *purposes'* found from EnglishTLC (English Taiwan Learner Corpus). *Fulfill goal/dream*, *achieve ambition/dream*, *realize goal* and *reach goal* can be clustered to express 'a condition or an object that is longed for.' Figure 1 shows the collocation cluster formed for the above concept. The key here is that not all verb-noun combinations in Figure 1 are acceptable since the intercollocability is not complete. While *fulfill* and *achieve* collocate with the four nouns on the right, *realize* collocates with *dream* and *goal* but not with *purpose* (as is indicated by the dotted line). *Reach* shares with the verbs in this cluster the property of collocating with *goal*, yet the similarity ends their since it does not collocate with *dream*, *ambition* or *purpose*. Thus, the miscollocation *reach* their *purposes* can be seen as an overgeneralization of the intercollocability that reach shares with the other verbs in this cluster. Due to the complex intercollocability, learners might easily assume that *realize* and *reach* also collocate with *purpose* and thus produce **reach* their *purposes.*



While Cowie and Howarth point out that limitations on intercollocability within a collocation cluster lead to learner miscollocations, we conversely view a collocation cluster as a potential source for finding the correct counterparts for miscollocations. In other words, a cluster can serve as a bridge that links miscollocations back to the correct collocates. This approach requires, then, that for a particular miscollocation, we need to determine a collocation cluster that it belongs to which could contain its correct counterpart collocation. How these collocation clusters can be created is detailed in the next section.

Methodology

To make the bridge possible, we use learner miscollocations as the starting point. We first illustrate the approach with the attested miscollocation *improve problems* as in *She tries to* ***improve*** *her students'* ***problems***.

Hypothesizing that a miscollocation arises from overextending the intercollocability in a collocation cluster, we assume that if this cluster can be determined, the correct counterpart collocation will be found in it. Thus, the first step is to take the miscollocation **improve problem* as a seed to generate the relevant collocation cluster. To do this, we use Collocation Explorer[132], a search engine that shows possible co-occurring words by retrieving collocates in BNC. The retrieved collocations are listed on the basis of their MI ranking. This generates 52 noun collocates for *improve* and 86 verb collocates for *problem.* To find correct replacements for *improve* in the miscollocation **improve problems* we determine from these verbs which ones take the same collocate nouns as *improve* does. It is found that two verbs, *resolve* and *reduce*, have the same noun collocates as *improve* (resolve/improve + situation/matter/way ; reduce/improve + quality/efficiency/effectiveness.) This constitutes the relevant collocation cluster and provides a link between *improve, resolve* and *reduce* by virtue of their shared noun collocates (See Figure 2).

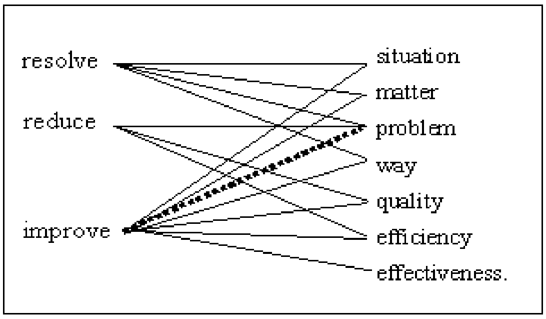


Figure 2. Cluster collocation for '**improve problem*'

[132] Collocation Explorer. http://research.iwillnow.org/project/bncrce/default.htm

Since unlike *improve*, both *resolve* and *reduce* have a high MI score with *problem*, we take them as acceptable substitutes for *improve* in the miscollocation **improve problem*. These two verbs, *resolve* and *reduce*, will thus be offered to learners as suggested corrections for the miscollocation ***improve** her students'* ***problems***.

To test the robustness of this approach, 84 miscollocations found from Liu's study (2002) are incorporated and tested. We randomly choose half to be the training data and the other half to be the testing data. Two experienced English teachers (one native and one non-native English speaker) manually examine the suggestions to judge their precision as suggestions for the tested 84 miscollocations.

Results and Discussion

In dealing with learner miscollocations, we are mainly interested in whether such clusters and intercollocability are plausible sources for identifying the correct substitutes for a miscollocation. Table 1 shows some examples of miscollocations and the corresponding correct collocations provided manually by the two human experts. That is, if our approach finds verb-noun combinations that match the correct collocations found by human experts, we will treat such matches as true positives.

Miscollocation	Correct Collocations
pay time	spend time, devote time
make damage	cause damage,
pay effort	spend effort, make effort
get knowledge	gain knowledge, acquire knowledge
make conclusion	draw conclusion, form conclusion, lead conclusion

Table 1. Examples of miscollocations and the correspondent correct collocations

Since one of the traditional methods of retrieving collocations is by statistical measures of word association strength, we compare our results with results from using one of the most common of these measures, mutual information (MI), to identify the collocation corrections (Pecina and Schlesinger 2006). For convenience, we label our approach as ON for 'overlapping nouns.' We list the precision of k-best suggestions for considering MI and overlapping nouns in Table 2, where MI means Mutual Information and ON means Overlapping Nouns. The precision values for each indicate that our approach, which considers overlapping nouns exclusively, consistently outperforms the MI approach. To see what influence the ON approach exerts in a probabilistic model, we then combine the two approaches to construct a hybrid model. The precision of k-best suggestions of this hybrid model is compared with that of the MI and ON used independently in Table 3.

K-Best	MI	ON
K=1	16.67	22.62
K=2	36.9	38.1
K=3	47.62	50
K=4	52.38	63.1
K=5	64.29	72.62
K=6	65.48	75
K=7	67.86	77.38
K=8	70.24	82.14
K=9	72.62	85.71
K=10	76.19	88.1

Table 2. K-best suggestion results of the 84 miscollocations for MI and ON approaches

As Table 3 shows, the hybrid model, which combines the two features (MI + ON), provides the highest proportion of true positives at every value of k. It is clear that our approach that considers overlapping nouns not only outperforms the MI model but boosts the precision of MI when used in combination with it.

K-Best	MI	ON	MI+ON
K=1	16.67	22.62	29.76
K=2	36.9	38.1	44.05
K=3	47.62	50	59.52
K=4	52.38	63.1	72.62
K=5	64.29	72.62	78.57
K=6	65.48	75	83.33
K=7	67.86	77.38	86.9
K=8	70.24	82.14	89.29
K=9	72.62	85.71	92.86
K=10	76.19	88.1	94.05

Table 3. K-best suggestion results of the 84 miscollocations for MI, ON and the hybrid model.

We will now look at some miscollocations with the found replacements in Tables 4, 5, 6, and 7 and discuss the intercollocability in more detail. Each of the Tables 4-7 shows the k-best suggestions provided by the MI and ON models for k=1-5 for the following four miscollocations, *_get knowledge_, *_pay time_, *_make conclusion_ and *_fill ambition_. The pluses (+) indicate the true positives in the tables. Each model finds correct suggestions for the correspondent miscollocations with the different ranking of k-best results. Figures 4, 5 and 6 refer to verb-noun collocation clusters formed for the three miscollocations, *_get knowledge_, *_pay time_ and *_make conclusion_.

Miscollocation: get knowledge		
K-Best	MI	ON
K=1	impart knowledge	provide knowledge
K=2	+acquire knowledge	+obtain knowledge
K=3	broaden knowledge	increase knowledge
K=4	detail knowledge	secure knowledge
K=5	possess knowledge	+gain knowledge

Table 4. The K-Best suggestions for *_get knowledge_.

Miscollocation: pay time		
K-Best	MI	ON
K=1	bide time	+invest time
K=2	waste time	date time
K=3	+spend time	+spend time
K=4	idle time	occupy time
K=5	while time	last time

Table 5. The K-Best suggestions for *_pay time_

Miscollocation: pay time		
K-Best	MI	ON
K=1	jump conclusion	support conclusion
K=2	+reach conclusion	+form conclusion
K=3	+draw conclusion	base conclusion
K=4	leap conclusion	+arrive conclusion
K=5	+lead conclusion	+reach conclusion

Table 6. The K-Best suggestions for '**make conclusion*

Miscollocation: pay time		
K-Best	MI	ON
K=1	+fulfill ambition	+fulfill ambition
K=2	harbor ambition	+achieve ambition
K=3	+achieve ambition	harbor ambition
K=4	lack ambition	lack ambition
K=5	+realize ambition	have ambition

Table 7. The K-Best suggestions for **fill ambition*

The two approaches seem to perform equally well since the total true positives found for the four miscollocations are nearly the same (8 for MI approach and 9 for ON approach.) Yet when k is 5, our approach finds *gain knowledge* for **get knowledge*, *reach conclusion* for **make conclusion* whereas the MI approach finds *lead (to) conclusion* for **make conclusion* and *realize ambition* for **fill ambition*. If we look at the k-best suggestions for k=1 to 2 only, it is evident that our ON approach finds more suitable alternatives.

The collocation cluster shown in Figure 4 expresses the concept of 'gaining possession of something.' Despite the fact that *knowledge*, *certificate*, *qualification* and *reputation* co-occur with *gain*, *acquire* only shares the overlapping nouns, *knowledge*, *qualification* and *reputation* whereas *get* only collocates with *certificate* and *qualification*. While it is impossible to determine from this data alone whether this incomplete intercollocability is the cause of the learning difficulty, giving rise to the miscollocation *get knowledge*, it does clearly provide us with a computational bridge from the miscollocation to candidate corrections for it.

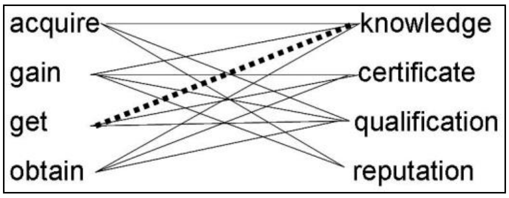


Figure 4. Collocation Cluster for '**get knowledge*'

There are 115 collocations obtained from English TLC that are used to mean 'gain knowledge.' 75 out of these 115 are instances where 'get' is employed, rather than the correct verb collocate, *gain*, *acquire* or *obtain*. This error ratio suggests that every 6 out of 10 learners wrongly assume that *get* co-occurs with *knowledge.* With this cluster information made available, the path to the corresponding corrections can be generated automatically.

Figure 5 shows the cluster created for the miscollocation **pay time*[133]. The intercollocability suggests that when showing the concept of 'to use up something or to give one's time, attention or self to a particular activity, pursuit or person,' what could be given or used up requires different verb collocates.

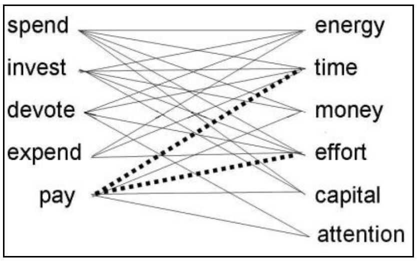


Figure 5. Collocation cluster for '*pay time' and 'pay effort.'

As a matter of fact, the cluster is a merged version of two separate clusters (see Figure 6). When what is being exerted is related to individuals' mental and physical status, such as *energy*, *time*, *effort* and *attention*, verbs used to describe the action are *spend*, *invest*, *expend* and *devote*. On the other hand, if what is being used up is detachable from individuals like *money* and *capital*, another set of verbs, *spend*, *invest* and *pay*, are the right collocates. The two clusters, although similar, reflect two different composition semantic sets. By constructing such neighboring clusters, we not only find a plausible explanation for learner miscollocation, we further take a practical yet crucial step toward offering solutions for such miscollocation problems.

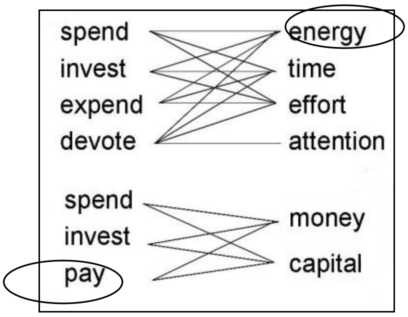


Figure 6. Two sub-clusters found for '*pay time.'

Another example of overgeneralizing the overlapping nouns is **make conclusion* for which its correspondent cluster is shown in Figure 7. All the possible collocations in this cluster show more or less 'the act of reasoning.' The word 'conclusion' connotes the attributes of 'inference being made,' 'comparison being drawn' or 'judgments being reached' first; a *conclusion* could be 'formed' only after the above processes. That is, 'arriving at a conclusion' is not an action that is done exclusively but a result or outcome of an act of process. Language learners might neglect the subtle semantic properties on the one hand and draw too general a conclusion on the other and thus produce a miscollocation like **make* a *conclusion*.

[133] Although we do not include '*pay effort' in the discussion, 15 instances of '*pay effort' are found from English TLC. Thus the miscollocation 'pay effort' is also indicated in the cluster.

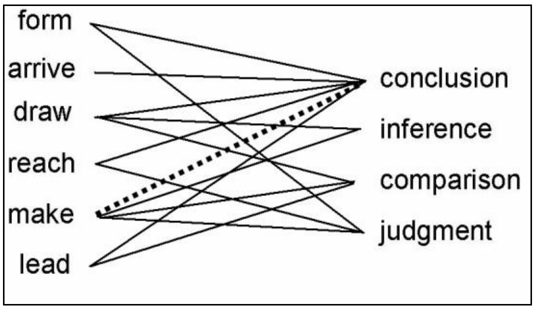


Figure 7. Collocation Cluster for **make conclusion*

One miscollocation that is not included in the 84 miscollocations is **communicate condolence* that Cowie and Howarth (1995) find after examining advance learners' writing. With the proposed approach in this paper, our model successfully finds the correct alternatives which are shown in Table 7 even though both MI and ON approaches yield the same results. With only a few verbs that occur in this cluster, the k-best suggestions offered by the two models only show up to K=3. Seen from another perspective, the small set of results for this specific miscollocation has no false positives; it shows perfect precision.

Miscollocation: communicate condolence		
K-Best	MI	ON
K=1	*express condolence	*offer condolence
K=2	*offer condolence	*send condolence
K=3	*send condolence	*express condolence
K=4, K=5		

Table 8. The K-Best suggestions of different models for **communicate condolence*

Conclusion

Collocations which can be clustered show a certain level of intercollocability that hinders learners from developing collocational competence (Cowie and Howarth, 1995). We borrow this concept and present the value and contribution of the reported approach in finding suggestions for specific miscollocations via the notion of intercollocability. The 84 miscollocations tested in this paper have shown that without incorporating semantic knowledge, our approach offers a promising means of correcting miscollocations, one type of lexical problem that arises due to the semantic confusion. Further, our approach, which considers overlapping nouns only, performs better than the traditional MI approach, and such information improves the precision even more when being combined with the probabilistic MI model.

The analysis of miscollocation helps researchers and teachers to understand further what learners lack in producing acceptable word combinations. What helps learners more, however, is something that is accessible when they are in need of such information. Thus, the next step of this research is to implement the approach as an application in the digital writing environment of IWiLL, an online language learning platform (Wible et al., 2000). The main issue of interest there will be how useful such information could be both in the writing process of non-native writers and in the feedback process of teachers. To what extent this could affect learners' writing, particularly in producing collocation, will also shed light on the pedagogical benefits of corpora in language teaching.

References

Aston, **G**. 1997. "Small and Large Corpora in Language Learning." In B. Lewandowska-Tomaszczyk & J. P. Melia (Eds.), *Practical Applications in Language Corpora*, 51-62. Lodz, Poland: Lodz University Press.

Chambers, A. 2005. "Integrating Corpus Consultation in Language Studies," In *Language Learning & Technology*, Vol. 9, No. 2, May 2005, 111-125.

Cowie, A.P. & P. Howarth. 1995. "Phraseological competence and written proficiency." In *Language and Education*, G. M. Blue and R. Mitchell (eds.), 80-93, Clevedon: BAAL in association with Multilingual Matters.

Curado, **A.** (2001) "Lexical Behaviour in Academic and Technical Corpora: Implications for ESP Development." In Language Learning & Technology 5: 106-129.

Granger, S. 1998. "Prefabricated patterns in advanced EFL writing: collocations and formulae." In *Phraseology: theory, analysis and applications*. Cowie, A. (ed.). Oxford University Press, Oxford, 145-160.

Horst, M., Tom Cobb & **Ioana Nicolae.** 2005. "Expanding Academic Vocabulary with an Interactive On-line Database." In *Language Learning and Technology, 9*(2), 90-110.

Howarth, P. 1998. "Phraseology and second language proficiency," In *Applied Linguistics*, 19(1), 22-44.

Kita, K. and Hiroaki Ogata. 1997. "Collocations in Language Learning: Corpus-based Automatic compilation of Collocations and Bilingual Collocation Concordancer," In *Computer Assisted Language Learning*. Vol.10, No. 3, 229-238.

Kaszubski, Przemyslaw. 2000. *Selected Aspects of Lexicon, Phraseology and Style in the Writing of Polish Advanced Learners of English: A Contrastive, Corpus-based Approach.*

Liu, Anne Li-E. 2002. *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners' English*. M. A. Thesis, Tamkang University, Taipei County, Taiwan.

Nesselhauf, Nadja. 2004. *Collocations in a learner corpus.* Amsterdam ; Philadelphia : J. Benjamins.

Pecina, P. & **Pavel Schlesinger.** 2006. "Combining Association Measures for Collocation Extraction," In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Main Conference Poster Sessions. 651–658.

Shei, C.-C., & Pain, H. 2000. "An ESL writer's collocational aid." In *CALL*, 13(2). 167~182.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics.* Oxford: Blackwell.

Wible, D., Chin-Hwa Kuo, Anne Li-E Liu, **Nai-lung Tsao.** 2000. "A Web-based EFL Writing Environment: Integrating Information for Learners, Teachers, and Researchers," In *International Conference on Computers and Education/ International Conference on Computer-Assisted Instruction (ICCE/ICCAI 2000)*, National Tsing Hua University, Taipei, November 2000. Recipient of Outstanding Paper Award.

POLISHING PAPERS FOR PUBLICATION: PALIMPSESTS OR PROCRUSTEAN BEDS?

John McKenny[134]

Karen Bennett[135]

Abstract

Portuguese academic discourse of the humanities is notoriously difficult to render into English, given the prevalence of rhetorical and discourse features that are largely alien to English academic style. The aim of this study ws to test the hypothesis that some of those features might find their way into the English texts produced by Portuguese scholars through a process of pragmalinguistic and sociopragmatic transfer. If so, this would have important practical and ideological implications, not only for the academics concerned, but also for editors, revisers, teachers of EAP, translators, writers of academic style manuals and all the other gatekeepers of the globalized culture.

The study involved a corpus of some 113,000 running words of English academic prose written by established Portuguese academics in the Humanities, which had been presented to a native speaker of English (professional translator and specialist in academic discourse) for revision prior to submission for publication. After correction of superficial grammatical and spelling errors, the texts were made into a corpus, which was tagged for Part of Speech (CLAWS7) and discourse markers (USAS) using WMatrix2 (Rayson 2003). The annotated corpus was then interrogated for the presence of certain discourse features using Wmatrix2 and Wordsmith 5 (Scott 2006), and the findings compared with those of a control corpus, Controlit, of published articles by L1 academics in the same or comparable journals.

The results reveal significant overuse of certain features by Portuguese academics, and a corresponding underuse of others, suggesting marked differences in value attributed to those features by the two cultures.

Keywords: academic discourse, humanities, Portuguese, English, research articles corpus

Introduction

English academic discourse, which emerged in the 17th century as a vehicle for the new rationalist/scientific paradigm (Halliday & Martin, 1993:2-21, 54-68; Martin, 1998), now holds hegemonic status on the world stage, and mastery of it is essential for any scholar wishing to pursue an international career. However, it may not be taken for granted that all cultures construe knowledge in the same way. In Portugal, which did not experience a Scientific Revolution as such, an older humanities-based tradition was perpetuated by an education system grounded on Scholastic and Rhetorical principles; as a result, Portuguese academic discourse in the humanities contains features that are markedly different from the hegemonic English style (Bennett, 2006, 2007a, b).

This study is designed to test the hypothesis that some of those discourse features may manifest themselves in the English-language research articles produced by Portuguese scholars, over and above the kind of cross-linguistic transfer that is expected on the level of grammar and lexis (Odlin 1989). On the basis of the results found, it discusses the role of the language reviser in polishing such texts for publication, and reflects upon the potential contribution of corpus studies in raising awareness of discourse differences on both sides of the cultural divide.

[134] John McKenny is Head of the Division of English Studies at the University of Nottingham, Ningbo China and is Head of the Centre for Research in Applied Linguistics, Ningbo. He previously worked as a Professor Adjunto at Viseu Polytechnic (Portugal) for twelve years and as Senior Lecturer at Northumbria University for 5 years. His Ph.D. thesis at Leeds University was entitled A corpus-based investigation of the phraseology in various genres of written English with applications to the teaching of English for academic purposes. He is currently co-editing with Tometro Hopkins a volume entitled Englishes of the British Isles to be published this year by Continuum International. This book is the first of a 15-volume series on World Englishes.

[135] Karen Bennett is a member of the Centre for Comparative Studies, University of Lisbon, where she researches in Translation Studies. Her PhD in English Academic Discourse: its Hegemonic Status and Implications for Translation is based upon her extensive experience as translator and teacher of Academic Discourse and Translation with the Catholic University of Portugal, University of Coimbra, Polytechnic of Leiria and British Council, amongst others. She has published a number of articles on this subject and others, including 'Galileo's Revenge: Ways of Construing Knowledge and Translation Strategies in the Era of Globalization' in Social Semiotics, Vol. 17, No. 2 (2007), and 'Epistemicide! The Tale of a Predatory Discourse' in The Translator, Vol. 13, No. 2, (2007).

Many epistemological issues were raised during the course of this experiment. If the researcher has strong intuitions as to why a group of writers write in a certain way based on long experience of teaching EAP, translating and polishing papers, should these intuitions be brought to bear at the outset of the corpus analysis? This runs counter to the position of Sinclair (2004) and Tognini-Bonelli (2001) who each recommend adopting a *tabula rasa* stance towards the data at the outset. As researchers we had to decide whether we were doing *corpus-based* or *corpus-driven* linguistics (Ooi 1998).

Methods and Results

The research was based on the comparison of two corpora each of around 113,000 words. The corpus under investigation, dubbed *Portac* consisted of a sample of articles from the area of the Humanities or Arts written by a group of senior Portuguese academics aiming to publish their work in English-language journals. The control corpus (*Controlit*) was a collection of articles already published by L1 academics in the same or comparable journals.

Two software suites were used for this study in a complimentary fashion. Wmatrix2 (Rayson 2003), available to scholars online, enables the investigator to compare two corpora and continually shift focus as trends become apparent; that is to say, researchers may quickly compare lexical, grammatical or semantic dimensions from the perspective of one or other of the corpora.. Wordsmith Tools 5 (Scott 1999) was used to carry out searches which are not available on Wmatrix2 such as the creation of frequency counts of words, or searches using a *wild card* (for example, for polysyllabic noun forms, a frequency list of all words ending in *ion). Results of corpus comparison in Wmatrix2 and in Wordsmith Tools are expressed in terms of Log Likelihood (henceforth LL), which measures the likelihood that a difference between the observed frequency of an item and its expected frequency is not random. The higher the LL value, the more significant is the difference between two frequency scores. An LL value of 3.8 or higher is significant at the level of $p < 0.05$ and an LL of 6.6 or higher is significant at $p < 0.01$.

Probably the most significant finding was the high degree of *nominalization* present in the writing of Portuguese academics compared to the control corpus. This was manifested in a number of ways. On the basic level, there was an *overuse of nouns,* both *singular* (LL 25.17) and *plural* (LL 69.81), and, as might be expected in such a context, a g*reater use of indefinite and definite articles* (LL 43.81 and LL 36.13 respectively). Concomitant with this, there was also a massive *underuse of pronouns* in *Portac,* 6154 (6.11% of all text) vs. 8671 in *Controlit* (8.49%), giving an astonishing Log Likelihood of 394.98. This may represent a straightforward consequence of nominalization; for, as Biber *et al*.(1999:92) conclude from analyzing various corpora totalling 40 million words, ‘a high frequency of nouns/…/corresponds to a low density of pronouns’. However, the *Portac* writers also seem to be selective about the pronouns they avoid: *he* (*LL* 232), *she (LL 104), him (LL 96), I (LL39, me (LL37), it (LL 25.74* were all underused, while *we* (39.41) and *us (*16.85) were overused. This would seem to indicate that there may be some other mechanism at work, as we discuss below.

Of the nouns employed, Portuguese authors appear to have a penchant for *polysyllabic abstract nouns of Latinate origin*. Using Wordsmith 5 to search on **ion*, 2184 instances of this suffix were obtained in *Portac* compared to only 1458 in *Controlit* (the Log Likelihood of such a difference is 163), while the results for *–icity*, *-ization* and *–ation* gave LL7.07, LL14.16 and LL50.71 respectively. Hofland and Johansson (1982:22) suggest that the high frequency of the indefinite article *an* found in written informative prose indicated a high proportion of Latinate vocabulary. The Portuguese writers’ overuse of *an* (LL 18.65) indicates their greater use of Latinate word tokens consonant with their mother tongue’s close filiation with Latin. *Adjectives* were also more prevalent in *Portac (*46.58), which once again indicates a heavy concentration of semantic content in the noun phrase.

Perhaps also related to the tendency for nominalization was a truly startling *overuse of the genitive*, both singular and plural (*‘s* and *s’*) (LL 211.64), and also the *of* alternative for expressing the same relationship (LL 34.03). In some cases, this may simply reflect the difficulty that non-native speakers have with English compound nouns (examples from *Portac* include ‘the world’s population’, where a native speaker might prefer ‘the world population’ or ‘Luanda’s slums’ instead of ‘the Luanda slums’). Elsewhere, however, it seems to directly derive from the tendency to over-nominalise. For example, the genitive in the noun phrase ‘a comment on the possibilities of the play’s staging’ was reconstrued by the reviser using a clausal form (i.e. ‘a comment upon how the play might be staged’).

As regards other syntactic features, *Portac* authors also generally produce *longer sentences* than their *Controlit* counterparts (mean sentence length 35.04 vs. 29.83 words), and make greater use of *embedding structures* (such as ‘We can see that…’ ‘It should be pointed out that...’ See Table 1). There is also a higher occurrence of *appositional forms introduced by adverbs* (LL. 67.76), such as ‘namely', ‘i.e.', ‘that is’, 'in other words’.

Search words with wild cards (*)	*Portac*	*Controlit*
It * * that	42	28
It is * * that	27	22
We*that	17	4
We** that	15	3

Table 1 Embedding Structures

All these features together make the English prose of Portuguese academics seem very dense and abstract in relation to that of their native speaker counterparts, and this may ultimately affect their chances of getting their work published. However, before looking at solutions to this problem, let us first discuss possible reasons for these differences.

Discussion

Although nominalization has been a central feature of English academic discourse since the emergence of scientific writing in the 17th century (see Halliday & Martin, 1993; Martin & Veel, 1998), there is evidence to suggest that the 'historic drift towards thinginess' (Halliday, 1998:211) may have gone into reverse in recent years. Certainly, public English generally seems to be becoming more 'conversational' and informal (Fairclough 1994, 1997), and one of the ways in which this is manifested is by a new preference for clausal structures above nominalizations (see Leech *et al.* 2001:294). Could it be that Portuguese academic writers are somewhat lagging behind in this respect, reluctant to accompany such innovation or less able to respond to the trend?

If this were all there were to it, then the problem would seem to be easily solvable through effective teaching, designed to raise L2 writers' awareness of nominalization and encourage a more clausal-based style. If, however, there are cultural reasons for the markedly different style employed by Portuguese authors, as we suspect, the issue becomes ideologically more complex.

Although Portuguese academic writing has not been systematically studied by descriptive or historical linguists, ongoing work in the area of Translation Studies (Bennett, 2006, 2007a, b) seems to suggest that there does in fact exist a Portuguese discourse of the humanities that is quite distinct in terms of its aims and values to those espoused by Anglophone culture, and which has its roots in the scholastic/rhetorical tradition perpetuated by a Catholic education system. It is therefore reasonable to assume that many of the differences between *Portac* and *Controlit* may be accounted for by a tendency on the part of Portuguese academics to transfer stylistic and rhetorical features that are valued in their own culture into their English writing.

In broad terms, these stylistic preferences include: a taste for 'copiousness' (manifested by a general 'wordiness' and redundancy); a preference for a high-flown erudite register over the demotic (evident in both syntactical structure and lexical choices), and a tendency towards abstraction and figurative language. There are also differences as regards textual organization: a propensity for indirectness means that the main idea is often embedded, adorned or deferred at all ranks; while cohesion is frequently achieved through elaborate synonyms rather than by ellipsis or pronoun substitution as might be preferred in English (Halliday & Hasan, 1976; Mateus et al. 1989: 146). This fact may have contributed to the low pronoun count that was found in the *Portac* corpus. Further corpus analysis is needed here to substantiate our strong intuitions. To our knowledge, the relative frequencies of different cohesive devices have not yet been systematically counted in either English or Portuguese. A corpus investigation of this area would provide a very useful contribution to research in Contrastive Rhetoric and Translation Studies.

Although it has not been possible to test for the presence of all these features in *Portac*, the findings listed above would seem to point to the persistence of the Portuguese value system in these English texts. For example, the heavy nominalization not only makes the prose sound more 'learned' and 'literary', it also has the effect of turning contingent observations into abstractions, a quality that is reinforced by the prevalence of polysyllabic Latinate words and lexical abstractions (i.e. nouns ending in *-ion*, *-icity*, *-ization*). The long sentences and embedding structures reproduce the copiousness and indirectness of Portuguese prose, while the proliferation of adjectives and appositional structures also serve to 'pad out' the discourse, creating an impression of abundance. Finally, the overuse of the first-person plural pronoun is a direct transposition of the Portuguese authorial 'we', used systematically even when the text has been penned by a single author (as was the case with all the *Portac* texts).

Conclusion

If the differences between *Portac* and *Controlit* can indeed be explained by the intrusion of Portuguese discourse features into the English prose produced by Portuguese academics, this raises important questions of both a practical and an ideological nature. Firstly, to what extent does this transfer jeopardize the chances of Portuguese academics being published in international journals? We know that verbosity, unnecessary complexity, abstraction and 'pomposity' are generally eschewed by arbitrators of style in English academic prose; but are editors and referees aware that other cultures may value these qualities differently? Would such an awareness alter their perception of the quality of the work submitted and therefore affect the international status of the authors in question?

Secondly, to what extent should texts like these be domesticated in order to bring them into line with the Procrustean norms imposed by the hegemonic culture? Are revisers, editors and proofreaders at liberty to erase or alter discourse features that transmit value and are therefore profoundly bound up with questions of identity? Or might this constitute a form of cultural imperialism, or even 'epistemicide' (Santos, 2005; Bennett, 2007b), all the more insidious because it undermines the very conceptual framework upon which the author's worldview is based? And what of the alternative, the 'palimpsest', that allows the thought patterns of the original version to be glimpsed beneath the surface structure? Can we guarantee that this will find a readership, even if it gets past the editors and referees? It is, after all, so much more tiring for readers to process sentences that do not fall in the way that one expects them to.

Corpus Linguistics may have a useful role to play in this debate. Communication is now understood to be far more complex than theoretical notions of 'standard English' would have us believe, and there have already been moves towards adopting more realistic language models within corpus-enabled learning environments. By raising awareness of some of the differences existing between the discourses produced at the centre and margins of the system, Corpus Linguistics can make a useful contribution to work currently being pursued in fields such as Critical Discourse Analysis, Contrastive Rhetoric and Ethnomethodology, where issues of value and power take centre stage. Corpus tools may also be used by EAP teachers in the preparation of didactic materials and by learners who wish to orient their own progress autonomously. Hopefully, this will not only empower those on the periphery that wish to make their voice heard, but also encourage the conservatives at the centre to question the basic premises upon which the whole concept of Western knowledge is based.

References

Bennett, K. 2006. 'Critical Language Study and Translation: The Case of Academic Discourse'. In *Translation Studies at the Interface of Disciplines,* J.F. Duarte, A.A.Rosa & T. Seruya (Eds.). Amsterdam & Philadelphia: John Benjamins. 111-127.

Bennett, K. 2007a, 'Galileo's Revenge: Ways of Construing Knowledge and Translation Strategies in the Era of Globalization'. In *Social Semiotics, Vol. 17, No. 2,* M. Salaama-Carr (Ed.), Abington: Taylor & Francis, 171-193

Bennett, K. 2007b, 'Epistemicide! The Tale of a Predatory Discourse'. In *The Translator, Vol. 13, No. 2,* Manchester: St Jerome. 151-169

Biber, D., Johansson, S., Leech, G., Conrad, S. and **Finegan, E.** 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Fairclough, N. 1994. Conversationalization of public discourse and the authority of the consumer. In R. Keat, N. Whitely, and N. Abercrombie (eds.) *The authority of the consumer.* London: Routledge.

Fairclough, N. 1997. Critical discourse analysis. In T. A. van Dijk (ed.) *Discourse studies : a multidisciplinary approach. Vol 2 Discourse as social action.* London: Sage. (258-284).

Halliday, M.A.K. 1998**.** 'Things and Relations: Regrammaticising Experience as Technical Knowledge'. In *Reading Science: Critical and Functional Perspectives on Discourses of Science,* Jim R. Martin. & Robert Veel (Eds).London & New York: Routledge.

Halliday, M.A.K. and **Hasan, R.** 1976. *Cohesion in English,* London & New York: Longman.

Halliday, M.A.K. and **Martin, J.R.** (Eds), 1993. *Writing Science: Literacy and Discursive Power,* Pittsburgh & London: University of Pittsburgh Press.

Kasper, G. 1992 Pragmatic transfer. *Second Language Research,* 8/3,201-31.

Martin, J.R. 1998. 'Discourses of Science: Recontextualisation, genesis, intertextuality and hegemony' in Martin, J. R. & Veel, R. (Eds.) *Reading Science: Critical and Functional Perspectives on Discourses of Science*.London & New York: Routledge.3-14.

Martin, J. R. and **Veel, R.** (Eds.) 1998. *Reading Science: Critical and Functional Perspectives on Discourses of Science.* London & New York: Routledge

Mateus, M.H.M., Brito, A.M., Duarte, I. and **Faria, I. H.** 1989. *Gramática da Língua Portuguesa.* Lisbon: Caminho.

Odlin, T. 1989 *Language Transfer: Cross-linguistic influence in Language learning*. Cambridge: Cambridge University Press.

Ooi, V. 1998. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Rayson, P. 2003. *Matrix: a statistical method and software tool for linguistic analysis through corpus comparison.* Unpublished Ph.D. thesis. Lancaster University.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Santos, B.S. 2005. 'General Introduction' to *Reinventing Social Emancipation. Toward New Manifestos*. In Santos, B.S. (Ed.), *Vol. 1. Democratizing Democracy: Beyond the Liberal Democratic Canon*. London: Verso, pp. xvii – xxxiii.

Scott, M. 1999. *Wordsmith Tools*. Oxford: Oxford University Press

Sinclair, J. 2004. *Trust the text: language, corpus and discourse*. London: Routledge

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

TEACHING COLLOCATIONS THROUGH DDL: DESIGN, IMPLEMENTATION AND PRELIMINARY RESULTS OF A CORPUS-BASED LEARNING EXPERIENCE

María Moreno Jaén[136]

Abstract

This exploratory study involves the design and implementation of a corpus-based module of lexical collocations, which is part of a virtual course devoted to the assessment and development of the lexical competence of university students in Andalusia (Spain). This learning experience bears upon corpus data not only as a necessary source of information for the selection of collocations and the design of teaching materials but also as a valuable resource for students to develop a better understanding of the nature and behaviour of collocations. Thus, after a careful process of selection of collocations where both statistical and pedagogical criteria are taken into consideration, we design a four-stage unit devoted to the teaching of these phraseological patterns. Students are first introduced to the notion of collocations and to basic aspects of DDL, a stage followed by a sequence of corpus-based materials devoted to noticing, recycling and producing collocations. After the pedagogical implementation of the module through a virtual platform (ILIAS), qualitative analyses are performed in order to obtain information concerning students' attitudes and opinions about the experience. In the light of these preliminary results, it seems that our corpus-based proposal may represent a beneficial contribution to the field of collocational development and teaching.

Keywords: Collocations, phraseology, data-driven learning, CALL, university students.

Introduction

This study reports on a teaching experience aimed at the development of the collocational competence of university students through the design and implementation of a corpus-based module of collocations[137]. Two basic assumptions underlie this innovative experience. First, and as suggested by a number of studies (Bahns 1993; Lewis 2000), collocations constitute an essential component of lexical competence, and therefore, it is an aspect which warrants special attention in the FL classroom. However, a systematic approach to their teaching still remains to be developed, a necessary task in order to meet our students' lexical needs insofar as the "depth knowledge" dimension of vocabulary clearly bears upon collocations. Our second fundamental assumption is that, as a number of authors (Philip 2007; Shin & Nation 2007; Walker 2007) have recently revealed, corpus linguistics may offer some answers not only in terms of what collocations to teach and how to teach them but also in the necessary task of equipping students with strategies they may find useful for the permanent and autonomous development of their collocational competence. Namely, the use of corpora as a source of language data can provide information about frequency and stability of lexical combinations. On the other hand, the value of corpora as a language awareness tool can also be very beneficial for exposing students to authentic language and thus helping them discover the way words co-occur in real contexts. Taking these assumptions on board, we have created a set of corpus-based teaching materials to be implemented within an online modular course (ADELEX) for Spanish university students of English. This paper endeavours to describe this learning experience in as much detail as space allows. Finally, a small scale qualitative study has allowed us to establish the validity and adequacy of the materials designed.

Design and implementation of a corpus-based module of collocations

Selection of collocations

The first aspect we addressed in our teaching programme was the selection of contents, that is, the list of the lexical collocations that we considered most relevant for our students. To this end, we compiled a bank of the most appropriate verb-noun and adjective-noun collocations. We looked into the first 400 nouns of a frequency

[136] María Moreno Jaén is a research assistant at the University of Granada. She holds an MA in English Studies (University of Granada) and is currently doing her PhD on the assessment of L2 students' knowledge of collocations at the same university. She has published a number of articles on computerised lexical analysis and second language teaching and testing. Her main areas of interest include Lexicology and Phraseology, EFL, Corpus Linguistics, Language Testing and CALL. She is a member of a Spanish R&D project entitled "ADELEX: Assessing and Developing Lexical Competence through New Technologies" (HUM2007-61766) where she has co-developed an online program for measuring lexical difficulty in texts (http://www.ugr.es/local/inped/ada).

[137] This research was conducted under the auspices of the research grant FPU (2005-1353) and within the R&D project entitled "ADELEX: Assessing and Developing Lexical Competence through New Technologies" (Ref: HUM2007-61766), both financed by the Spanish Ministry of Education and Science.

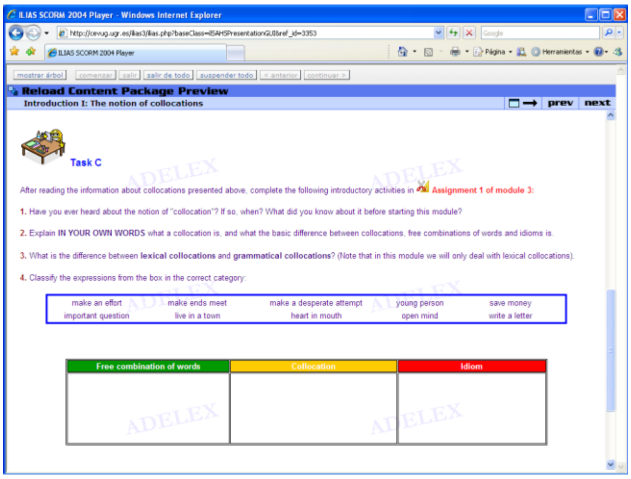
list which contains the first 7,000 words of English (López-Mezquita 2005), and subsequently retrieved their verbal and adjectival collocates. For the extraction of the most relevant collocates of the nouns previously selected, and for the purpose of rigour, the search was performed by retrieving and comparing data from both the BNC (running the program Sketch Engine[138]) and the BOE (running the LookUp software provided by Collins Cobuild[139]). The lists of collocates obtained from both programs were calculated using T-score measures and results were normalised for the purpose of comparison.

In terms of the criteria applied for our selection, we advocate an eclectic approach given that not only statistical significance but also pedagogical (and to some extent intuition-based) factors should be taken into consideration for the final selection. Hence, since corpus-based software can only perform statistical calculations, the results provided by both programs had to be manually analysed in order to identify and reject frequent combinations which, in our opinion, do not constitute interesting collocations from a pedagogical point of view. From our experience, this is the most vital aspect when performing a corpus-based selection of collocations insofar as it is absolutely crucial to differentiate between collocations which may be necessary for our students and those which, despite their high frequency, may be of little use. In our case, we considered interesting those verb-noun and adjective-noun collocations which met the following criteria: 1) they were highly frequent combinations according to corpus evidence (a frequency cut-off point of t-score = 5 was established), 2) they were used in a wide range of texts types and contexts, 3) they were semantically transparent, and 4) they were arbitrarily restricted in their commutability and/or combinability being, thus, necessary for learners to store them as single units since they are unexpected for them (Schmitt & Underwood 2004).

Design and implementation of materials

Drawing on the resulting list of collocations gathered, we designed a three-week virtual module which put a strong emphasis on data-driven learning (DDL) as an efficient approach to collocational instruction. Bearing upon Nation's (2001) suggestions for vocabulary teaching, our module was divided into four stages: 1) introduction to the notion of collocation and to corpus-based learning; 2) awareness raising and noticing of collocations; 3) practice and recycling; 4) learners' production and autonomous learning.

First of all, a number of **introductory activities** were designed in order to make students aware of the concept and importance of collocations in the process of SLA. By asking them, for instance, to translate into the L2 a number of L1 expressions which contained basic nouns accompanied by one of their most common collocates, they were prompted to think about the phraseological nature of language and the need to learn vocabulary in chunks rather than in isolation. For this purpose, once learners had read about the nature of collocations to obtain a more precise knowledge of their function, they were also asked to discriminate between collocations and free combinations on the one hand and idioms on the other (Fig. 1).


ILIAS SCORM 2004 Player - Windows Internet Explorer
http://cevug.ugr.es/ilias3/ilias.php?baseClass=ilSAHSPresentationGUI&ref_id=3353
Google
ILIAS SCORM 2004 Player
Página
Herramientas
mostrar árbol
comenzar
salir
salir de todo
suspender todo
< anterior
continuar >
prev
next

Reload Content Package Preview

Introduction I: The notion of collocations

Task C

After reading the information about collocations presented above, complete the following introductory activities in Assignment 1 of module 3:

1. Have you ever heard about the notion of "collocation"? If so, when? What did you know about it before starting this module?

2. Explain **IN YOUR OWN WORDS** what a collocation is, and what the basic difference between collocations, free combinations of words and idioms is.

3. What is the difference between **lexical collocations** and **grammatical collocations**? (Note that in this module we will only deal with lexical collocations).

4. Classify the expressions from the box in the correct category:

make an effort | make ends meet | make a desperate attempt | young person | save money
important question | live in a town | heart in mouth | open mind | write a letter

Free combination of words	Collocation	Idiom

Figure 1. Activity for classification of lexical combinations

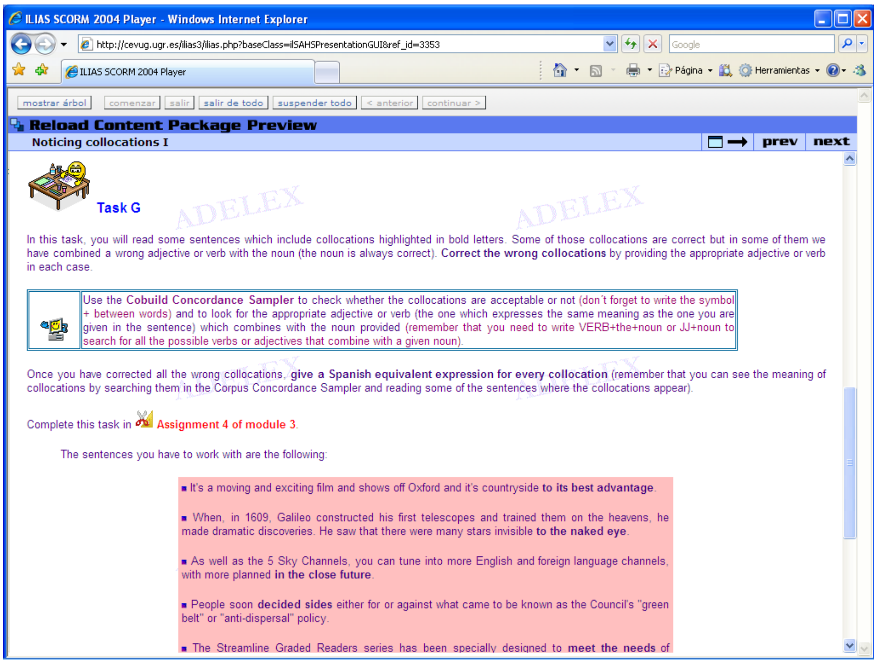
[138] http://www.sketchengine.co.uk [Last visited 10/05/2008]

[139] http://www.collins.co.uk/books.aspx?group=154 [Last visited 10/05/2008]

In this preparatory phase, learners were also introduced to corpus-based materials and techniques, as these represent the basic methodological tools used throughout the module. Therefore, after reading some basic information about what corpora and concordances are, learners were instructed in the use of Cobuild Concordance and Collocations Sampler[140]. This tool is a very useful resource for practical reasons, given that it is a free access program learners may use autonomously in the future, and also from the methodological perspective as it provides a reasonable number of concordances in a very clear KWIC display. It was, therefore, by using this tool in a number of introductory tasks that learners became familiar with corpus consultation techniques.

In the second stage we aimed at raising learners' awareness of the importance of **identifying and noticing collocations** in the input they receive, so that they could subsequently enlarge their collocational competence in an autonomous and lifelong learning way. It is widely acknowledged today that concordances are one of the most valuable resources we now have at our disposal to expose students to authentic input and eventually to explore the language inductively. This will be done in conjunction with DDL, which encourages the student to become "a research worker whose learning needs to be driven by access to linguistic data" (Johns 1991: 2). In fact, DDL is an approach particularly suitable not only to help students notice and explore linguistic patterns which are made salient by the concordancer because of their frequency and stability, but also to make them aware of the combinations which are not naturally used by native speakers.

Therefore, in order to foster learners' noticing of recurrent patterns and also in an attempt to familiarise them with some of the most frequent and useful collocations of English, we created activities which mainly consisted in direct corpus consultation designed to alert students to patterns of lexical co-occurrence. These activities involved correction of mistakes and translation of the resulting correct collocations into the L1 (Fig. 2), completion of grids where learners were required to observe the frequency of patterns and the kinds of combinations they obtained (some of the resulting combinations were free combinations and others were idioms which students were expected to identify), reformulation of lexical combinations by providing more accurate and/or elaborate adjective-noun and verb-noun collocations, etc. It should be highlighted here that, since one of our main objectives was to train students in the direct and autonomous use of online concordancers and, therefore, we could not manipulate the data they would obtain from the program, it was essential to follow a very careful design procedure where all the results the concordancer was expected to provide had to be checked in advance by the teacher in order to make sure learners would find no problems in completing the activities.


ILIAS SCORM 2004 Player - Windows Internet Explorer
http://cevug.ugr.es/ilias3/ilias.php?baseClass=ilSAHSPresentationGUI&ref_id=3353
Google
ILIAS SCORM 2004 Player
Página
Herramientas
mostrar árbol
comenzar
salir
salir de todo
suspender todo
< anterior
continuar >
Reload Content Package Preview
Noticing collocations I
prev
next
ADELEX

Task G

In this task, you will read some sentences which include collocations highlighted in bold letters. Some of those collocations are correct but in some of them we have combined a wrong adjective or verb with the noun (the noun is always correct). **Correct the wrong collocations** by providing the appropriate adjective or verb in each case.

Use the **Cobuild Concordance Sampler** to check whether the collocations are acceptable or not (don´t forget to write the symbol + between words) and to look for the appropriate adjective or verb (the one which expresses the same meaning as the one you are given in the sentence) which combines with the noun provided (remember that you need to write VERB+the+noun or JJ+noun to search for all the possible verbs or adjectives that combine with a given noun).

Once you have corrected all the wrong collocations, **give a Spanish equivalent expression for every collocation** (remember that you can see the meaning of collocations by searching them in the Corpus Concordance Sampler and reading some of the sentences where the collocations appear).

Complete this task in **Assignment 4 of module 3**.

The sentences you have to work with are the following:

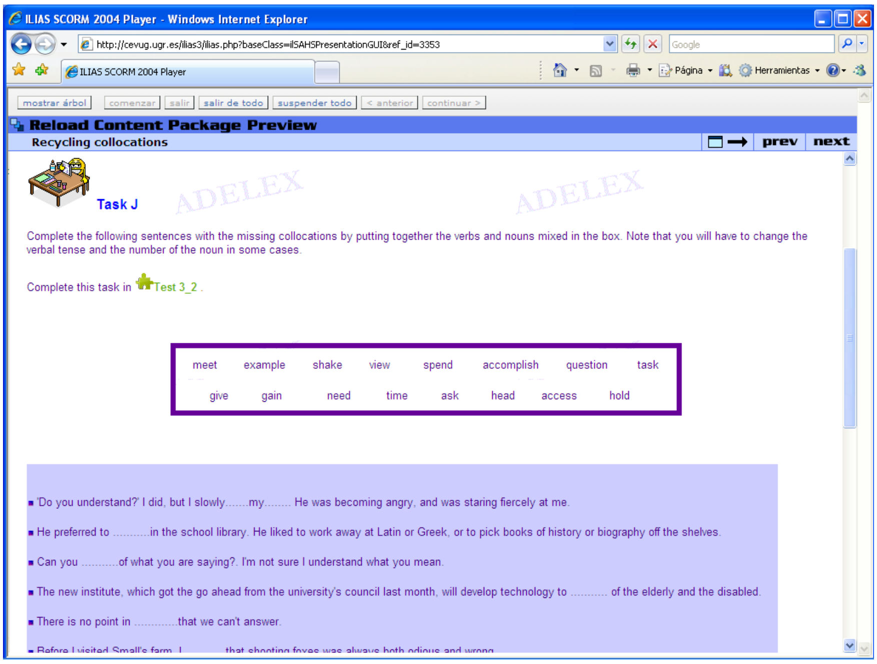
- It's a moving and exciting film and shows off Oxford and it's countryside **to its best advantage**.
- When, in 1609, Galileo constructed his first telescopes and trained them on the heavens, he made dramatic discoveries. He saw that there were many stars invisible **to the naked eye**.
- As well as the 5 Sky Channels, you can tune into more English and foreign language channels, with more planned **in the close future**.
- People soon **decided sides** either for or against what came to be known as the Council's "green belt" or "anti-dispersal" policy.
- The Streamline Graded Readers series has been specially designed to **meet the needs** of

Figure 2. Activity for correction of wrong collocations by corpus consultation

[140] http://www.collins.co.uk/Corpus/CorpusSearch.aspx [Last visited 10/05/2008]

The third stage of the module was devoted to **reviewing and** thus **recycling** the collocations learners had already encountered in the previous tasks of the module. Taking into account, on the one hand, the widely accepted hypothesis that vocabulary learning only takes place after learners have had between six and twelve encounters with new words (Jenkins & Dixon 1984), and considering, on the other hand, that we are dealing with multi-word expressions which require students to establish even more mental associations for memorisation than it is required when learning single words, it seems obvious that a stage where learners are prompted to revisit the contents already studied is a necessary step which should be present in any programme devoted to lexical development.

For the purpose of recycling collocations, we designed a number of activities, some of which required learners to combine a number of nouns with their collocates in order to fill gaps in sentences extracted from the BNC (Fig. 3).



Task J

Complete the following sentences with the missing collocations by putting together the verbs and nouns mixed in the box. Note that you will have to change the verbal tense and the number of the noun in some cases.

Complete this task in Test 3_2 .

meet example shake view spend accomplish question task
give gain need time ask head access hold

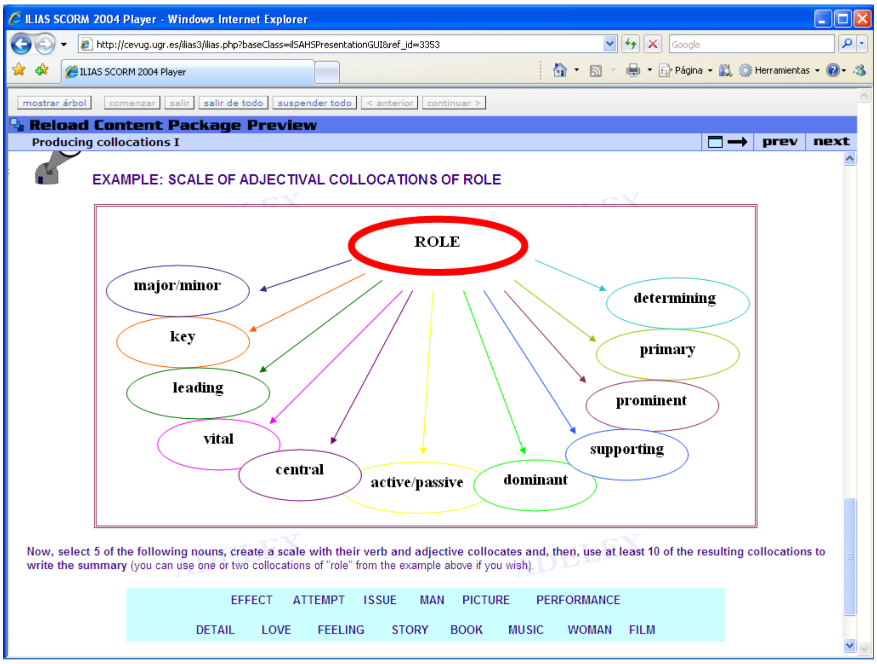
- 'Do you understand?' I did, but I slowly.......my........ He was becoming angry, and was staring fiercely at me.
- He preferred to ....in the school library. He liked to work away at Latin or Greek, or to pick books of history or biography off the shelves.
- Can you ....of what you are saying?. I'm not sure I understand what you mean.
- The new institute, which got the go ahead from the university's council last month, will develop technology to .... of the elderly and the disabled.
- There is no point in ....that we can't answer.
- Before I visited Small's farm, I ....that shooting foxes was always both odious and wrong.

Figure 3. Activity for recycling collocations

Finally, the fourth stage of the module was concerned with the **productive use of collocations**. The activities in this stage attempted to go beyond the boundaries of traditional DDL techniques which bear upon concordance observation. Here, learners were also prompted to develop corpus-based strategies which would enable them to carry out production activities successfully. This strategic competence would undoubtedly contribute to promote processes of autonomous learning. Once students reached this last stage and given that they were already acquainted with the use of concordancers and DDL techniques, it seemed advisable to introduce them to new tools for corpus exploration, in an attempt to equip them with as many resources and strategies for collocational learning as possible.

"Phrases in English" (PIE) [141] is another useful and freely available resource for furthering collocational knowledge, which enables the user to retrieve collocations containing as many words as s/he may preselect. After showing students how to enter a query in PIE and the type of information it provides in turn, they were asked to search for the most frequent collocates of a number of nouns provided in the activity and, subsequently, they had to compile scales where they needed to arrange the collocates by frequency (Fig. 4). The final part of this task was the production of a written summary where some of these collocations had to be used. In our view, this is a useful activity not only as a means to develop strategies for autonomous learning of collocations, but also to help students improve their writing skills and their productive command of the L2. Very briefly, other activities included in this section which provide learners with opportunities to use all the corpus-based resources they are familiar with are those dealing with translation of L1 into L2 sentences containing new (i.e. not studied in this module) collocations, an activity which, in our view, represents an authentic task to encourage productive use of the language (Fig. 5).

[141] http://pie.usna.edu/ [Last visited 10/05/2008]



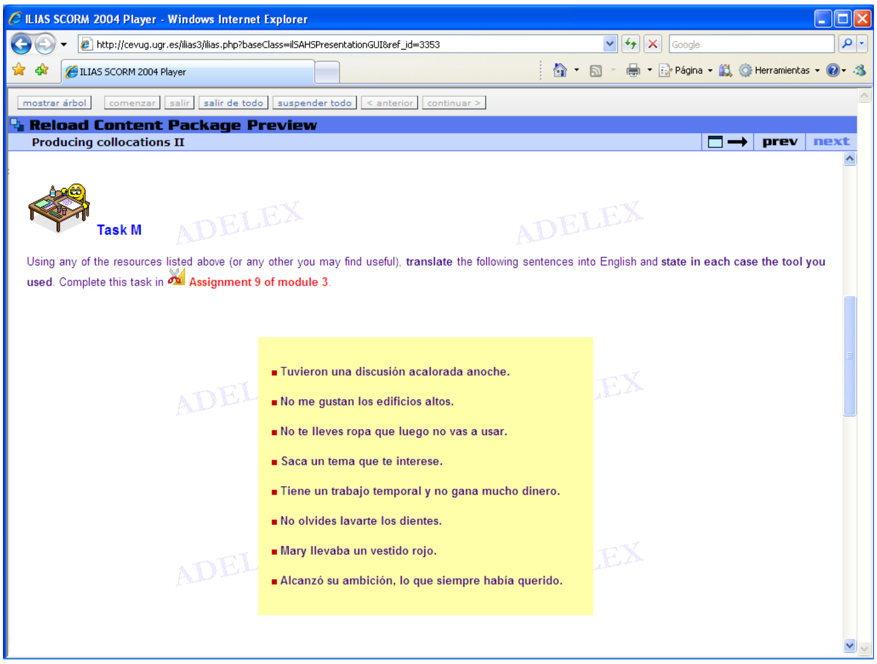
EXAMPLE: SCALE OF ADJECTIVAL COLLOCATIONS OF ROLE

Now, select 5 of the following nouns, create a scale with their verb and adjective collocates and, then, use at least 10 of the resulting collocations to write the summary (you can use one or two collocations of "role" from the example above if you wish).

EFFECT ATTEMPT ISSUE MAN PICTURE PERFORMANCE

DETAIL LOVE FEELING STORY BOOK MUSIC WOMAN FILM

Figure 4. Scale of adjectival collocations of "role" ordered by frequency


ILIAS SCORM 2004 Player - Windows Internet Explorer
http://cevug.ugr.es/ilias3/ilias.php?baseClass=ilSAHSPresentationGUI&ref_id=3353
Google
ILIAS SCORM 2004 Player
Página
Herramientas
mostrar árbol
comenzar
salir
salir de todo
suspender todo
< anterior
continuar >
Reload Content Package Preview
Producing collocations II
prev
next

Task M

Using any of the resources listed above (or any other you may find useful), **translate** the following sentences into English and **state in each case the tool you used**. Complete this task in **Assignment 9 of module 3**.

- Tuvieron una discusión acalorada anoche.
- No me gustan los edificios altos.
- No te lleves ropa que luego no vas a usar.
- Saca un tema que te interese.
- Tiene un trabajo temporal y no gana mucho dinero.
- No olvides lavarte los dientes.
- Mary llevaba un vestido rojo.
- Alcanzó su ambición, lo que siempre había querido.

Figure 5. Translation activity in the stage for productive use of collocations

Results and discussion

Once this pedagogical treatment was administered to students, we carried out a pilot study where some qualitative and quantitative statistical analyses were performed in order to evaluate the effectiveness of our experience. Due to space constraints, we will only deal in this paper with the qualitative analysis performed and, for the sake of brevity, only the results of a few items of our questionnaire will be addressed in detail.

The questionnaire was answered by 19 out of the 20 students who took this module, and it was administered immediately after the module had been completed. The questionnaire consisted of 19 items which can be divided into 3 main topics:

A) Questions 1 to 4: Gathering information about personal variables concerning learners (age, degree s/he is studying, year of studies, and total number of years learning English).

B) Questions 5, 6 and 14 to 19: Gathering information about the efficacy and usefulness of the learning module (the concept of collocations, the type and number of collocations included, the strategies for collocational learning).

C) Questions 7 to 13: Gathering information about the methodology, i.e. the role and management of concordances and corpus-based techniques for language learning (level of difficulty in the use of concordancers, efficacy of searches performed, DDL techniques as opposed to traditional deductive learning, level of autonomy provided by corpus-based resources).

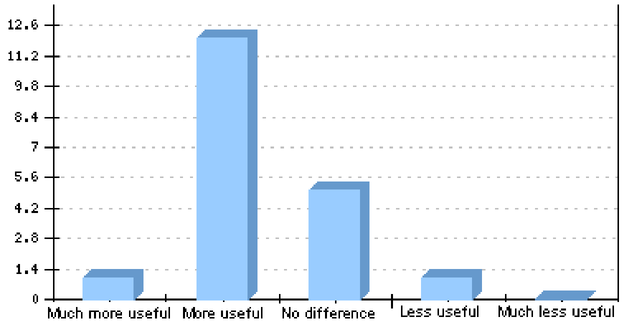
Limited space only allows a brief note on the main results obtained from our analysis, but we think they may be sufficient for the reader to see what the general opinions of our learners were at the end of the module. As regards section B, which tried to assess the efficacy and usefulness of the learning module, 100% of students considered they had obtained a clear or a very clear idea of what collocations are, 84.21% of learners considered they had learnt quite a lot of collocations, 100% found the collocations included in the module were useful or very useful for their real needs, and 78.95% considered they had learnt many or quite a lot of strategies which they could use in the future to continue learning collocations.

Section C, whose aim was the evaluation of the methodology of this module, showed that 68.42% of students found the concordancer easy or very easy to use (the remaining 31.58% said it had been difficult to use but only in the beginning). On the other hand, concordancers were regarded as adequate or very adequate tools for learning collocations by 100% of learners, 89.47% considered DDL to be more useful or much more useful than traditional deductive learning when dealing with collocations and 100% said concordancers would be useful for them on many occasions or in particular courses in the future.

Together with these issues, learners also answered two more very interesting questions which are shown below in a more detailed manner.

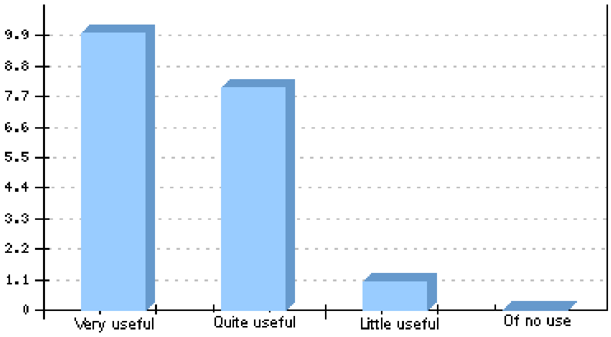
Question 12: As compared to other resources (dictionaries, textbooks, etc.), when learning collocations concordancers are:

Much more useful	More useful	No difference	Less useful	Much less useful
5.26%	63.16%	26.32%	5.26%	0%



Question 18: In general, you have found the module of collocations:

Very useful	Quite useful	Little useful	Of no use
52.63%	42.11%	5.26%	0%



In the light of the results obtained from our qualitative study, and despite the fact that they have been very succinctly portrayed here, we may observe that, in general, learners evaluated this new teaching experience very positively. Statistical results show that students expressed a high level of satisfaction with regard to both contents and methodology, being all the questions answered positively by over 65% of subjects. In effect, question 18 seems to be a very revealing indicator of the strong agreement students show concerning the usefulness and relevance of our module, since 94.74% of them (that is 18 out of the 19 surveyed subjects) regarded it as quite useful or very useful. In sum, findings from this small-scale study provide support for the conclusion that the corpus-based module proposed here offered a successful learning experience for students and, therefore, it may contribute to the field of teaching collocations in a beneficial way.

Nevertheless, a final note should be added in terms of the limitations of this evaluation. We are fully aware that for a more reliable and valid study to be performed, three considerations should be taken into account in the future. Firstly, learners should be instructed both in learning collocations and in the use of corpus-based resources for a longer period of time. Secondly, it should be piloted with a larger number of students. Thirdly, both written and oral tasks should be implemented. This exploratory study is though a first step to show the potential value of teaching collocations online.

Conclusion

By and large, the innovative research reported here represents an attempt to address the issue of teaching collocations from a more rigorous and systematic approach than it has been done in the past, something for which corpus-based techniques seem to be particularly suitable. It is also intended to make the most of corpus data both for the selection of collocations and for their pedagogical development, in a time when the advantages of corpus-based pedagogy are widely documented, but comparatively very little hands-on work with corpora has been developed in teaching scenarios (Braun 2007). Likewise, this pilot study also represents an initial attempt to delve into the qualitative nature of a corpus-based learning experience by gathering and analysing students' opinions about their learning preferences, an aspect which needs to be further developed in the field of data-driven learning.

References

Bahns, J. 1993. “Lexical collocations: a contrastive view.” *English Language Teaching Journal* 47/1: 56-63.

Braun, S. 2007. “Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora.” *ReCALL* 19/3: 307-328.

Jenkins, J. R. and **Dixon, R.** 1984. “Vocabulary learning.” *Contemporary Educational Psychology* 8: 237-260.

Johns, T. 1991. “Should you be persuaded: Two samples of data-driven learning.” In *Classroom Concordancing*, T. Johns and P. King (eds.). Birmingham University: English Language Research Journal 4, 1-13.

Lewis, M. (ed.). 2000. *Teaching Collocation. Further Developments in the Lexical Approach.* Hove: Language Teaching Publications.

López-Mezquita Molina, M. T. 2005. *La Evaluación de la Competencia Léxica: Tests de Vocabulario. Su Fiabilidad y Validez.* PhD dissertation. Electronic publication in CD-Rom. Granada: University of Granada.

Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Philip, G. 2007. "Decomposition and delexicalisation in learners' collocational (mis)behaviour." Paper presented at the *4th Corpus Linguistics Conference*, Birmingham, 27-30 July, 2007.

Schmitt, N. and **Underwood, G.** 2004. "Exploring the processing of formulaic sequences through a self-paced reading task." In *Formulaic Sequences: Acquisition, Processing and Use*, N. Schmitt (ed.). Amsterdam: John Benjamins, 173-190.

Shin, D. and **Nation, P.** 2007. "Beyond single words: The most frequent collocations in spoken English." *English Language Teaching Journal* 0/2007 (Advanced Access online).

Walker, C. 2007. "Collocation: From the corpus to the classroom." Paper presented at the *4th Corpus Linguistics Conference*, Birmingham, 27-30 July, 2007.

BAWE: AN INTRODUCTION TO A NEW RESOURCE

Hilary Nesi[142]

Abstract

The British Academic Written English (BAWE) corpus was developed with ESRC funding as part of the project entitled 'An investigation of genres of assessed writing in British Higher Education' (2004-2007). The project aimed to identify the characteristics of proficient student writing, and to compare these across disciplines and levels of study. The corpus consists of just under 3000 student assignments of a good standard (6,506,995 words), at all levels from first year undergraduate to taught masters degree, and in many disciplines. Information about discipline and level is provided in the header for each assignment file, alongside other types of contextual information which did not influence collection policy such as gender, year of birth, native speaker status, and years of UK secondary education. We believe that BAWE is currently the only complete corpus of its kind in the public domain. It offers opportunities to investigate student writing which has been judged to conform to departmental requirements, but which differs markedly from expert and near-expert academic writing in terms of its communicative intent.

Keywords: academic, assignment, essay, EAP, genre

Background to the project

The project 'An investigation of genres of assessed writing in British Higher Education' grew out of a concern that too little was known about the types of writing students produced in British universities, and a concern that inappropriate genre models were used for academic writing courses.

The research article is as popular a genre for analysis today (e.g. Ozturk, 2007; Bruce, 2008) as in the 1980s (e.g. Swales 1983, 1984). The discourse of doctoral theses has also been investigated fairly thoroughly (e.g. Thompson, 2005; Charles, 2006). This focus on published articles and theses is understandable, since they represent the standard many academic writers aspire to, and they are readily available in the public domain. Nevertheless they do not represent the bulk of what is written in academic contexts, i.e. the texts produced by students on taught degree programmes, for assessment, generally with the intention of demonstrating academic knowledge and skills as opposed to presenting research findings.

Of course the university assignment is not an entirely neglected genre, and there have been a number of excellent studies of small collections of student writing, usually within jusr one or two disciplines and with reference to one particular discourse feature (see, for example, Woodward-Kron, 2002; North, 2005). Before the development of the BAWE corpus, however, no fully documented collection existed which might enable large scale comparisons of assignments across disciplines and levels of study. Two such corpora are under development in the United States (the Michigan Corpus of Upper-level Student Papers (MICUSP), and the 'Viking' corpus at Portland State University), but at the time of writing both of these contain less than a million words.

Our initial attempt to create a small corpus of student assignments was not entirely successful, and provided some insight into why such a corpus did not yet exist. Our pilot project ran from May 2001 to November 2002, during which time we collected 499 assignments from 70 student writers. The contributors, however, tended to come from a limited range of disciplines (largely from the humanities, with very few from the hard sciences) and there was a disproportionate number of assignments from the first year of study (44%) (see Nesi, Sharpling and Ganobcsik-Williams, 2004). The project did not adopt any particular collection policy, and simply accepted any assignment offered by any willing student. This helps to explain why the hard sciences and the later years of study were not well represented, as fewer scientists were interested in contributing, they produced less written work, and there was diminishing availability of assignments in the upper levels (students could contribute work written in

[142] Hilary Nesi joined Coventry University as Professor in English Language in October 2007, having worked for twenty years in the Centre for English Language Teacher Education at the Unversity of Warwick. She led the project to create the BASE corpus of British Academic Spoken English (2001–2005) and also the ESRC funded project 'An Investigation of Genres of Assessed Writing in British Higher Education' (2004-2007), which involved the creation of the BAWE corpus. She is chief academic consultant for the Essential Academic Skills in English (EASE) series of multimedia interactive self-access materials on CD-ROM, based around video clips of authentic academic discourse drawn from the BASE corpus, and she is currently involved in several projects relating to the use of English as a medium of instruction in international higher education.

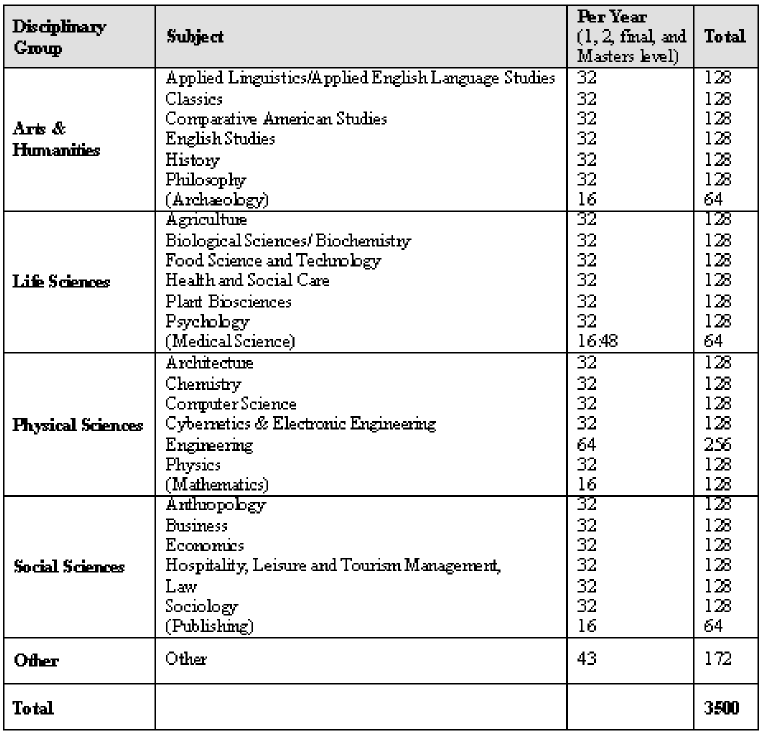
preceding years, but could not contribute work that had not yet been assessed). It was evident that it would be necessary to devise a more systematic approach to data collection to fulfil the aims of the main project, which received funding from the ESRC in 2004..

For this project we proposed to integrate ethnographic, multidimensional and functional linguistic approaches to text description, each of which suggested a different method of sampling (as discussed in Gardner, forthcoming). Ethnographic aspects of the study favoured cluster sampling and the targetting of specific university discourse communities, but random sampling seemed an appropriately objective way of collecting data for computational analysis, and purposive sampling, involving the targetting of specific text types, promised to provide the richest array of data for genre analysis.

Our final collection policy involved stratified sampling, a compromise which took into account these conflicting approaches to corpus analysis, together with the practical constraints on policy implementation. We did conduct interviews with staff and students (see Nesi and Gardner, 2006; Gardner and Powell, 2006), but we rejected the idea of sampling selected clusters of contributors because we did not have the resources (or the persuasive power) to guarantee contributions from sufficient numbers of individuals within specified departmental communities. We considered random sampling, but even if it had been possible to identify a random sample of potential student contributors, our experience with the pilot corpus had taught us that it would be impossible to force contributions from them. We abandoned more purposive sampling, although we wanted to gather several instances of each assignment type we encountered, because it soon became clear that it would be impossible to create a multi-million word corpus if we set restrictions on the genre of contributions, as well as on their grade, discipline and year of study.

Corpus holdings

We used a 4-by-4 matrix to guide data collection. This combined four years of study with four broad disciplinary groupings, and we intended to fill each of the 16 cells with a roughly equal quantity of assignments, rejecting all but a few contributions which were superfluous to these requirements (we retained an ‘other’ category, to round up numbers). The following table represents our ideal corpus structure in more detail, and our plan to collect 3,500 assignments across 28 disciplinary fields.



Disciplinary Group	Subject	Per Year (1, 2, final, and Masters level)	Total
Arts & Humanities	Applied Linguistics/Applied English Language Studies Classics Comparative American Studies English Studies History Philosophy (Archaeology)	32 32 32 32 32 32 16	128 128 128 128 128 128 64
Life Sciences	Agriculture Biological Sciences/ Biochemistry Food Science and Technology Health and Social Care Plant Biosciences Psychology (Medical Science)	32 32 32 32 32 32 16:48	128 128 128 128 128 128 64
Physical Sciences	Architecture Chemistry Computer Science Cybernetics & Electronic Engineering Engineering Physics (Mathematics)	32 32 32 32 64 32 16	128 128 128 128 256 128 128
Social Sciences	Anthropology Business Economics Hospitality, Leisure and Tourism Management, Law Sociology (Publishing)	32 32 32 32 32 32 16	128 128 128 128 128 128 64
Other	Other	43	172
Total			3500

Table One: the plan for BAWE corpus collection.

Our matrix was not designed to represent proportionally the quantity of writing produced in each discipline and at each level, or to ensure perfect representation of all the genres produced in the target disciplines. Students usually write more in their final year(s), and some disciplines are understood to be more discursive than others (as indicated in British university rules concerning PhD thesis length – usually a maximum of 80,000 words in the Humanities and Social Sciences, but only 50,000 words in the Sciences). Also we knew we could not collect assignments for every module in every discipline, and that module tutors were liable at any time to introduce new tasks with different generic expectations. We realized we might miss some unusual genres, especially if only a few students selected a particular writing task, or if they received low grades (we only accepted assignments graded 60% or above). Nevertheless steps were taken to encourage variety in the corpus in terms of both assignment type and authorship, by prompting contributors to submit additional work belonging to a different genre, if possible, whilst preventing individuals from contributing more than three assignments from any single module.

Assignments were collected at Oxford Brookes, Reading and Warwick, and, in the final year of the project, Coventry University (to make up numbers in disciplines which still lacked sufficient contributions). Most cells of our matrix were not quite filled, as can be seen from Table Two.

Disciplinary Grouping		**Yr 1**	**Yr 2**	**Yr 3**	**Masters**	**Total**
Arts and Humanities	students	101	83	61	23	268
	assignments	239	228	160	78	705
	texts	254	232	160	82	728
	words	468,353	583,617	427,942	234,206	1,714,118
Life Sciences	students	74	71	42	46	233
	assignments	180	193	113	197	683
	texts	186	203	92	246	727
	words	299,370	408,070	263,668	441,283	1,412,391
Physical Sciences	students	73	60	56	36	225
	assignments	181	149	156	110	596
	texts	201	156	159	121	637
	words	300,989	314,331	426,431	339,605	1,381,356
Social Sciences	students	85	88	75	62	313[1]
	assignments	207	197	162	202	777[2]
	texts	215	205	165	210	804[3]
	words	371,473	475,668	440,674	688,921	1,999,130[4]
Total students		**333**	**302**	**234**	**167**	**1039**[1]
Total assignments		**807**	**767**	**591**	**6587**	**2761**[2]
Total texts		**856**	**796**	**576**	**659**	**2896**[3]
Total words		**1,440,185**	**1,781,686**	**1,558,715**	**1,704,015**	**6,506,995**[4]

[1] Includes 3 students of unknown level.

[2] Includes 9 assignments of unknown level.

[3] Includes 9 texts of unknown level.

[4]. Includes 22,394 words of unknown level

The number of texts recorded in the table exceeds the number of assignments, because some assignments turned out to consist of more than one independent text, submitted together to receive a single grade.

Table Three provides a more complete picture of the disciplines represented in the corpus. In this table 'discipline' is not synonymous with 'department', because some assignments in the same field came from more than one university, and departments with slightly different names have been conflated (*Computer Science* and *Computing*, for example). We recognize that 'discipline' is a difficult concept to define, however, and that 'variation in epistemology and discourse occurs not only across disciplines, but also within disciplines' (Nesi and Gardner, 2006: 101).

Disciplinary Grouping	**Discipline**	**1**	**2**	**3**	**4**	**Total**
Arts and Humanities	Archaeology	23	21	15	17	76
	Classics	33	27	15	7	82
	Comparative American Studies	29	26	13	6	74
	English	35	35	28	8	106
	History	30	32	31	3	96
	Linguistics	27	31	24	33	115
	Other	19	22	9	0	50
	Philosophy	43	34	25	4	106
	Total	**239**	**228**	**160**	**78**	**705**
Life Sciences	Agriculture	35	35	30	34	134
	Biological Sciences	52	50	26	41	169
	Food Sciences	26	36	32	30	124
	Health	35	33	12	1	81
	Medicine	0	0	0	80	80
	Psychology	32	39	13	11	95
	Total	180	193	113	197	683
	Total	**180**	**193**	**82**	**228**	**683**
Physical Sciences	Architecture	2	4	2	1	9
	Chemistry	23	24	29	13	89
	Computer Science	34	13	30	10	87
	Cybernetics & Electronics	4	4	13	7	28
	Engineering	59	71	54	54	238
	Mathematics	8	5	12	8	33
	Meteorology	6	9	0	14	29
	Other	0	1	0	0	1
	Physics	37	14	14	3	68
	Planning	8	4	2	0	14
	Total	181	149	156	110	596
	Total	**181**	**149**	**155**	**111**	**596**

Social Sciences	Anthropology	14	12	6	17	49
	Business	32	33	31	50	146
	Economics	30	30	23	13	96
	HLTM	14	21	29	29	93
	Law	37	37	31	28	134*
	Other	0	2	3	4	9
	Politics	37	33	15	25	110
	Publishing	11	4	0	15	30
	Sociology	32	25	24	21	110†
	Total	**207**	**197**	**162**	**202**	**777‡**
Total		**807**	**767**	**591**	**587**	**2761‡**

* Includes 1 of unknown year.
† Includes 8 of unknown year.
‡ Includes 9 of unknown year.

Table Three: number of assignments by discipline and year

The corpus was encoded according to the guidelines of TEI P4 *(*Sperberg-McQueen and Burnard, 2004), but since the TEI standard was devised for a wide range of texts, a special DTD containing only a subset of all TEI elements and attributes was created for BAWE (see Heuboeck, Holmes and Nesi, 2008). Information of the following types was encoded:

- header information
- document structure and hierarchy
- types of front and back matter
- functional features within running text
- character formatting
- anonymized personal information (related to student, university or third parties)

The header provides information about the discipline and level of each assignment, alongside other types of contextual information which did not influence collection policy, such as gender, first language and years of UK secondary education. However although we recorded the gender of each contributor, gender proportions vary so much from cell to cell that corpus-wide comparisons of assignments written by male and female wrters are potentially unreliable. For similar reasons we do not recommend comparisons of first language groups: the proportions vary across disciplines, and there are far more non-native speakers at Masters level. In any case it is dangerous to make assumptions about a contributor's English language proficiency or genre knowledge on the basis of their mother tongue; in British university contexts a contributor's choice of first language sometimes reflects affiliation rather than proficiency, and prior apprenticeship within the British education system (recorded in terms of the number of years of UK secondary schooling) is a factor to be considered independently of mother tongue.

Findings

The following broad 'genre families' were identified in the corpus:

Case Study: A description of a particular case with recommendations or suggestions for future action, written to gain an understanding of professional practice (e.g. in business, medicine, or engineering).

Critique: A text including a descriptive account, explanation, and evaluation, often involving tests, written to to demonstrate understanding of the object of study and to demonstrate the ability to evaluate and / or assess the significance of the object of study.

Design Specification: A text typically including an expression of purpose, an account of component selection, and a proposal; and possibly including an account of the development and testing of the design.

Empathy writing: A letter, newspaper article or similar non-academic genre, written to demonstrate understanding and appreciation of the relevance of academic ideas by translating them into a non-academic register, for a non-specialist readership.

Essay: A discussion, exposition, factorial, challenge or commentary, written to develop the ability to construct a coherent argument and develop critical thinking skills.

Exercise: Data analysis or a series of responses to questions, written to provide practice in key skills and to consolidate knowledge of key concepts.

Explanation: A descriptive account and explanation, written to demonstrate understanding of the object of study and the ability to describe and/or assess its significance.

Literature Survey: A summary including varying degrees of critical evaluation, written to demonstrate familiarity with the literature relevant to the focus of study.

Methodology Recount: A description of procedures undertaken by the writer, possibly including Introduction, Methods, Results, and Discussion sections, written to develop familiarity with disciplinary procedures and methods, and additionally to record experimental findings.

Narrative Recount: A fictional or factual recount of events, written to develop awareness of motives and/or the behaviour of organisations or individuals (including oneself).

Problem question: A text presenting relevant arguments or possible solution(s) to a problem, written to practise the application of specific methods in response to simulated professional scenarios.

Proposal: A text including an expression of purpose, a detailed plan, and persuasive argumentation, written to demonstrate the ability to make a case for future action.

Research Report: A text typically including a Literature Review, Methods, Findings, and Discussion, or several 'chapters' relating to the same theme, written to demonstrate the ability to undertake a complete piece of research, including research design, and to appreciate its significance in the field.

One obvious conclusion that can be drawn from this categorisation scheme is that university students write for a range of purposes, not all of them identical to the purposes of academics. Some assignments are generically similar to texts produced in the professions, but only the Research Report bears much generic resemblance to the thesis or research article.

The distribution of the genre families in the corpus is presented in Table Four. The essay is the best represented category, although in the Physical and Life Sciences it is outnumbered by submissions belonging to other genre families (Methodology Recounts, Design Specifications, and Critiques). Also, some genre families are rare or totally absent from some disciplinary groupings, particularly the Arts and Humanities.

	Arts and Humanities	**Life Sciences**	**Physical Sciences**	**Social Sciences**	**Total**
Case Study	0	91	37	66	194
Critique	48	84	76	114	322
Design Specification	1	2	87	3	93
Empathy Writing	4	19	9	3	35
Essay	602	127	65	444	1238
Exercise	14	33	49	18	114
Explanation	9	117	65	23	214
Literature Survey	7	14	4	10	35
Methodology Recount	18	158	170	16	362
Narrative Recount	10	25	21	19	75
Problem Question	0	2	6	32	40
Proposal	2	26	19	29	76
Research Report	9	22	16	14	61
Total	724	720	624	791	2859

Table Four: Distribution of genre families by disciplinary group

Multidimensional analysis revealed the corpus to be carefully written and information-rich, but there were also significant differences among genre families, as can be seen from Table Five. The entirely negative scores on the 'involved' and 'narrative' dimensions indicate a high informational focus and a low level of narration, whilst the entirely positive scores for 'explicit' and 'abstract' qualities indicate lexically dense text containing passives, past participial clauses, and other features typical of academic prose. Mixed scores on the 'persuasive' dimension, however, indicate variation in the degree of argumentation (Proposals being the most persuasive, and Literature Surveys the least). Student writing simply does not need to 'create a research space' in the manner of research article introductions, because the centrality of the topic is not usually in question, and the tutor is duty-bound to read the text.

	Involved	**Narrative**	**Explicit**	**Abstract**	**Persuasive**
Essay	-14.327	-2.4788	6.234	5.920	-1.8345
Methodology Recount	-15.856	-3.6533	4.506	7.304	-2.5011
Critique	-14.833	-3.0714	5.988	6.381	-1.6127
Explanation	-15.411	-3.5878	5.042	5.848	-2.2744
Case Study	-16.402	-2.8617	5.772	4.450	-0.4519
Exercise	-12.098	-3.8543	4.628	5.678	-1.3301
Design Specification	-13.090	-4.0223	4.079	6.750	0.6702
Proposal	-16.421	-3.7855	6.326	4.793	1.2799
Narrative Recount	-4.818	-1.1128	3.814	3.957	-0.7439
Research Report	-16.186	-3.1156	5.524	7.198	-2.4064
Problem Question	-11.950	-2.7730	5.222	6.429	1.6295
Literature Survey	-17.907	-2.6214	6.311	5.047	-3.4343
Empathy Writing	-11.500	-2.7369	4.533	4.472	0.7713

Table Five: Multiple Range Test Scores for Genre Families

Multidimensional analysis also revealed significant differences between the four disciplinary groupings in terms of their information load, and significant differences between first and final year undergraduate assignments on all but the 'persuasive' dimension.

Conclusion

Clearly the BAWE corpus is a very rich resource, offering a currently unique opportunity to investigate thousands of academic texts which have been judged to conform to departmental requirements (on the evidence of the grade awarded), but which differ markedly from professional academic writing in terms of their communicative intent. Several close analyses of the corpus are planned or in press, and proposals for further investigations will be welcomed by the research team.

Acknowledgements

The project *An Investigation of Genres of Assessed Writing in British Higher Education*, including the development of the British Academic Written English corpus, was funded by the Economic and Social Research Council (project number RES-000-23-0800), under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for English Language Teacher Education, Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes). Other members of the project team were Sian Alsop, Dawn Hindle, Maria Leedham, Signe Ebeling, Alois Heuboeck, Jasper Holmes, Richard Forsyth and Laura Powell. The multidimensional analysis was conducted by Doug Biber and his team at Northern Arizona University.

References

Bruce, I. 2008. “Cognitive genre structures in Methods sections of research articles: A corpus study” *Journal of English for Academic Purposes* 7/1: 38-54

Charles, M. 2006. “Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines” *English for Specific Purposes* 25/3: 310-331

Gardner, S. Forthcoming. “Integrating ethnographic, multidimensional, corpus linguistic and systemic functional approaches to genre description: an illustration through university history and engineering assignments”. *Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop,* Universität des Saarlandes, Saarbrücken, July 2007.

Gardner, S. and Powell, L. 2006. ‘An investigation of genres of assessed writing in British Higher Education’. Paper presented at the annual seminar *Research, Scholarship and Practice in the area of Academic Literacies*, University of Westminster, 30 June 2006.

http://www.coventry.ac.uk/researchnet/external/content/1/c4/33/84/v1193312407/user/genresbhe_handout.pdf [Access date 27/05/2008].

Heuboeck, A., Holmes, J. and Nesi, H. 2008 *The BAWE Corpus Manual*. http://www.coventry.ac.uk/researchnet/external/content/1/c4/51/60/v1212053950/user/BAWE.pdf [Access date 27/05/2008].

Nesi, H. and Gardner, S. 2006. “Variation in disciplinary culture: University tutors’ views on assessed writing tasks”. In R. Kiely, P. Rea-Dickins, H. Woodfield and G. Clibbon (eds.), *Language, Culture and Identity in Applied Linguistics*, British Studies in Applied Linguistics Vol. 21. London: Equinox Publishing, 99-117.

Nesi, H., Gardner, S., Forsyth, R., Hindle, D., Wickens, P., Ebeling, S., Leedham, M., Thompson, P. and Heuboeck, A. 2005. “Towards the compilation of a corpus of assessed student writing: an account of work in progress” *Proceedings from the Corpus Linguistics Conference* Series 1/1. www.corpus.bham.ac.uk/PCLC/NesiStudentWriting.doc [Access date 27/05/2008].

Nesi, H, Sharpling, G. and **Ganobcsik-Williams, L.** 2004. “The design, development and purpose of a corpus of British student writing” *Computers and Composition* 21/4: 439-450.

North, S. 2005. “Different values, different skills? A comparison of essay writing by students from arts and science backgrounds” *Studies in Higher Education* 30/5: 517-533

Ozturk, I. 2007. “The textual organisation of research article introductions in applied linguistics: Variability within a single discipline” *English for Specific Purposes* 26/1: 25-38.

Sperberg-McQueen, C. M. and Burnard, L. (eds.). 2004. *TEI P4 – Guidelines for Electronic Text Encoding and Interchange, XML-compatible edition.* http://www.tei-c.org/P4X/ [Access date 27/05/2008].

Swales, J. M. 1983. “Developing materials for writing scholarly introductions”. In *Case Studies in ELT*, R. R. Jordan (ed.). London: Collins ELT.

Swales, J. M. 1984. “Research into the structure of introductions to journal articles and its application to the teaching of academic writing”. In *Common Ground: shared interests in ESP and communication studies*, R. Williams, J. Swales & J. Kirkman (eds.) Oxford: Pergamon Press.

Thompson, P. 2005. **“**Points of focus and position: Intertextual reference in PhD theses” *Journal of English for Academic Purposes* 4/4: 307-323

Woodward-Kron, R. 2002. **“**Critical analysis versus description? Examining the relationship in successful student writing” *Journal of English for Academic Purposes* 1/2: 121-143

CREATING AN ORAL CORPUS FOR TEACHING PURPOSES – WITH STUDENTS OF SPANISH AS AUTHORS

Carlota Nicolás Martínez[199]

Abstract

The aim of this paper is to demonstrate and evaluate the design and the creation of a Corpus of Oral Language for Teaching Purposes carried out by advanced students of Spanish. In this on-going project at the Faculty of Letters, the students are entirely responsible for the creation of materials. The instructor has two crucial duties: to report on the state of the art in oral corpora to allow participants to make decisions about the structural peculiarities o

f the new corpus and to teach them how to produce transcription from oral sources.

As an introduction to the design of the corpus the instructor illustrates the contents and structure of Spanish corpora. Following an examination of the didactic potential of the corpus, students provide a heading for each transcribed text in view of applications in Spanish L2. Students must also learn the complex art of transcribing from oral sources, which involves several didactically significant features. In short, students encounter four aspects of the language study: corpus linguistics, information linguistics, prosody and the analysis of spoken language.

On completion of this project students undergo a transformation in their sensitivity to listening to spontaneous speech. In particular, they become more aware of features of speech such as minimal segments, like interjections and vocalizations, and larger segments, like turns and sequences. Although statistical data are not yet available, it appears that the most frequent errors in transcription center around lexical groups serving to structure information.

Keywords: spoken corpora, Spanish L2, transcription markup, prosodic patterns, error analysis, spoken language studies

The aim of this paper is to analyze a corpus of oral language for teaching purposes created by students of Spanish at various levels of knowledge of the language. This is an on-going project at the Facoltà di Lettere e Filosofia of the Università di Firenze in which students are entirely responsible for the creation of the corpus. In the beginning, the professor has two crucial duties, that is, to report on the state of the art in spoken corpora, which allows students to make decisions on the design and contents of the new corpus, and to teach them how to produce a transcription from an oral source.

After the students have completed the transcriptions there is be a second phase, which, however, will only be mentioned in this paper, and which focuses entirely on the result that is, the study of spoken language. During this second phase the teacher invites students to focus on a variety of bigger or smaller issues according to the students' proficiency, and above all, according to the theoretical approach considered most suitable. For instance, some studies have been carried out on lexical items having a discoursive function, or on more general aspects, such as the signalling of *Comment* and *Topic* and other kinds of dialogical aids in utterances (within the framework of E. Cresti's theory.

There are many reasons why I have addressed these issues in several university courses. The present work follows the path of that carried out by other work in this field. My study focuses on spoken language. There is no need to demonstrate that every person who learns an L2 has very close contact with this diamesic(al) [???] variety. As a consequence, I believe that students should feel seriously interested in the extent to which they need to approach

[199] Carlota Nicolás Martínez is a researcher in Spanish at the Facoltà di Lettere e Filosofia of the Università di Firenze. She was former lecturer at the Università degli Studi di Pisa, and foreign expert at Universities of Beijing and Tianjin. She was awarded the title of European PhD from the Universidad Autónoma de Madrid. Her fields of interest are bilingual lexicography, language teaching and spoken language. She has published an Italian-Spanish dictionary, Dizionario Italiano-spagnolo, Spagnolo-italiano, and several articles: "El diccionario bilingüe en la enseñanza del léxico"; "Una propuesta de utilización de corpus orales en el enseñanza de segundas lenguas". She has also edited Atti delle Giornate di Studi su Strumenti Didattici Università degli Studi di Firenze (13–15 maggio 2002); Atti del Convegno Internazionale CLiP 2003 and Ricerche sul Corpus del parlato romanzo C-ORAL-ROM. Studi linguistici e applicazioni didattiche per l'insegnamento di L2 (2008). She was a researcher member of the European Project C-ORAL-ROM. She is also a member of the Scientific Committee of the International Seminar Computers, Literature and Philology CLiP and a member of the Editorial Board of the Lingue d'Europa. Strumenti Didattici series published by Alinea.

the study of this variety. On the one hand, this study requires the application of theories such as the knowledge of notions of prosodic patterns (on a transcription-basis). On other, it involves a rather systematic investigation which requires a great deal of practical work - what may be called "handcrafted" work - which brings students into direct contact with the spoken language.

In order to present corpora, teachers have to develop a three-step process: firstly, the analysis of the global design of other existing corpora; then, their headers with meta-data; finally, their transcription markup.

The present study will focus on the work the professor carries out with intermediate students, as it is more suitable for the kind of examination considered here.

To provide an introduction to the design of the corpus, the professor illustrates the contents and structure of both simple and complex Spanish corpora, such as CORDE and C-ORAL-ROM. These are the only spoken corpora which are completely available and complex in design. They both permit users to access all transcribed texts and they also include an extensive introduction. C-ORAL-ROM offers transcriptions, audio in alignment, a concordance programme and many statistical tables, and it also is the major reference point for the work we are interested in, since the markup which will be used is, with minor variations, that of C-ORAL-ROM. However, the transcription which we carry out is mainly prosodical and orthographic. These two aspects are useful when learning Spanish. As far as orthography is concerned, it is very important to follow the rules. Students must avoid transcribing what they perceive as features of spoken language. They must be able to convert what they hear into orthographically correct structures.

Every attempt at comprehending spoken language the learner has to deal with is of great didactic value. The learner's ability to interpret texts and his/her learning how to represent them in prosodic patterns are also relevant.

Students who have been working on this so far belong to three different levels. Each course has received a different amount of information on the subject according to different levels of students. This is particularly true as regards the definition of corpora, the transcription of spoken texts and the filling of headers. Different teaching methods corresponds to a different attitudes toward the teaching process itself. Questions arise such as how we can approach a work in order to gain results which would correspond to students' knowledge of Spanish or to their general background on language.

At present, there has been no chance to have the same work repeated. Nevertheless, I have already been experimenting with different courses according to students’ levels. In the future I will try to develop this aspect as a necessary step in the training of students. In fact, students’ acquiring of an awareness of their own improvement in the learning process and commitment to it is an extremely relevant aspect to consider. This is also due to the fact that we believe this will help them develop their own ability in learning a language on their own (autonomy of learning) outside of a university environment.

As far as the didactics is concerned, it may be pointed out that a high degree of commitment from students is extremely important. Students' commitment is stimulated by the awareness that they are participating in a co-operative task which is part of a broader, rather ambitious project. Indeed, the goal of the project is to become a part of the heritage of the Facoltà di Lettere e Filosofia. At the same time, another important aspect of students' commitment is the fact that their effort is perceived as something which will help future students of Spanish in many different ways.

Nonetheless, it is hard to let students access this work from the beginning. The first problem occurs with the seminar nature of the lessons examined here. Italian universities do not commonly develop seminar-based lessons. Moreover, Spanish courses are very often crowded, which makes it harder to develop a successful teacher-learner relationship. In order to underline the co-operative nature of this work, we have acknowledged the importance of a rigid definition of every phase of the process, so as to create a secure environment for the learning process which helps students not to feel lost in the process itself. In addition, individual work requires that students do the best they can, putting all their knowledge into their work, using thier language skills actively. Teachers, then, should try in every way to get the best of them/from them, making sure they are really concentrating on the work they are doing.

The three main phases of the process are:

After the professor has handed the recordings of spoken spontaneous language that previously were made in Spain back to students according to their levels, lab work begins to introduce students to a program for speech analysis, *Cool Edit*. *Cool Edit* renders the listening phase of the transcription process easier. *Word* editing tools will also be used, as they are necessary to track the transcription process carried out by the student.

A rigid calendar of work is established. Several meetings are needed to revise the two or three versions of the transcription. This is something which is done by the professor and the student together. At advanced levels, the student who has done the transcription also meets his/her student supervisor.

During the first meeting, the professor points out errors and checks students' knowledge of the basic concepts. Then, the teacher decides how to go on with the work. During the second revision, the professor tends to focus on a deeper analysis of more concrete aspects of spoken texts and also helps students to decipher some difficult fragments of the texts themselves.

The final draft of the work, after it has been double-checked by the professor and, in higher level courses, by a student supervisor, is the subject of an oral examination, whose purpose is to discuss the difficulties encountered by the student while working. The student also answers specific questions on his/her choice with the introduction of transcription marks. The errors in listening are also checked, which the student might have made due to his/her inability to identify words or strings of words. These errors are examined by the student and the professor together, paying particular attention to the position or the nature of words missing from his/her own transcription and thus re-directing his/her attention towards certain features of spoken language.

There are several Spanish corpora available. ; there transcriptions are preceded by more or less the same header, and also the two corpora mentioned above contain headers with a great deal of information. All the varieties in the header have been considered in order to consider which could best suit our work. Thus, it has been quite interesting to analyze the data gathered in other corpora from different perspectives. In particular, the analysis of the header reveals the objectives targeted by each corpus. The degree of objectivity in data collection and the people who gathered the data have also been considered. Of course, the new data collected by the different groups vary according to their level. Only those objective data such as the length of the audio, number of words in the transcription, and those connected with the title are the same for all groups. It should be pointed out, however, that even students from the lowest levels fill the header with the data which are considered useful for the teaching process and which characterize our corpus. Enriching all the headers with much more useful data for teaching demonstrates to students that a header is more than a simple data container. It is thanks to the header that students become aware of the existence of statistical work and of automatic extraction of information work and, in deciding those data to include, that they acknowledge their value.

Students must also learn how to transcribe oral sources, hard as it may seem. Transcription rules are the same for all students and all marks are used with the same value. The very act of transcribing involves several features:

- Students recognize the value of transcription markers and become aware that a text transcribed from an oral corpus is an informational object taken out of a whole from which information can be extracted. Consequently, the transcription markers must respect certain scientific conventions.

- Learners become sensitive to the prosodic traits of the oral text, and these traits must be preserved in the transcription.

Specific features of spoken language are identified. Particularly, students become more aware of those features of speech such as minimal segments (interjections and vocalizations) as well as larger segments (turns and sequences) or/and aspects of a proper discoursive strategy (such as false-starts, reformulations, and overlappings).

Furthermore, through an observation of the transcriptions of other corpora, advanced students may observe that these vary according to specific linguistic theories . Once we have understood this, students may be invited to ponder the extent to which transcription marks chosen for the corpus they have elaborated are successful.

The accomplishment of a transcription from a recording and of the compilation of a header should bring students in contact with four different aspects of language learning: corpus linguistics, information linguistics, prosody and the analysis of spoken language.

I consider the present study as a work-in-progress. Future improvements could take into consideration the following:

- The use of a platform which would make poszsible an interaction between professor and students, and among students.

- The use of an audio analyzer for the alignment of text and voice.

- The use of standard textual markup which would render this corpus more suitable for the scientific community.

There are two aspects which are worthy of consideration both for improving teaching features and analysing corpora. Firstly, student errors, as they help to focus on those aspects which require more accurate learning. Although statistical data are not available yet, it seems that the most frequent errors in transcription are influenced by lexical groups serving to structure information. The errors also offer a possibility to improve the quality and choice of transcription marks and of meta-data of the header. The present work has tried to underline this by showing some of the problems which may occur while working. Secondly, a survey carried out among students might demonstrate that, while completing this project, they/students change their attitudes towards listening to spontaneous speech.

The work considered in this study has a general feature, that is, its ability to progress and to determine someone’s progress incessantly.

ORAL LEARNER CORPORA AND ASSESSMENT OF SPEAKING SKILLS

John Osborne[200]

Abstract

The aim is this paper is to examine how empirical findings from learner corpora can help to inform practical assessment of foreign language speaking skills within the Common European Framework of Reference. The present study is based on a corpus of spoken productions by learners of English (various L1s) performing comparable tasks, collected and assessed as part of a European project in collaborative assessment of oral language proficiency. Each recording is independently rated on the CEF scales, and is transcribed and analysed for fluency features such as speech rate, pauses, retracing and length of utterances.

The main research questions are (1) what degree of convergence there is between different measures of fluency and (2) to what extent these are reflected in the raters' perception of oral language proficiency. Results so far indicate that there is individual subject variation in specific features of spoken fluency, so that any single measure taken in isolation – say, the number of pauses – is an unreliable indication of proficiency, but that if fluency is measured as a bundle of features, then it becomes more meaningful to relate proficiency bands to objective fluency measures. Although these measures are too time-consuming to be easily implemented in everyday language assessment, they can be used to benchmark representative samples illustrating CEF levels.

Keywords: Corpora, oral proficiency, fluency, assessment, Common European Framework

Oral production and the CEFR

The descriptors used in the Common European Framework (Council of Europe 2001) combine quantitative criteria - how many things the second language learner can do in the target language - and qualitative criteria – how well he/she can do them. These qualitative aspects of language use are described through carefully formulated "can do" statements covering, in the case of spoken language use, such things as accuracy, fluency and coherence. However, these qualitative scales are not themselves directly quantifiable, and it is left to the expertise and experience of the evaluator to determine, individually or collectively, whether a given L2 learner is able to express him/herself "spontaneously at length with a natural colloquial flow" (CEF level C2) or "fluently and spontaneously, almost effortlessly" (CEF level C1). In practical terms, this may not be of major importance, if the objective of language assessment is to reflect how a learner's speaking skills will be perceived by a potential listener. Nevertheless, as Hulstijn (2007) observes, it is a somewhat shaky ground on which to construct a European-wide scale of language proficiency.

Specifically, the CEFR descriptors for spoken fluency include the following formulations:

- natural, effortless, unhesitating flow (C2)
- can express him/herself fluently and spontaneously, almost effortlessly (C1)
- often showing remarkable fluency and ease of expression (B2+)
- fairly even tempo / few noticeably long pauses (B2)
- some problems with formulation resulting in pauses and 'cul-de-sacs' (B1+)
- pausing for grammatical and lexical planning and repair is very evident (B1)
- pauses, false starts and reformulation are very evident (A2+)
- very noticeable hesitation and false starts (A2)
- much pausing to search for expressions, to articulate less familiar words, and to repair communication (A1)

[200] John Osborne is a Professor of English language and linguistics at the university of Savoie, Chambéry. His principal research interests are in linguistics and second language learning, particularly the sources of persistent errors in more advanced learner language, and of disfluency in spoken production. He is currently involved in two corpus-based projects: the "Scientext" corpus (a corpus and tools to carry out a linguistic study of authorial position and reasoning in scientific texts), in collaboration with the universities of Grenoble and Lorient (France), and the PAROLE corpus of spoken learner language, described in this paper.

The interpretation and application of these descriptors to specific samples of L2 production pose three kinds of problem: (1) it is debatable how many native speakers could maintain a "natural, effortless, unhesitating flow", particularly in the context of an oral examination; (2) the distinctions between neighbouring levels often rely on downtoners and semantic niceties ("effortless" vs. "almost effortlessly"; "very evident" vs. "very noticeable"), making it difficult to apply them systematically; (3) the descriptors identify a number of disfluency phenomena – pauses, false starts and reformulation – which may be more or less "evident" up to B1 level, but imply that these will be absent, or at least not "noticeable" at B2 level and above. There is thus a need for more information about the extent to which these phenomena are present in oral production at all levels, including native-speaker production. North (2007) remarks that the CEFR fluency descriptors are in fact based on exploratory research, since they are much inspired by Fulcher (1996). However, the 200-300 word descriptors proposed for each band in Fulcher's fluency rating scale only appear in a much condensed form in the 20-30 word CEFR scales. In addition, Fulcher's descriptors are themselves already based on an interpretive coding system, which codes the data for such things as "pauses which appear to allow the student to plan the content of the next utterance" (Fulcher 1996: 216).

Oral production in L2 and L1

The data used for this study come from two related sources: a European project in collaborative assessment of oral language proficiency (*WebCEF*), and a parallel corpus of oral learner language (the *PAROLE* corpus). These two sources are briefly described below.

The WebCEF project

The aim of this project, financed under the Socrates-Minerva programme, is to provide web-based tools for evaluating oral language skills with reference to the Common European Framework. The project has two main components: a "showcase" of selected samples that have been jointly assessed and annotated by a team of language teachers and assessors from various European countries, and a "community of practice" which provides a workspace for language teachers throughout Europe to upload samples of oral production, to assess and annotate them, and to compare their assessments with those of other European colleagues. The tasks on which the oral samples are based are also available on the workspace, along with any documents used during the task, so that users can re-use existing tasks to record samples produced in similar conditions. Although the central objective of the project is to provide tools for collaborative evaluation of speaking skills, it will also result in a steadily growing database of oral learner language in English, French, Italian, Dutch, German, Finnish, Polish and other languages. More information about *WebCEF* is available on the project website: http://www.webcef.eu

The PAROLE corpus

Some of the initial tasks used in the *WebCEF* project were first used to collect data for the *PAROLE* corpus (Parallèle, Oral, en Langue Etrangère) at the University of Savoie. This corpus consists of 15-20 minute recordings of speakers of L2 English (L1 French and German), of L2 French (various L1s) and of L2 Italian (L1 French), along with recordings from native speakers of these three languages carrying out the same tasks. The recordings are transcribed in CHAT format (MacWhinney 2000) and annotated for pauses (filled and unfilled), retracings and errors, with a view to comparing (dis)fluency characteristics across languages, between native speakers and non-natives, and between non-native speakers at different levels of proficiency. For a fuller description of the corpus, see Osborne (2007), Osborne & Rutigliano (2007). The main purpose of the *PAROLE* project is to investigate the sources of disfluency in second language production, but this necessarily involves identifying and measuring the characteristics of fluent and disfluent speech.

The convergence between these two projects allows for previously assessed *WebCEF* samples to be analysed for their fluency features, and for previously analysed samples from the *PAROLE* corpus to be collectively assessed on the CEF scales, thus providing for each sample a set of quantitative measures and a measure of perceived fluency, as rated by a group of experienced assessors. Comparison of these should then enable us to investigate (1) what degree of convergence there is between different measures of fluency and (2) to what extent these are reflected in the raters' perception of oral language proficiency.

Measures of oral fluency

The principal problem in evaluating spoken L2 fluency is teasing apart individual, task-related and developmental factors. Hesitation features are frequent in native speech, and native speakers vary considerably in the extent to which they display such features. The additional demands of oral production in an L2 may be expected to introduce extra sources of disfluency, but there is no reason to suppose that individual differences related to a person's speaking style will disappear when the person happens to be speaking another language. Longitudinal studies (Dechert 1980; Towell *et al.* 1996; Freed *et al.* 2004*)* provide useful evidence of how individual and developmental factors may be related, but what is needed for the present purposes is a measure, or set of measures, that will reflect an individual speaker's fluency at a given time.

Temporal measures of fluency

A range of temporal measures have been used in previous studies of spoken production, including primary variables such as speech rate, articulation rate, length of runs or length of silent pauses, and secondary variables such as filled pauses, drawls, repetitions and false starts (Grosjean 1980; Raupach 1980). Those used in the present study are:

1. Speech rate: measured in words per minute or syllables per minute. Since the ratio of syllables to words varies between languages, it is important to take this into account in cross-language comparisons.
2. Pauses: measured by the average length of pauses (filled and unfilled) and by the percentage of pause time in relation to the total time used by the speaker.
3. Length of utterances: measured by mean length of utterances (MLU) and by the mean length of fluent runs in between pauses.
4. Retracing: measured by the number of retracings per 100 words, either without modification (simple repetition) or with repetition (self-correction).

From these measures, a fairly constant picture emerges: whatever single measure is taken, there is dispersion among speakers of the same group (native speakers, more advanced non-native speakers, less advanced non-native speakers), and there is overlap between the groups. Table 1 below, showing the percentage of hesitation time in relation to total speaking time, provides an illustration of this phenomenon. In each of the native-speaker groups, there is a considerable spread between the least hesitant speaker and the most hesitant, and the most hesitant native speaker hesitates more than the least hesitant of the non-native speakers for the same language. This pattern is repeated for all of the other temporal measures. The greatest variations among native speakers (up to 4:1) are observed for retracings, percentage of pause time and mean length of runs; the smallest variations (less than 2:1) for words per minute, average length of pauses and mean length of utterances. There is no measure for which there is no overlap between native and non-native speakers (i.e. there is always a non-native speaker whose measure exceeds that of the least fluent native speaker of the same language).

	all speakers	**most hesitant**	**least hesitant**
L1 English	31.9	36.1	12.9
L1 French	23.1	34.7	15.9
L1 Italian	17.2	27.0	6.8
L2 English (NNS1)	29.6	34.5	22.3
L2 French (NNS1)	30.5	33.2	27.4
L2 Italian (NNS1)	34.5	46.0	24.6
L2 English (NNS 2)	59.0	67.5	51.9
L2 French (NNS 2)	43.6	49.5	38.0
L2 Italian (NNS 2)	51.1	63.4	34.3

Table 1: percentage of hesitation time in relation to total speaking time

Does this mean that temporal fluency is so much an idiosyncratic characteristic of speakers that it is fruitless trying to use it as a measure of L2 proficiency? If an overall fluency measure is calculated, by taking six specific temporal measures for each speaker (words per minute, average length of pauses, percentage of pause time, MLU, mean length of runs and number of retracings) and averaging them out, then the differentiation between groups appears much more clearly, and there is no longer any overlap between native and non-native speakers. In other words, temporal fluency appears to be a real characteristic of more proficient speakers, but it is a bundle of features and cannot be reliably assessed by any single measure.

Qualitative aspects of fluency

Impressions of fluency may not derive only from temporal characteristics of speech, but also from more qualitative factors. These are incorporated into the descriptors for some, but not all of the CEF levels: "pauses only to reflect on precisely the right words" (C2); "fluency and ease of expression in even longer complex stretches of speech" (B2+); "very short, isolated, mainly pre-packaged utterances" (A1).

Fillmore (2000: 51) describes this kind of fluency as "the ability to talk in coherent, reasoned and 'semantically dense' sentences". Its quantification requires a measure of how a speaker packages content into his/her production, and to obtain this it is useful to distinguish between three types of units:

- Utterances: typically an utterance is made up of an independent clause and all its dependent clauses (i.e. a T-unit). Run-on coordinate clauses are coded as a single utterance if the conjunction is not preceded or followed by a pause of more than 300 ms. Isolated clause fragments are counted as utterances.
- Syntactic units: the total number of clauses plus the number of adjunct phrases. A clause is defined following Berman & Slobin (1994: 660) as "any unit that contains a unified predicate".
- Information units: this a measure of the amount of information encoded by the speaker. Three types of unit are distinguished: macro-statements, micro-events and circumstances/attributes. Thus, irrespective of their syntactic differences, the two following examples would be coded as having the same informational content: (1) *There is a boy. He is in a zoo. He is eating a chocolate. I think it is his last chocolate.* (2) *There is a boy in a zoo, eating his last chocolate.*

Combining these three types of unit, it is possible to measure syntactic complexity (syntactic units/utterance) and semantic density (information units/utterance). In addition to syntactic and semantic density, perceptions of fluency may be affected by what Kormos & Dénes (2004) call "high-order" fluency features, such as accuracy and lexical diversity, noting that "accuracy is positively related to temporal variables that are influential in fluency judgements". Altogether, we thus have four types of qualitative fluency measures:

1. Propositional content: measured by the number of information units per minute, per 100 words or per utterance. An additional measure of "granularity" (Noyau et al. 2005) indicates how much detail a speaker provides, by looking at the number of micro-events in relation to the number of macro-statements.
2. Syntactic density: measured by the number of syntactic units per 100 words. The number of syntactic units per utterance can also be calculated, to give a measure of "condensation"
3. Vocabulary range: measured by Vocd (Malvern & Richards 1997) and by the proportion of words used which fall outside the first 2000 word frequency band.
4. Accuracy: measured by the rate of errors (errors per minute or per 100 words).

These measures show the same general pattern as the purely temporal measures described above; there is individual variation within each group, and overlap between the groups. The greatest overlap is found for granularity, which not only shows considerable variation within groups, but also within an individual speaker, who may choose to present two consecutive macro-events with very different degrees of granularity. Condensation (the number of syntactic units per utterance) appears to increase in the earlier stages of language learning and then level off. The clearest differentiation between groups is found, not surprisingly, for rate of errors, but even here there is a small overlap, with the most error-free non-native production showing fewer performance errors than some native speakers. As is the case for the temporal measures, some non-native speakers may perform better than native speakers on specific features, but the overlaps disappear when all features are taken into consideration.

Measurement and perception of fluency

A number of previous studies (Lennon 1990, Freed 2000, Kormos & Dénes 2004, Mizera 2006) have investigated relationships between perceived fluency, as indicated by native-speaker judgements, and various temporal or other quantifiable measures of fluency. Two studies are of particular interest, in that they explicitly compare fluency measures with the ratings given by expert assessors on oral proficiency tests. Fulcher's (1996) study, already mentioned above, related learners' disfluency phenomena to the bands they achieved on the ELTS oral test. A more recent study, by Iwashita *et al*. (2008), analyses data collected during piloting of the speaking section of the TOEFL iBT test. The recordings were measured for seven spoken language features, grouped into three categories: linguistic resources (grammatical accuracy and complexity, vocabulary), phonology (pronunciation, intonation and rhythm) and fluency, measured by filled pauses, unfilled pauses, repair, total pausing time, speech rate and mean length of runs. Of these six measures of fluency, similar to those used in the present study, three – speech rate, number of unfilled pauses and total pause time – were found to be clearly related to proficiency level, while three – filled pauses, repair and mean length of runs – were not.

Our results so far indicate a similar pattern; independent CEF ratings correlate best with speech rate (wpm) and percentage of pause time, and least well with retracing and granularity. For any individual speaker, a single measure taken in isolation is not necessarily a reliable indication of proficiency, so that overall fluency is best measured as a group of features.

Benchmarking oral samples

The measures described above are time-consuming to carry out. They are therefore not a practical option for day-to-day assessment of oral production, unless they can be automatised. A degree of automatisation is technically possible, for example in the detection of pauses, but usable tools for doing this are not presently available to language assessors. The value of these measures thus lies more in their potential use for benchmarking samples of L2 oral production, to provide empirically based evidence for fluency features at different CEF levels. There are two ways of doing this. One is to take a comprehensive range of measures, excluding only those which are shown to be more related to individual variation than to proficiency level, and average them out to obtain an overall measurement of fluency. The second way, potentially more economical, would be to select one or two measures that are shown to be particularly good predictors of perceived fluency. A measurement of information density, such as the number of information units per minute, might be a good candidate for this, since it combines a temporal measure with a quantification of propositional content.

References

Berman, R. and Slobin, D. 1994. *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale NJ: Lawrence Erlbaum.

Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Dechert, H. 1980. Pauses and intonation as indicators of verbal planning in second language speech productions: Two examples from a case study. In *Temporal Variables in Speech,* H. Dechert and M. Raupach (eds.). The Hague: Mouton, 271-285.

Fillmore, C. 2000. On fluency. In *Perspectives on Fluency*, H. Riggenbach (ed.). Ann Arbor: University of Michigan Press, 43-60. First published in *Individual Differences in Language Ability and Language Behavior*, C. Fillmore *et al*. (eds.) 1979. New York: Academic Press 85-101.

Freed, B. 2000. Is fluency in the eyes (and ears) of the beholder? In *Perspectives on Fluency*, H. Riggenbach (ed.). Ann Arbor: University of Michigan Press, 243-265.

Freed, B., Segalowitz, N. and Dewey, D. 2004. Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition* 26/2, 275-301.

Fulcher, G. 1996. Does thick description lead to smart tests? A data-based approach to rating scales construction. *Language Testing* 13, 208-238.

Grosjean, F. 1980. Linguistic structures and performance structures: Studies in pause distribution. In *Temporal Variables in Speech,* H. Dechert and M. Raupach (eds.). The Hague: Mouton, 91-106.

Hulstijn, J. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal* 91/4, 663-667.

Iwashita, N., Brown, A., McNamara, T., and O'Hagan, S. 2008. Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29/1, 24-49.

Kormos, J., and **Dénes, M.** 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System* 32, 145-164.

Lennon, P. 1990. Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40/3, 387-417.

MacWhinney B. 2000. *The CHILDES project: tools for analyzing talk*. Mahwaw NJ: Lawrence Erlbaum.

Malvern, D. and Richards, B. 1997. A new measure of lexical diversity. In *Evolving models of language*, A. Ryan & A. Wray (eds.). Clevedon: Multilingual Matters, 58-71.

Mizera, J. 2006. *Working Memory and L2 Oral Fluency*. Unpublished PhD dissertation, University of Pittsburgh http://etd.library.pitt.edu/ETD/available/etd-04262006-124945/unrestricted/Mizera.Dissertation.pdf [Access date 13/02/2008]

North, B. 2007. The CEFR illustrative descriptor scales. *Modern Language Journal* 91/4, 656-659.

Noyau, C., de Lorenzo, C., Kihlstedt, M., Paprocka, U., Sanz Espinar, G., and Schneider, R. 2005. Two dimensions of the representation of complex events structures : granularity and condensation. Towards a typology of textual production in L1 and L2. In *The Structure of Learner Varieties*, H. Hendriks (ed.). Berlin: Mouton de Gruyter, 157-201.

Osborne, J. 2007. Investigating L2 fluency through oral learner corpora. In *Spoken Corpora in Applied Linguistics*, M.C. Campoy & M.J. Luzón (eds.). Frankfurt: Peter Lang, 181-197.

Osborne, J. and Rutigliano, S. 2007. Constitution d'un corpus multilingue d'apprenants d'une L2: recueil et exploitation des données. In *Acquisition et didactique, Actes de l'atelier didactique, AFLS 2005,* H. Hilton (ed.). Chambéry : LLS, Collection Langages, 141-156.

Raupach, M. 1980. Temporal variables in first and second language speech production. In *Temporal Variables in Speech*, H. Dechert and M. Raupach (eds.). The Hague: Mouton, 263-270.

Towell, R., Hawkins, R. and N. Bazergui, N. 1996 The development of fluency in advanced learners of French. *Applied Linguistics* 17, 84-19.

AN ONLINE SYSTEM FOR ERROR IDENTIFICATION IN BRAZILIAN LEARNER ENGLISH

Tony Berber Sardinha[201]
Tania M G Shepherd[202]

Abstract

The purpose of this paper is to introduce a proposal for an online system for the identification of apprentice error in written English as a foreign language. It is argued that errors are, to a large extent, systematic and, to a certain extent, predictable. This is especially so, when the population of apprentices being investigated is homogeneous, the case of the data in question, which stems entirely from argumentative essays on a limited number of topics written by English Language undergraduates at Brazilian universities, all speakers of Brazilian Portuguese. The errors/inadequacies in each of the argumentative essays were manually tagged by one of the researchers, who matched doubtful cases against the BNC. A simple tag set based on Nicholls (1999) was adopted, including, < to signal the beginning of errors; vertical bars were used to separate the error itself from a possible adequate alternative for the same error. Where acceptable alternatives could not be provided, questions marks were included. It was hoped that this simplified annotation might mirror the procedures normally adopted when a non-native EFL teacher is correcting written work. The annotated data was first fed into a pre-processor, which then extracted various probabilities, namely, of each word being used erroneously; of 3-word bundles (Biber & Conrad, 1999) preceding and following a word being used erroneously; and of 3-word collocational frameworks (Renouf & Sinclair, 1991) occurring around each word being used erroneously and, finally, the probability of each part of speech being used erroneously. On the basis of this information, an application is being developed, which takes either a single learner composition or an entire learner corpus and outputs a list of each word followed by its probability of having been used erroneously. It is hoped that this tool may be helpful to both teachers and students as a means of detecting possible errors in learner writing.

Keywords: EFL writing, error, error annotation, automated identification, probability

Introduction and justification

The purpose of the present paper is to discuss certain theoretical principles and practical steps behind a proposal of an online system for the identification of apprentice error in written English as a foreign language. To this end, the initial stage of the paper provides a brief overview of error and error correction within Applied Linguistics. The research focus subsequently concentrates on to existing forms of error annotation of digitalized corpora. The methods section describes the corpus and the procedures for error annotation for the present research. Finally, the paper includes a brief discussion of several of the implications of the findings.

In the history of methods and approaches for the teaching of foreign languages, error correction has oscillated from being an important asset for both the practitioner and the apprentice, to being totally dispensable (Ferris, 2004: 3). Depending on the Applied Linguistic creed adopted, the ‘what’, ‘when’ and ‘how much’ to correct have been, and continue to be, a topic of debate. Specifically in terms of teaching and learning of writing in EFL (English as Foreign Language), errors have also received different treatment dictated by the methodology in fashion at any one time. For example, with the advent of the Communicative Approach, the focus of writing instruction shifted from the final written product to the various stages involved in the process of writing (Hyland, 2002:11-13). Error correction during the more orthodox phases of this era suffered what Ferris (2004: 4) has defined as ‘benign negligence’.

[201] Tony Berber Sardinha is an Associate Professor of the Applied Linguistics program and a researcher at the Institute for Research in Reading at the Catholic University in São Paulo, Brazil. Tony completed his Ph.D. at the University of Liverpool and took post-doctoral studies at Northern Arizona University. He has published and supervised on a wide range of research issues relating to the interface of foreign language teaching and corpus linguistics, including major seminal works on Corpus Linguistics and Portug uese in Brazil. He coordinates both the Br-ICLE project and the Bank of Brazilian Portuguese Language.

[202] Tania Shepherd is an Adjunct Professor in Applied Linguistics and English Language at the State University of Rio de Janeiro, Brazil. Tania completed her PhD in English Language at the University of Birmingham and carried out post-doctoral studies at the Catholic University of São Paulo. Her research interests and publications include the fields of TEFL, teacher development, l earner corpora and discourse analysis.

However, there is empirical evidence that signaling an error is beneficial for apprentices because it allows them to reflect over the possibilities of correction. In their 1992 research, Green and Hetch decided to verify whether pointing out an error to learners would benefit them in any way. The results of their research indicated that a large number of learners succeeded in correcting possible errors, if their errors were pointed out to them. This number increased considerably if learners could actually explain why the error was mad in the first place. In the late nineties, Lightbown (1998: 180) argued that to direct the learners’ attention to erroneous aspects of their production is frequently beneficial and thus ought to be done.

In terms of the ways in which one can signal an error, there is not always a great deal of convergence among researchers. If each and every error is corrected, learners may feel that they are being transformed into nothing more than simple copiers of the right formulas produced by their teachers (Knoblauch e Brannon, 1984:118). On the other hand, too few corrections may mean that learners are not being given the right (and much needed) support for the language work (Reid, 1994). Correction and praise for the ´right’ attempts at producing language are welcome; however students seem to prefer direct corrective feedback from which they might be able to improve their texts (Hyland and Hyland, 2001).

The truth is that whatever the pedagogic trend, someone (either teacher or peers) will tend to make some form of comment on pieces of writing by L2 learners. Learners seem to expect this, and the majority pay attention to the comments or corrections as they try to use them in the revisions of their original draft (Ferris, 2003:122).

The continued interest in the controversial areas of error and error correction may be verified in the papers of an entire issue of the *Journal of Second Language Writing*, published in 1999. This edition was entirely devoted to the discussion of the efficacy or otherwise of error correction in the teaching of L2 or FL. In addition, already at the beginning of the 21st century, there are entire sections in Kroll (2003), as well as in Hyland and Hyland (2006) which tackle yet again the problem of error and possible ways of correcting. In terms of the interface between corpus linguistics and error correction, the topic seems to be alive and kicking as was recently seen in the workshop for automated error correction presented at 2008 CALICO (*The Computer-Assisted Language Instruction Consortium*)..

Having problematized the need to focus on error, and specifically on written error, it is necessary to move on to a working definition of error.

Error: how to recognize and signal

Marking written errors made by separate individuals is a tiring and rather imprecise activity. Different annotators will select different items as erroneous. This occurs because thus far a consensus has not been achieved as to what error is. Error in L2 or FL has been described as “a linguistic form, ... which, in the same context would in all likelihood not be produced by the learner's native speaker counterparts.’ (Lennon, 1991: 182). Thus, the decision regarding what may not be produced erroneously seems to vary from annotator to annotator. The situation becomes even more complicated because certain annotator prioritize one error over the other.

There are at least three trends in error spotting and signaling. An annotator may correct each and every error. This occurs to prevent bad language habits (Higgs and Clifford, 1982; Lalande, 1982). Alternatively, as a result of focusing on process writing, the annotator fails to correct any error (Corder, 1981; Krashen, 1984; Selinker, 1992; Truscott, 1996). In this latter case the positive results seem to be short-lived. The third trend is to correct only patterned errors, because collective patterned errors seem to be able to be improved (Bates, et al., 1993; Ferris, 1995).

While analyzing specifically those individual error made in written work, Ferris (2004-70) suggests that teacher comments on written work might be conveyed in a number of ways. In a hypothetical structure like ‘...they could go anywhere they **want**.’, in which the focus of correction is the form *want*, teachers’ annotation could come in the form of direct annotation (~~want~~. Wanted), in the form of error marking **(want),** of error coding (**want** VT = verb tense) or in the shape of an overall comment at the end of the text, which could call the student’s attention to the need to revise his|her verb tenses.

In terms of collective errors, or errors being made by a population of learners, corpus linguists including Daigneaux *et al*. (1998), de Cock (1998) e Granger (1998 *et seq*.), have also identified frequency errors, which they have called *overuse* (excessive use of a particular lexicogrammatical form) and *underuse* (insufficient use of the same). In addition, the same corpus linguists have called certain pragmatic errors as *infelicities*.

In this paper, it has been established that errors are ‘systematic use (lexico-grammatical and pragmatic) which, in a similar context (similar communicative situation) would in all likelihood not be produced by the learner's native speaker counterparts.’ Yhus, for the present research the focus has been on collective misuse. In other words, it will concentrate on patterned errors.

There are a number of well-reputed systems for automated (or semi-automated) identification of error in L2, namely the Louvain system, a hierarchical lexico-grammatical system based on a finite tag set (Dagneaux et al, 1998;) and there are systems which evaluate units which are bigger than the individual word, i.e. apprentice errors are at the level of multiword units (Granger 2003, Scott and Tribble, 2006). Biggert et al. (2004) also developed a system which consisted of manually constructed detection rules and statistical differences between correct and incorrect texts. The system was able to identify errors in word order and split compounds in Swedish. Our system, which will also include a pedagogical interface for use by both learners and practitioners, is based on the probability of a word or group of words being used erroneously. It is explained albeit briefly in the section below.

Methodology

As Ellis (1994: 18) has claimed, errors are, to a large extent, systematic and, to a certain extent predictable. This paper argues that systematicity and predictability are more obvious, when the population of apprentices being investigated is homogeneous, the case of the data in question, which stems entirely from speakers of Brazilian Portuguese, English Language undergraduates at Brazilian universities, writing argumentative essays on a limited number of topics. The essays, taken from Br-ICLE, the Brazilian sub-corpus of the International Corpus of Learner English (www2.lael.pucsp.br/corpora/bricle), totaling nearly 100 thousand words, were written either without any time constraints or under timed exam conditions.

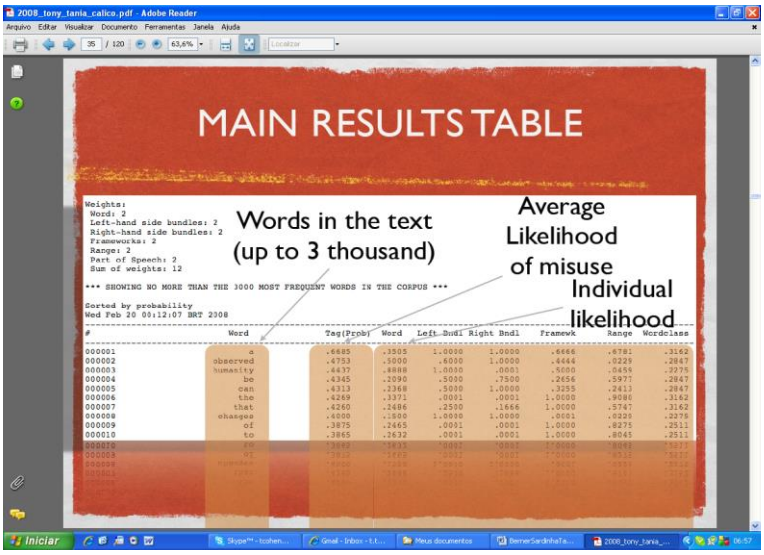
The errors/inadequacies in each of the argumentative essays were manually tagged by one of the researchers, who matched doubtful cases against the BNC. A simple tag set based on Nicholls (1999) was used, i.e., < were used to mark the beginning of errors; vertical bars were used to separate the error itself from a possible adequate alternative for the same error. Where acceptable alternatives could not be provided, questions marks were included (< error | suggestion, if possible >. In this way the annotation might be said to mirror the procedures normally adopted when a non-native EFL teacher is correcting written work.

As an example the following is an excerpt from one of the students’ texts:
‘This perspective may be concerning when people has not access to a level of education which provides them condition to develop a critical and independent thought’. This was hand-coded as
‘This perspective < may be concerning | ? > when people < has not | have no > access to a level of education which < provides them condition | provides them with the means > to develop < a | > critical and independent < thought | thinking >’, which read as *‘This perspective ? when people have no access to a level of education which provides them with the means to develop critical and independent thinking ’*

This hand-coded data formed the training corpus for our error detection tool. The data was first fed into a pre-processor, which then extracted the following information from the corpus:

- The probability of each word being used erroneously;
- The probability of a 3-word bundle (Biber & Conrad, 1999) preceding a word being used erroneously;
- The probability of a 3-word bundle following a word being used erroneously;
- The probability of 3-word collocational frameworks (Renouf & Sinclair, 1991) occurring around each word being used erroneously;
- The probability of each part of speech being used erroneously.

On the basis of this information, an application was developed, which takes either a single learner composition or an entire learner corpus and outputs a list of each word followed by its probability of having been used erroneously, as can be seen on the table below. The table includes different columns, including the average likelihood of misuse, which is calculated by attributing equal weights to left-hand and right-hand side bundles, collocation frameworks occurring around each erroneous use of word and finally part of speech used wrongly.



```
Weights:
 Word: 2
 Left-hand side bundles: 2
 Right-hand side bundles: 2
 Frameworks: 2
 Range: 2
 Part of Speech: 2
 Sum of weights: 12

*** SHOWING NO MORE THAN THE 3000 MOST FREQUENT WORDS IN THE CORPUS ***

Sorted by probability
Wed Feb 20 00:12:07 BRT 2008
----------------------------------------------------------------------------------------------------
#                        Word        Tag(Prob)  Word   Left Bndl Right Bndl   Framewk     Range  Wordclass
----------------------------------------------------------------------------------------------------
000001                      a        .6685     .3505    1.0000    1.0000      .6666     .6781     .3162
000002               observed        .4753     .5000     .6000    1.0000      .4444     .0229     .2847
000003               humanity        .4437     .8888    1.0000     .0001      .5000     .0459     .2275
000004                     be        .4345     .2090     .5000     .7500      .2656     .5977     .2847
000005                    can        .4313     .2368     .5000    1.0000      .3255     .2413     .2847
000006                    the        .4269     .3371     .0001     .0001     1.0000     .9080     .3162
000007                   that        .4260     .2486     .2500     .1666     1.0000     .5747     .3162
000008                changes        .4000     .1500    1.0000    1.0000      .0001     .0229     .2275
000009                     of        .3875     .2465     .0001     .0001     1.0000     .8275     .2511
000010                     to        .3865     .2632     .0001     .0001     1.0000     .8045     .2511
```

Figure 1: Main results table as an output

Analysis

In practical terms for each word in each composition in the corpus we ask

a) how often has the word been misused in the other compositions

b) how often have the words around it (its collocates) signaled misuses

c) how often has its part of speech been misused?

Let us take as an example the excerpt

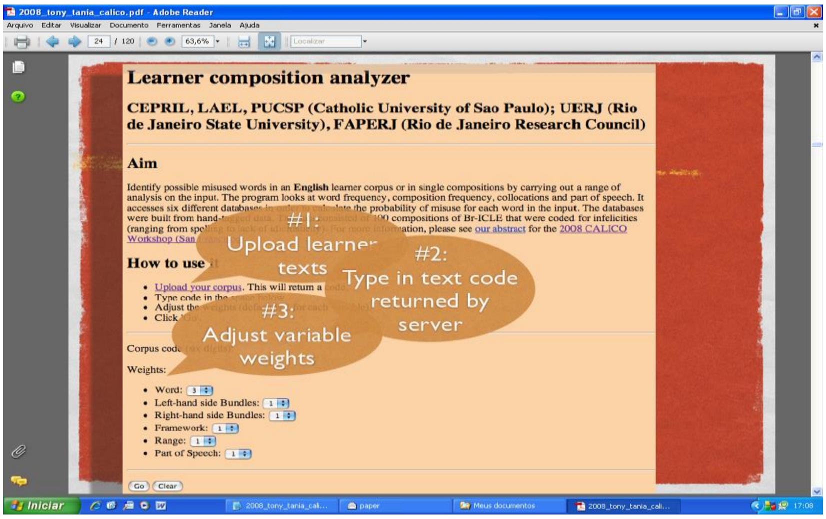
> "There have been many changes in the world throughout the years that caused many effects on humanity, some of which have made people question: do we still have the capacity to dream?"

The program focuses on the item *throughout* and checks whether it has been used wrongly in other compositions. The answer in this case is that it has been used wrongly 23% of the time. Then it asks the question whether the bundle which precedes the wrong item (*in the world*) has preceded other misused words in other compositions. The answer in this case is 21% of the time. The same verification is made in relation to the bundle which follows the misused word (*the years that*), In this case the answer is negative. In terms of framework analysis, it checks whether the particular frame is found around any misused words anywhere in the data. The answer is that it has been found 33% of the time. It finally verifies whether its part of speech has been misused and it calculates the average of 25% of the time. In fact, 17% of the time, *throughout* is used wrongly in the corpus. The suggested form by the annotator is *over the years*, because it is 67 times more frequent in the BNC than *throughout the years.*

The final part of the work was to evaluate the performance of the tool with 20 compositions of the Br-ICLE database. Recall and precision were evaluated . In other words, in terms of recall we asked how many of the total misuses are picked up, or rather how comprehensive is the program. In terms of precision, the question asked was how accurate the program is or how many potential misuses are real misuses. The analyses aimed at the top 10, 20 and then 50 words signaled by the program as potential misuses. These top words were checked against human analysis, looking for matches.

Preliminary Conclusions

The program is 17% accurate and 39% comprehensive. For one in every 6 words it signals a real misuse, and it picks up 4 out 10 misuses (among top 50 words). This is far from ideal, as it was expected that the performance could reach at least 70 % . It was also found that the program's performance is better on texts containing a large number of misuses. Unusually well-written compositions have thrown the program off. As a next step for improving the program, an attempt will be made to reduce the span of the chunk, or to code the texts at word level. This may reduce the number of false alarms. To improve precision, a larger lexicon is needed, one which contains more hand-coded compositions. If the program 'knows' a word is problematic for students, it is likely to be correct in its diagnosis. To improve recall, more hand-coded compositions are also needed. If the program 'knows' which words are misused in more compositions, it is likely to identify a larger number of these misuses. Finally, once tested satisfactorily, the system will be made available online for both students and teachers at www2.lael.pucsp.br/corpora.



Learner composition analyzer

CEPRIL, LAEL, PUCSP (Catholic University of Sao Paulo); UERJ (Rio de Janeiro State University), FAPERJ (Rio de Janeiro Research Council)

Aim

Identify possible misused words in an **English** learner corpus or in single compositions by carrying out a range of analysis on the input. The program looks at word frequency, composition frequency, collocations and part of speech. It accesses six different databases in order to calculate the probability of misuse for each word in the input. The databases were built from hand-tagged data. The data consisted of 100 compositions of Br-ICLE that were coded for infelicities (ranging from spelling to lack of idiomaticity). For more information, please see our abstract for the 2008 CALICO Workshop (San Francisco).

How to use it

- Upload your corpus. This will return a code.
- Type code in the space below.
- Adjust the weights (default 1 for each variable).
- Click 'Go'.

Corpus code (six digits):

Weights:

- Word: 3
- Left-hand side Bundles: 1
- Right-hand side Bundles: 1
- Framework: 1
- Range: 1
- Part of Speech: 1

Go Clear

Figure 2 Future layout of the homepage of the learner composition analyzer

It will be possible to upload compositions on the homepage above, adjust various weights and obtain candidate words for misuse. It is believed that this tool may be helpful to both teachers and students as means of detecting possible errors in learner writing.

References

Bates, L, Lane, J. and **Lange, E**. 1993. *Writing clearly: responding to ESL compositions.* Boston: Heinle and Heinle.

Biber, D. and **Conrad, S.** 1999. "Lexical bundles in conversation and academic prose." In *Out of Corpora - Studies in Honour of Stig Johansson*, H. Hasselgard & S. Oksefjell (eds.), Amsterdam/Atlanta, GA: Rodopi, 181-190.

Bigert, J., **Kann, V.**, **Knutsson, O** and **Sjobergh, J**. 2004."Grammar Checking for Swedish Second Language Learners". *CALL (Computer Aided Language Learning) for theNordic Languages.*

Corder, S. P. 1981. *Error Analysis and Interlanguage*. Oxford: OUP.

Dagneaux, E., Denness S. and **Granger S.** 1998. Computer-aided Error Analysis. *System: An International Journal of Educational Technology and Applied Linguistics* 26(2), 163-174.

De Cock, S., Granger, S., Leech, G., and **McEnery, T**. 1998. "An automated approach to the phrasicon of EFL learners", in S. Granger (ed.) *Learner English on Compute.r* London and New York: Addison Wesley Longman., 67-79.

Ellis, R. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Ferris, D. 2004. *Treatment of Error in Second Language Student Writing.* Michigan: The University of Michigan Press.

Ferris, D. 2003. "Responding to Writing". In Kroll, B (ed.) *Exploring the Dynamics of Second Language Writing.* Cambridge: Cambridge University Press.

Ferris, D. 1995. Student reactions to teacher response in multiple-draft composition classrooms, *TESOL Quarterly, 29*, 33-53.

Granger, S. 2003. "Error-tagged Learner Corpora and CALL: A Promising Synergy". *CALICO Journal, 20*/4: 465-480.

Higgs, T.V, and **Clifford, R.** 1982. "The push toward communication". In Higgs, T (ed.), *Curriiculum, competence and the foreign language teacher*. Studies in The Case for Learning. pp 133.

Hyland, K. 2002. *Teaching and Researching Writing.* London: Longman.

Hyland, K. and **Hyland, F.** 2006.*Feedback in Second Language Writing* Cambridge: Cambridge University Press.

Hyland, K. and **Hyland, F**. 2001. "Sugaring the pill: praise and criticism in written feedback. *Journal of Second Language Writing* 10 (3), 185-212.

Krashen, S.D. 1984.*Pprinciples and Practice in Second Language Acquisition.* Oxford: Pergamon Press.

Knoblauch, C. and **Brannon, L**. 1984. *Rhetorical traditions and the teaching of writing.* Upper Montclair, NJ: Boyton Cook.

Kroll, B. 2003. 'Teaching the next generation of second language writers". In Kroll, B. (ed.) *Exploring the dynamics of second language writing.* Cambridge: Cambridge University Press

Lalande, J. F. 1982. Reducing composition errors: an experiment. *Modern Language Journal* **66**, pp. 140–149.

Lennon, P. 1991. "Error: Some Problems of Definition, Identification, and Distinction". *Applied Linguistics*, 12(2):180-196.

Lightbown, P. 1998. "The importance of timing in focus on form". In Doughty, C e Williams, J. (eds.) *Focus on form in classroom second language acquisition.* New York: Cambridge University Press.

Nicholls, D. 1999. "The Cambridge Learner Corpus - Error coding and analysis for writing dictionaries and other books for English learners." Paper presented at the *Learner Corpus Workshop*, Showa Women's University, Tokyo, July 30,1999.

Reid, J. 1994. "Responding to ESL students' texts: the myths of appropriation. *TESOL Quarterly (28),* 273-292.

Renouf, A. and **Sinclair, J. M**. 1991. "Collocational frameworks in English." In *English Corpus Linguisitcs - Studies in honour of Jan Svartvik,* K. Aijmer & B. Altenberg (eds.). London: Longman,128-144.

Scott, M. and Tribble, C. 2006. Textual patterns: keywords and corpus analysis in language education. Amsterdam: John Benjamins.

Selinker, L. 1992. *Rediscovering interlanguage*. London: Longman.

Truscott, J. 1996. "The case against grammar correction in L2 writing classes. *Language Learning.* 46:327-69.

INVESTIGATIVE LEARNING IN DIFFERENT TYPES OF ACADEMIC WRITING IN THE SPACE CORPUS

Josef Schmied[203]

Abstract

This contribution demonstrates how academic texts can be used in inductive, genre-based corpus-linguistic learning, where the learner as investigator develops hypotheses about the text based on corpus-specific features. The SPACE Corpus consists of corresponding Specialised and Popular ACademic English texts in two types of academic writing, specialised original research articles (i.e. specialised expert-to-expert communication) and popular articles based on these research articles in the New Scientist (i.e. popular expert-to-layman communication). A brief case study focuses on the qualitative evaluation and quantitative measurement of linguistic determinants like modal and evaluative sentence adverbs. These are crucial in academic writing since they determine the effect of the "science story" on the reader in the specialized and the popular domains. We would expect relatively more "extreme" adverbial usage in the more popular texts, since popular writers may have to exaggerate or "sex up" their texts to make them more interesting.

Keyword: ELT, investigative learning, adverb functions, English for Academic Purposes, writing training, metadiscourse, expert-to-expert/layperson communication

Specialised and Popular ACademic English (SPACE): a new corpus for learning about academic writing

The linguistic basis of our discussion is a new corpus of corresponding Specialised and Popular ACademic English texts, hence the acronym SPACE corpus (Schmied 2006 and 2007 or Haase 2007). The texts in the corpus can be classified as specialised and popular research articles, if we want to follow a genre approach, which has become popular since Swales (1990) and is still (Swales 2004). The texts are extracted from different academic disciplines (like natural sciences, psychology and medicine) and from two different types of sources: On the one hand, we have texts in popular science journals, in which the non-specialist versions refer explicitly to their source materials in specialist on-line publications. On the other hand, we have academic online databases like *arXiv* (arxiv.org) and publications in the *Proceedings of the National Academy of Sciences* (PNAS, pnas.org) that are the basis of their popular academic adaptations in journals like *New Scientist* (and sometimes even more popular reflections in national and international English-language newspapers). The *New Scientist* has been considered as a leading international interdisciplinary journal for a long time. It contains numerous short articles that make current specialist research results available to the non-specialist "academic layperson"; the articles are sometimes presented by specific science editors, sometimes by an anonymous member of the scientific writers' team. Both types of academic writing can be taken as acknowledged models for their respective genre.

The focus presented in this contribution is more on learning than on teaching, in particular on autonomous learning and on inductive learner-centred analyses of the texts according to several dimensions: readership (specialised vs. popular), discipline (esp. the continuum of natural sciences), adverbial type (e.g. modal or evaluative, cf. 2 below), and others (e.g. some cohesive devices depend typically on the position in the text, like beginning or end).

The basic assumption is that there are differences in English usage for different academic "cultures" based on readership (i.e. expert to expert, expert to layperson) as well as different discipline cultures, like physics (with the subcategories quantum physics, particle physics and astrophysics) or biosciences (with the subcategories biochemistry, genetics and microbiology). Additionally, a few texts in psychology and medicine for comparative purposes have been compiled (so far).

[203] Josef Schmied studied mainly English and geography at Erlangen-Nürnberg (Germany) and the University of Kent at Canterbury (UKC, England). He taught at the Universities at Bamberg (PhD 1985), Bayreuth (Habilitation 1991) and Dresden before coming to Chemnitz to set up a new department here in April 1993. Research Interests: His main research interests are in Language & Culture (sociolinguistics, English in Africa and SE Asia, Academic English) and in Language & Computers (corpus-linguistics, e-learning, www English and Wiki+).

The basic set-up of the SPACE Corpus can be seen in table 1:

	readership	**specialised**	**popular**	relationship **popular:special**
texts	#	words	words	
Physics	**0001-0046**			
quantum-		115981	15327	13%
particle-		26384	4574	17%
astro-		167553	18119	11%
bioscience	**0047-0107**			
biochemistry		53212	7114	13%
genetics		183312	19101	10%
microbiology		53139	5176	10%
Σ		599581	69411	12%
average		99930,1667	11568,5	

Table 1: Set-up of the Specialised and Popular ACademic English (SPACE) Corpus

Table 1 makes it also clear that the (proportions of the) text categories are uneven in length, partly because the popular publications of the same “science story” are usually only one tenth of the original specialised version, although there are great differences between disciplines and individual texts.

Modal and evaluative adverbials as crucial signals of author involvement and commitment in academic writing

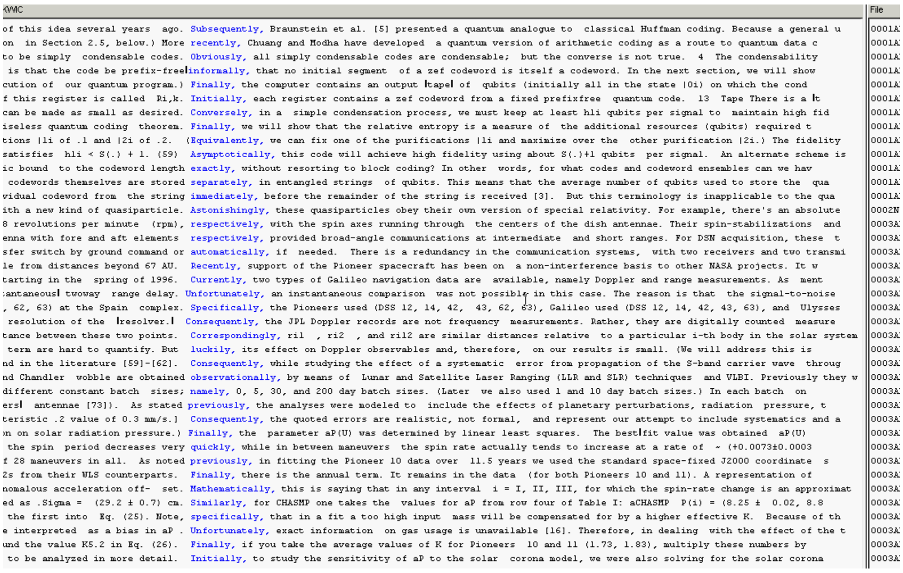
Author stance and engagement are crucial variables in academic interaction (Hyland 2005 and 2006). Thus, it is not surprising that standard grammars discuss them under different name(s) in great detail. Quirk at al. (1985: 620-631) distinguish between two types of content disjuncts, indicating degree of truth and value judgment. The big *Cambridge Grammar* (Huddleston/Pullum 2002: 767-773) discusses modal and evaluative adjuncts, but the corresponding brief section in the student version (Huddleston/Pullum 2005: 78-81) only mentions modifiers and supplements and emphasizes that supplements are set apart intonationally from the rest of the clause. The *Longman Grammar of Spoken and Written English* (Biber et al. 1999: 562) is the only one that demonstrates in its frequency and distribution tables that many stance (e.g. *generally*, *indeed*) and linking (*however*) adverbs are salient features of academic prose. For qualitative analyses, a considerable body of secondary literature is also available (e.g. Hoye 1997).

Corpus-induced investigate learning

The concept of learner-as-researcher has a long tradition at TALC conferences and corpus-ELT publications (Gavioli 2005). In contrast to other learning resources like grammars and dictionaries, corpora (including derived concordances and frequency or distributional statistics) do not provide explanations of language phenomena, but only data – if these data are extracted properly from an appropriate corpus. This is why the learner-as-researcher has to learn to ask questions first. These questions may be derived from the investigative researchers own writing as well as from the literature on the respective linguistic topic. Thus contrary to wide-spread impressions, investigative learning does not start from a *tabula rasa*. Learners do not find inspiration in the data alone, but a good combination of qualitative and quantitative analyses makes them realize much more than any grammar book can tell them. With only some direct training in corpus-linguistic methods and a brief warning including the usual caveats about the limitations of the corpus materials, students can learn about real-language usage, which may be interesting theoretically and useful practically at the same time. This is what this contribution is trying to exemplify for Academic Writing.

Investigating modal and evaluative adverbials in the SPACE Corpus

Students with some training in investigative language analysis can apply corpus tools like AntConc and use the information provided in the reference grammars mentioned above to test old and new hypotheses. Thus they can try a lexical approach and use the word lists or examples available in the grammars mentioned above to search for specific tokens, they can take a word-class approach and search a tagged corpus for all adjectives (_RB tags or similar) or the can take a morphological-punctuational approach and search for **ly,* since the word-class and intonation (symbolized in the query by the prototypical –*ly* ending and a comma afterwards) provide a large number of cases. Whereas the adverb tag query obviously leads to overcollection (since there are many more adverbs that do not have clause functions in English), the **ly,* search clearly leads to undercollection (since many adverbs are not morphologically marked or separated by comma from the main clause). The results of the **ly,* query can be seen in Table 2.



```
KWIC                                                                                                                                   File
of this idea several years  ago. Subsequently, Braunstein et al. [5] presented a quantum analogue to  classical Huffman coding. Because a general u   0001A
on  in Section 2.5, below.) More recently, Chuang and Modha have developed  a quantum version of arithmetic coding as a route to quantum data c     0001A
to be simply  condensable codes. Obviously, all simply condensable codes are condensable;  but the converse is not true.  4  The condensability    0001A
 is that the code be prefix-free|informally, that no initial segment  of a zef codeword is itself a codeword. In the next section, we will show     0001A
cution of  our quantum program.) Finally, the computer contains an output |tape| of  qubits (initially all in the state |0i) on which the cond     0001A
f this register is called  Ri,k. Initially, each register contains a zef codeword from a fixed prefixfree  quantum code.  13  Tape There is a |t   0001A
can be made as small as desired. Conversely, in a  simple condensation process, we must keep at least hli qubits per signal to  maintain high fid  0001A
iseless quantum coding  theorem. Finally, we will show that the relative entropy is a measure of  the additional resources (qubits) required t     0001A
tions |li of .1 and |2i of .2.  (Equivalently, we can fix one of the purifications |li and maximize over the  other purification |2i.) The fidelity 0001A
satisfies  hli < S(.) + 1. (59)  Asymptotically, this code will achieve high fidelity using about S(.)+1 qubits  per signal.  An alternate scheme is 0001A
ic bound  to the codeword length exactly, without resorting to block coding? In other  words, for what codes and codeword ensembles can we hav     0001A
 codewords themselves are stored separately, in entangled strings  of qubits. This means that the average number of qubits used to store the  qua  0001A
vidual codeword from  the string immediately, before the remainder of the string is received [3].  But this terminology is inapplicable to the qua  0001A
ith a new kind of quasiparticle. Astonishingly, these quasiparticles obey their own version of special relativity. For example, there's an absolute 0002N
8 revolutions per minute  (rpm), respectively, with the spin axes running through  the centers of the dish antennae. Their spin-stabilizations  and 0003A
enna with fore and aft elements  respectively, provided broad-angle communications at intermediate  and short ranges. For DSN acquisition, these  t  0003A
sfer switch by ground command or automatically, if  needed.  There is a redundancy in the communication systems,  with two receivers and two transmi 0003A
le from distances beyond 67 AU.  Recently, support of the Pioneer spacecraft has been on  a non-interference basis to other NASA projects. It w     0003A
tarting in the  spring of 1996.  Currently, two types of Galileo navigation data are  available, namely Doppler and range measurements. As  ment    0003A
:antaneous| twoway  range delay. Unfortunately, an instantaneous comparison  was not possible in this case. The reason is that  the signal-to-noise  0003A
, 62, 63) at the Spain  complex. Specifically, the Pioneers used (DSS 12, 14, 42,  43, 62, 63), Galileo used (DSS 12, 14, 42, 43, 63), and  Ulysses 0003A
 resolution of the  |resolver.|  Consequently, the JPL Doppler records are not frequency  measurements. Rather, they are digitally counted  measure 0003A
tance between these two points.  Correspondingly, ril  , ri2  , and ril2 are similar distances relative  to a particular i-th body in the solar system 0003A
 term are hard to quantify. But  luckily, its effect on Doppler observables and, therefore,  on our results is small. (We will address this is    0003A
nd in the literature [59]-[62].  Consequently, while studying the effect of a systematic  error from propagation of the S-band carrier wave  throug 0003A
nd Chandler  wobble are obtained observationally, by means of  Lunar and Satellite Laser Ranging (LLR and SLR) techniques  and VLBI. Previously they w 0003A
different constant batch  sizes; namely, 0, 5, 30, and 200 day batch sizes. (Later  we also used 1 and 10 day batch sizes.) In each batch  on      0003A
ers|  antennae [73]).  As stated previously, the analyses were modeled to  include the effects of planetary perturbations, radiation  pressure, t    0003A
teristic .2 value of 0.3 mm/s.]  Consequently, the quoted errors are realistic, not formal,  and represent our attempt to include systematics and a  0003A
on on solar radiation pressure.) Finally, the  parameter aP(U) was determined by linear least squares.  The best|fit value was obtained  aP(U)     0003A
 the spin  period decreases very quickly, while in between maneuvers  the spin rate actually tends to increase at a rate of  ~ (+0.0073±0.0003      0003A
f 28 maneuvers in all.  As noted previously, in fitting the Pioneer 10 data over  11.5 years we used the standard space-fixed J2000 coordinate  s    0003A
2s from their WLS counterparts.  Finally, there is the annual term. It remains in the data  (for both Pioneers 10 and 11). A representation of     0003A
nomalous acceleration off-  set. Mathematically, this is saying that in any interval  i = I, II, III, for which the spin-rate change is an approximat 0003A
ed as .Sigma =  (29.2 ± 0.7) cm. Similarly, for CHASMP one takes the  values for aP from row four of Table I: aCHASMP  P(i) = (8.25 ±  0.02, 8.8    0003A
 the first into  Eq. (25). Note, specifically, that in a fit a too high input  mass will be compensated for by a higher effective K.  Because of th  0003A
e interpreted  as a bias in aP . Unfortunately, exact information  on gas usage is unavailable [16]. Therefore, in dealing  with the effect of the t 0003A
und the value K5.2 in Eq. (26).  Finally, if you take the average values of K for Pioneers  10 and 11 (1.73, 1.83), multiply these numbers by      0003A
 to be analyzed in more detail.  Initially, to study the sensitivity of aP to the solar  corona model, we were also solving for the solar corona    0003A
```

Table 2: AntConc concordance for **ly,* in the first few SPACE Corpus files

Incomplete as the results in Table 2 are, students can take the concordance material as a starting point for functional analyses on the basis of traditional grammar terminologies. The concordance above lists the sentences from the first to the last corpus texts. Adverbs in contxt are not ordered according to the frequency of the adverb (which can be done by a simple mouse click), but from the last column we can see that there are many more cases extracted from the third specialised file (0003AX) than from the others. A qualitative analysis goes through the sample sentences one after the other: Thus, *subsequently* obviously has linking functions, *recently* temporal, *obviously* modal, etc. Modal and evaluative adverbs can, however, also be classified according to their strength or propensity - and our corpus has been tagged accordingly. On a scale from 1 to 9 with a neutral 5 in the middle, *obviously*, *exactly* or *astonishingly* could (despite their clear semantic differences) be given a high value (like 9), whereas *similarly* could be seen as neutral (like 5), etc. A short list already shows clearly that extreme values are most prominent than neutral values, since "extreme" adverbs make the "science story" more interesting.

When students are asked to hypothesize about the distribution of these types and values across the readership types and disciplines in the SPACE Corpus, they quickly come up with the suggestion that more popular texts tend towards more extreme values to attract the readers attention or in new style "to sex up a story". However, when they have determined the absolute figures for the different readership types for instance, they will soon realise that absolute figures mean nothing when the text sizes are so different (as we see in the last column of Table 1 giving proportions).

Of course, it will take many more steps until student researchers will arrive at a comprehensive summary of the issue, as can be seen in table 3.

Adverb types	Specialized			Popular	Relationship
	AX	PN	AX+PN	NS	NS:(AX+PN)
evaluative	191	160	351	69	20%
modal	34	14	48	11	23%
linking	131	129	260	7	3%
domain-specific	35	22	57	3	5%

Table 3: Types of sentence adverbs in the SPACE Corpus

This finally allows them a tentative answer to the questions:

But there are more questions. Students can, for instance, ask themselves whether the variation between discipline subcategories are less pronounced than between the major disciplines categorized here. They may ask which linguistic variables (e.g. linking or modal adverbs) change in which direction when "less" scientific disciplines are included in the study.

Conclusion: Metadiscourse as awareness raising and personal writing support

In this article, we intended to show that the newly available SPACE corpus provides a good basis for investigative learning. We hope to have indicated some possible steps for student activities that can be introduced in an exemplary manner relatively quickly so that students are equipped with appropriate tools and a database for many more explorations on their own.

The adverbs analysed here are a good example of what has been called metadiscourse in academic writing. This includes the writers' management of the perception of the academic discourse by the readers in terms of information management (e.g. sentence organisation) as well as stance (e.g. the personal commitment in the argumentation). Here small words like *consequently* and *unfortunately* play an important role. This is not only of theoretical importance to advanced students, but can have very practical implications, since investigative learners can use the database and the results of their queries directly to improve their own academic writing and success. If these examples lead writers to a greater awareness of linguistic devices that can make texts more coherent, more reader-friendly or more interesting for the reader, the concept of metadiscourse can make the interaction of writer, text and reader more effective. As Hyland (2004: 133f) put it:

The ability of writers to control the level of personality in their texts, claiming solidarity with readers, evaluating their material, and acknowledging alternative views, is now recognized as a key feature of successful academic writing.

Although there a great diversity of approaches to teaching academic writing (Partridge 2004), our investigative examples have at least given a glimpse into the possibilities of student text explorations that clearly have the potential of combining theory and practice in university teaching programmes – and if this combination even includes a small project might give students some real work after their studies, it does not do any harm either.

References

Biber, D., Johansson, S., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Gavioli, L. 2005. *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.

Haase, C. 2007. "Corpora and Academic English: Compilation, Analysis, and Teaching." *ReCall* 193 Special Issue on Incorporating Corpora in Language Learning and Teaching.

Hoye, L. 1997. *Adverbs and Modality in English.* London: Longman.

Huddleston, R. & Pullum G. 2002. *The Cambridge Grammar of the English Language.* Cambridge: CUP.

Huddleston, R., & Pullum, G. K. 2005. *A Student‘s Introduction to the English Language.* Cambridge: University Press.

Hyland, K. 2004. “Disciplinary Interactions: Metadiscourse in L2 postgraduate writing”. *Journal of Second Language Writing 13: 133-151.*

Hyland, K. 2005. “Stance and Engagement: A Model of Interaction in Academic Discourse.” *Discourse Studies 72: 173-92.*

Hyland, K. 2006. *English for Academic Purposes: an Advanced Resource Book*. New York: Routledge.

Paltridge, B. 2004. “Academic Writing”. *Language Teaching* 37: 87-105.

Quirk et al. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Schmied, J. 2006. “Specialist vs. Non-Specialist Academic Discourse: Measuring Complexity in Lexicon and Syntax”. In *Discourse and Interaction 2. Paper presented at the Brno Seminar on Linguistic Studies in English: Proceedings*, ed. by Povolná, R. & Dontcheva Navratilová, O. , 143-152. Brno: Masaryk University.

Schmied, J. 2007. “The Chemnitz Corpus of Specialised and Popular Academic English”. In Pahta, P./ Taavitsainen, I./ Nevalainen, T./ Tyrkkö, . eds. Towards Multimedia in Corpus Studies. University of Helsinki, Contacts and Change in English Series: Studies in Language Variation, Contacts and Change in English, Vol. 2. http://www.helsinki.fi/varieng/journal/volumes/02/schmied/

Swales, J. M. 1990. *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, J. M. 2004. *Research Genres. Exploration and Application*. Cambridge: Cambridge University Press. The SPACE Corpus is available with a detailed manual: www.tu-chemnitz.de/phil/english/REAL/SPACE

SEEKING NEEDLES IN THE WEB HAYSTACK:
FINDING TEXTS SUITABLE FOR LANGUAGE LEARNERS

Serge Sharoff[204]
Svitlana Kurella[205]
Anthony Hartley[206]

Abstract

While modern communicative methods of language teaching rely heavily on authentic, typical and recent materials, traditional graded readers often fall short of these requirements. The project reported in this paper is aimed at (1) designing methods for retrieving web texts that are suitable for a particular group of learners, and (2) using them in actual language teaching. Given that very little is currently known about the features that constitute texts suitable for individual language learning needs in a variety of languages, the paper reviews options for selecting texts according to their lexicon, grammatical features and readability statistics on the basis of trial runs with students studying English, Chinese, German and Russian. Among other sources, we used simplified Wikipedia texts (http://simple.wikipedia.org/) and their counterparts in the main English wikipedia. In addition to selecting texts for reading exercises, we experimented with the same model applied to selecting texts suitable for grammatical gap-fill exercises. For instance, it was used for finding texts rich in modal verbs or conjunctions. The use of authentic running texts instead of artificial single-sentence examples improves students' motivation in such exercises and helps in contextualising grammatical rules.

Keywords: graded readers, LSP, multilingual resources, reading skills, text selection

Introduction

While modern communicative methods of language teaching rely heavily on reading materials that are authentic, typical and recent, traditional collections graded for reading difficulty often fall short of meeting these requirements. Graded readers for some languages, e.g., Mandarin Chinese, frequently contain adaptations of authentic texts that restrict the range of characters, simplify idiomatic expressions, rewrite syntactic constructions not covered in text books, etc. Such texts have their pedagogical value, but they do little to prepare students for reading 'real' texts. As for the typicality of genres, texts in graded readers often have a large proportion of classic literary texts – e.g., Goethe and Schiller for German, or Pushkin and Tolstoy for Russian – while texts more relevant to accomplishing everyday tasks, such as administrative or argumentative texts, are missing. Finally, for domains undergoing rapid changes, such as software or international trade, all languages, including English, are seriously lacking in up-to-date reading resources oriented to language learners.

Nowadays it is relatively easy to collect a large corpus from the Web, using either search engines (Sharoff, 2006) or web crawlers (Baroni and Kilgarriff, 2006). Modern text classification methods help in selecting subsets of web corpora belonging to particular domains or genres (Sharoff, 2007). It is also possible to use entire domain- or genre-specific collections, e.g., Wikipedia or Reuters. However, the content of what is collected is often beyond the reading skills of language learners. Authentic corpora are commonly regarded as a useful resource for tailoring reading materials to what is expected by the language learner (Leech, 1997), but as far as we know there have been no studies that have actually implemented an automatic text selection procedure for a variety of languages and put it into teaching practice.

Existing research on automatic text selection is mostly devoted to teaching English reading skills in the context of US school education (Schwarm, Ostendorf, 2005; Collins-Thompson, Callan, 2004). Traditional readability measures, such as Flesch Reading Ease, Flesch-Kincaid and Gunning Fog have also been deployed in this context (DuBay, 2004). Some recent projects (Heilman *et al*., 2008; Kilgarriff *et al*., 2008; Kotani *et al*., 2008) do address the problems of selecting texts (or examples) aimed at non-native speakers of English. The grading procedure is typically based on lexical coverage or frequent words, e.g. words such as *essay* are selected as predictive for higher-grade texts (Collins-Thompson, Callan, 2004). When grammatical features are used, these are based on

[204] Serge Sharoff is a lecturer in the Centre for Translation Studies (CTS), University of Leeds. He is involved in several projects related to corpus collection and corpus-based technologies for language learning and translation

[205]Svitlana Kurella is a PhD student in CTS. Her project aims at developing an effective corpus-based methodology for acquiring reading abilities in Polish and Ukrainian based on the knowledge of a second language (L2, here Russian). She is also involved in teaching Russian at Leeds.

[206] Anthony Hartley is the Director of CTS. His research interests are in Machine Translation, controlled languages and quality of translation and interpreting.

advanced English parsers (Kotani *et al*., 2008). However, it is difficult to extend such approaches to languages other than English, since adequate parsers are rarely available and, if they are, their sets of features usually differ to such a degree that a new approach would be needed for each language. Moreover, these projects provide little information on the actual use of graded texts.

The project reported in this paper is aimed at (1) designing methods for retrieving web texts that are suitable for a particular group of learners for a variety of languages, and (2) using them in actual language teaching. The emphasis of this study was on (1) finding measures that work across a variety of languages without requiring complex resources, such as parsers, and on (2) their integration into the language learning process.

Features for text selection

In consultation with language teachers we identified a variety of features that might be expected to make a text difficult to read, and tested for their effectiveness in predicting the difficulty of each text:

1. lexical coverage by word bands from respective frequency lists, i.e. top 1000, top 2000, top 3000 words; the General Service List (GSL) was used for English, frequency lists from Web corpora (Sharoff, 2006) for Chinese, German and Russian
2. average sentence length (ASL)
3. average word length in syllables (ASW) – pinyin count was used for Chinese
4. Flesch Reading Ease (FRE)
5. coverage by more frequent part of speech (POS) trigrams
6. average number of conjunctions per sentence
7. average number of lexical verbs per sentence
8. average number of passive verbs per sentence
9. average number of modal verbs per sentence
10. average number of prepositions per sentence
11. average number of punctuation marks per sentence

Since reliable part of speech taggers are available for all languages under consideration, it is possible to use them to detect known grammatical complexities without the need for full parsers. Identifying frequent POS trigrams is an easy way to determine whether a text deviates from the standard language model, cf. the importance of language modeling in (Schwarm, Ostendorf, 2005) and (Kilgarriff *et al*., 2008). The number of lexical verbs is an indirect measure of sentence complexity, while passives and modals present well-known problems for language learners. The use of conjunctions is a way of detecting the complexity of discourse development. Kotani *et al*. (2008) assessed discourse complexity by the number of pronouns, but reported that this feature does not reduce the error rate. We used FRE (with language-specific formulas for German and Russian) to test for any correlation between the traditional readability measures mentioned earlier and our approach:
FRE_en = 206.835 - (1.015 x ASL) - (84.6 x ASW)
FRE_de = 180 - ASL - (58.5 x ASW)[207]
FRE_ru = 206.835 - (1.3 x ASL) - (60.1 x ASW)[208]

Machine Learning experiments

We calculated statistics for a range of “easy” texts from the Simple English Wikipedia website (http://simple.wikipedia.org/), and their counterparts from the main English Wikipedia website. Guidelines for contributors to the Simple English Wikipedia advise the use of Basic English vocabulary, the active voice and ‘Basic English verb[s] in past, present or future only’. Further texts were added from other sources; s-sherlock is a simplified version of 'The Boscombe Valley Mystery' as published in the Penguin Readers series (sherlock is the original text). The resulting file was processed by the Weka implementation of Principal Component Analysis (PCA) to identify the most significant correlation between the features. It resulted in two main components representing a linear combination of normalised original features:

[207] http://de.wikipedia.org/wiki/Lesbarkeitsindex
[208] http://ru.wikipedia.org/wiki/Индекс_читабельности

0.415prepositions+0.386lexverbs-0.352fre+0.334passiveverbs+0.32 top3000...
-0.416top2000-0.412top3000-0.41top1000+0.375punctuation+0.36 conjunctions...

These linear combinations can be roughly labelled, respectively, as the grammatical and lexical dimensions of difficulty (ranging from easy to difficult). Figure 1 displays a sample of texts in terms of the two dimensions. The majority of Simple Wikipedia texts (prefix s- in the chart) are easier on both dimensions than their main Wikipedia counterparts. For example, the simplified and standard definitions of ‘induced abortion’ read:

1 a. An induced abortion is when a person does something to end the pregnancy.

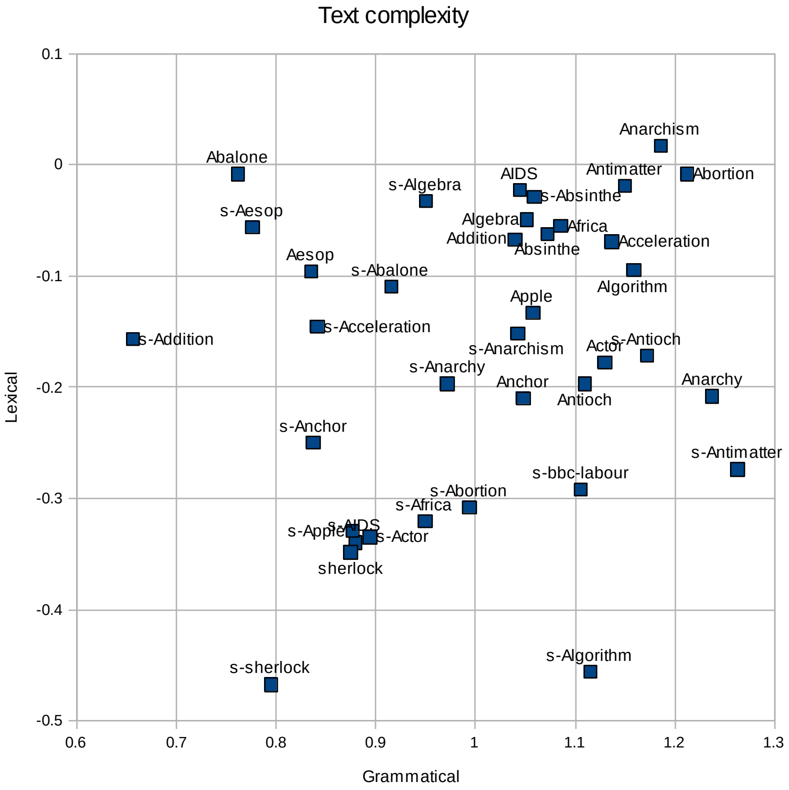
1 b. Induced abortion is the removal or expulsion of an embryo or fetus by medical, surgical, or other means at any point during human pregnancy for therapeutic or elective reasons.

In a few cases, the PCA results suggest that ‘simplified’ texts are not necessarily easier along both dimensions. For example, s-aesop is classified as grammatically easier yet more difficult lexically, a decision supported by the following extracts:

2 a. Aesop’s fables are still taught as moral lessons and used as subjects for various entertainments, especially children’s plays and cartoons.

2 b. The various collections that go under the rubric “Aesop’s Fables” are still taught as moral lessons and used as subjects for various entertainments, especially children’s plays and cartoons.

In a small number of cases, the PCA results suggest that the ‘simplified’ text is in fact more difficult along both the lexical and the grammatical dimension. However, closer inspection of the texts in question indicates that the classifier has in fact detected true complexity that is belied by the ‘simplified’ label. For example, the difficulty of s-absinthe is shown to be slightly greater than that of absinthe in the main English Wikipedia. This has been indeed acknowledged by its readers, who have added the following tag to s-absinthe: ‘The English used in this article may not be easy for everybody to understand.’ Similarly, inspection of s-antioch reveals that it has been edited from antioch by simple deletion. None of the complexity of the main entry has been modified and, indeed, some of the deleted text is grammatically simpler than the retained text, even if its content has been judged to be not worth preserving.



We conducted a similar exercise for Chinese and Russian, this time selecting original texts published in quality newspapers for comparison with texts published on the Chinese and Russian BBC websites and considered by language teachers to be significantly easier for native English students. Their PCA transforms yielded the following combinations of parameters:

Chinese:

0.522asl+0.475lexverbs+0.422prepositions-0.388fre+0.387conjunctions...

-0.495top500-0.494top1000-0.482top2000+0.379asw-0.324fre...

Russian:

0.461asl-0.444fre+0.433lexverbs+0.332conjunctions+0.316asw...

-0.556top1000-0.55top500-0.533top2000-0.238conjunctions-0.135asw...

This supports the intuition of language teachers that texts with a higher number of less frequent words, conjunctions, prepositions and longer sentences tend to be more difficult for language learners.

Language teaching experiment

To ground our research in the practice of language teaching we selected some texts classified as moderately difficult according to the dimensions reported above. These texts were presented as reading exercises to language students – British students for all languages except English, and foreign students attending pre-sessional English language courses. Language teachers confirmed that the texts selected were appropriate for their students. They also designed multiple-choice questions to test understanding of key aspects of the content and the argumentation in each text. For instance, a text from http://news.bbc.co.uk/1/hi/magazine/4149835.stm (shown as s-bbc-labour in Figure 1) was assessed by ten questions, of which the following is one example:

1 What is “unpaid overtime” (lines 3-4)?

a extra work which no-one pays them for

b getting no pay when they are away from work

c rest and holiday periods which are not paid

d longer holidays without pay

e work which they should not be doing

The main task of the test was to check the assumptions of the language teachers about the appropriateness of certain texts according to students’ level of language competence. The test results confirm that the teachers’ predictions of the difficulty level correspond to the automatic score: the average number of correct answers for the English was 6.2 (σ=2.03). Questions to test global comprehension (Alderson, 2000) in our test proved more difficult than those for testing local comprehension: the former gave 40% vs. 90% of correct answers for the latter.

Applications of automatically graded texts

The proposed method of grading texts by their difficulty finds various potential applications in the process of learning and teaching foreign languages. The first application is creating graded readers for extra-curricular reading. Teachers' experience shows that successful students regularly read authentic texts in foreign languages. Weaker learners, on the other hand, struggle to find texts suitable for their level of linguistic competence and therefore are often put off by the excessive difficulty of the majority of authentic texts available on-line. However, reading outside the classroom can be crucial for making progress in language learning at the intermediate level, especially outside the country of the target language. Thus, this use meets the needs of many students.

A second application will be beneficial for the language teachers. Selecting texts for the classroom is a well-known problem which, until now, has largely relied on intuitive decision-making and teachers' experience. This requires a teacher to solve many tasks at once to find: a text on a desired topic; a text that is suitable for certain grammar or/and lexical tasks; a text suitable for a certain group of students; a text that can be discussed; a text of a certain genre; etc. This list of requirements can vary and is open-ended. Most teachers are happy if the text can fit at least two of these requirements, and if they do not have to amend it. Finding suitable texts within the short time that is usually available for lesson preparation is a very demanding process. Automatic text selection and grading should relieve and support teachers to a great extent. Our future work will include developing a multipurpose tool not only to give a teacher the opportunity of selecting suitable texts according to subject and difficulty (that alone would be a great advantage), but also to give them the possibility of picking up texts with specific grammatical and lexical phenomena on which they are working in class.

This point relates to a third practical application of the proposed method: automatic creation of grammatical or lexical exercises which can help the teacher to develop meaningful tasks or to support the student in exploratory activities on the basis of the authentic content. We took our model already used for selecting texts for reading exercises and experimented with applying it to the selection of texts suitable for grammatical gap-fill exercises. For instance, it was used to find texts rich in modal verbs or conjunctions. The use of authentic running texts in such exercises instead of artificial single-sentence examples improves students' motivation and helps in contextualising grammatical rules. For instance, a text from http://teacher.scholastic.com/activities/wwatch/hurricanes/witnesses.htm was found to have a significant coverage by a large variety of modal verbs (some legal texts had greater coverage, but little variety of lexical items), so that it allowed a gap fill exercise like:

I screamed, "Mom, Dad, we _______ get out! The water's rising!" I packed one outfit in my book bag, and my parents grabbed a few things. We _______n't find our dog, Bear. But we just _______ leave.

Conclusions and further research

The project will be developed further along the two main lines of enquiry identified above. In terms of feature selection, we would like to experiment with other features indicative of text difficulty, such as nominalisations or the number of different participants in a text, as well as language-specific features, e.g. the use of oblique cases in Russian or genitive in German. Such features are not always marked explicitly, e.g. a list of nouns is needed to find nominalisations, we are not aware of a reliable German tagger that marks genitive constructions. Also, more research is needed on finding the most discriminative features. In our experiments the PCA transform did not select the coverage by POS trigrams, a feature considered to be capturing the language model. In terms of using selected texts in reading exercises, we would like to make more experiments with testing text comprehension, for instance, using MCQs to check understanding of texts that are considerably more or less difficult according to our model. Also it is important to investigate the balance between local and global comprehension required of an individual text. For instance, reading an instruction for operating a safety-critical device implies paying attention to exact understanding of every step, while reading a magazine might be less demanding. Automatic text selection has to take such purposes into account as well.

References

Alderson, J. Charles. 2000. Assesing Reading. Cambridge.

Baroni, M., Kilgarriff, A. 2006. Large linguistically-processed Web corpora for multiple languages. In: *Companion Volume to Proc. of the European Association of Computational Linguistics*, Trento, 87-90.

Collins-Thompson, K., Callan, J. 2004. A language modeling approach to predicting reading difficulty. In *Proc. of HLT/NAACL*, 193-200

DuBay, W. 2004. The principles of readability. Impact Information, California. http://www.impact-information.com/impactinfo/readability02.pdf

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Kilgarriff. A., Husak, M, McAdam, K, Rundell, M, Rychly, P. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Euralex08.*

Leech, G. 1997. Teaching and language corpora : A convergence. In A. Wichmann, S. Fligelstone, A. M. McEnery, & G. Knowles (eds), *Teaching and Language Corpora*, London, 1-23.

Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus*, M. Baroni and S. Bernardini (eds.). Bologna: Gedit, 63-98. http://wackybook.sslmit.unibo.it/

Sharoff, S. 2007. Classifying Web corpora into domain and genre using automatic feature identification. In *Proc. of the Third Web as Corpus Workshop*, Louvain-la-Neuve, September, 2007.

Schwarm S., Ostendorf. S. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. *Association for Computational Linguistics*, 2005.

AUTOMATIC CLOZE GENERATION: GETTING SENTENCES AND DISTRACTORS FROM CORPORA

Simon Smith[209]
Scott Sommers[210]
Adam Kilgarriff[211]

Abstract

Cloze exercises are widely used in language teaching, both as a learning resource and an assessment tool. It has been shown that they can cultivate and test a wider range of skills than immediately meets the eye. Cloze has a particularly useful role to play in Taiwan, and other Asian countries, where students of English expect and are expected to memorize a lot of vocabulary. Cloze encourages acquisition of vocabulary through context, rather than the memorization of synonyms or translations. Unfortunately, it is time-consuming and difficult for teachers and materials designers to make up large numbers of cloze exercises.

The present paper briefly reviews the literature on cloze in language learning, including systems which generate cloze items automatically, and an algorithm for automatically generating cloze exercises from corpora is presented. It is a bottom-up algorithm, which takes as input from the teacher-user a lexical item which will form the correct answer to the cloze exercise. It outputs a sentence, extracted from a corpus, which contains the lexical item (with the item itself deleted) and a set of distractors is generated. The distractors have a similar semantic distribution to that of the lexical item, but cannot replace it to form a correct answer in the context of the sentence extracted.

Keywords: cloze, FBQ, Sketch Engine, corpus linguistics, ELT

Introduction

As EFL teachers in Taiwan, we have found cloze exercises (or “fill in the blank” questions, FBQ) to be of great use in our classes, as an instructional as well as an assessment tool. This is especially true, we have found, for very large classes in which many students are reluctant to speak out. Most of the literature (including papers to be mentioned presently) deals with the role of cloze in language proficiency assessment. However, cloze exercises generated by the means we describe in this paper could be used for either purpose, with equal effectiveness.

Cloze is defined by Jonz (1990) as “the practice of measuring language proficiency or language comprehension by requiring examinees to restore words that have been removed from otherwise normal text.” The idea is traditionally attributed to Taylor (1953), when it was used as a test of text readability. The term itself derives from the concept of closure in Gestalt Theory used to describe the human tendency to mentally complete figures even when parts of that figure are missing. Taylor and other cloze researchers have used the term to describe a sample of naturally occurring text in which words are deleted and respondents asked to use semantic clues in filling in these deleted words. By the 1970s, the concept had been incorporated into educational assessment and

[209] Simon Smith is an Assistant Professor in the English Language Center of Ming Chuan University, Taiwan, where he teaches EFL and Linguistics, and conducts research on the use of corpora and corpus tools in the teaching of English and Chinese. He has many years of English teaching experience in Taiwan and China private and university centres. He holds a BA in Chinese & Linguistics, from Leeds; an MSc in Machine Translation, from Manchester/UMIST; and a PhD in Statistical Language Modelling, from Birmingham. In his postdoctoral year, at Institute of Linguistics, Academia Sinica, he worked under Professor Huang Chu-Ren on Chinese corpora and Chinese lexical semantics, including the Chinese Wordnet project.

[210] Scott Sommers is a Lecturer with the Ming Chuan University English Language Center. He has written and spoken on a wide range of topics including the history of English in Asia, commercial education, and social aspects of blogging. His research interests in ELT include assessment, data-driven learning and needs analysis. His personal blog, Scott Sommers’ Taiwan weblog can be found at www.scottsommers.blogs.com.

[211] Adam Kilgarriff is a research scientist working at the interface between corpus linguistics, language technology, lexicography and language teaching. He has published over fifty articles in this area. He has since worked as computational linguist for Longman Dictionaries (1992-95) and as a research scientist at Brighton University (1995-2003). Since 2003 he has been running his own company, Lexical Computing Ltd., which develops language corpora and provides web services for language research and dictionary publishers. He is a consultant to Oxford University Press and Macmillan Publishers amongst others. He has chaired the Association of Computational Linguistics Special Interest Group on the Lexicon (2000-2003), served on the Board of the European Association for Lexicography (2002-06) and, since 2007, has been the founding Chair of the Association of Computational Linguistics Special Interest Group on Web as Corpus. He has served as invited speaker at conferences in three continents.

subsequently into the assessment of English proficiency among second and foreign language learners (Alderson 1978, Oller 1973).

When constructing cloze tests, EFL researchers use a number of different procedures for text selection and word deletion. Deletion procedures generally follow one of three standardized formats. The historical format established by Taylor calls for the deletion of words at regular intervals regardless of their linguistic properties. A second, similar, approach is random word deletion. A third format uses the linguistic properties of words to determine which words get deleted. In this case, the focus might be syntactic (particular parts of speech, such as prepositions, are candidates for deletion) or, as in the work reported here, it might be on the semantics of deleted items.

Manual cloze generation

It is difficult for teachers to think up cloze exercises from scratch. Having composed or located a convincing and authentic carrier sentence, which incorporates the desired **key** (the correct answer, or deleted item), it is also necessary to generate **distractors** (wrong answers suggested to the student). This is not a trivial task, as two important constraints apply. On the one hand, the distractors must be incorrect (inserting them in the blank must generate an incorrect sentence). On the other hand, the distractors must in some sense be viable alternatives for completion of the carrier sentence: near synonyms of the key, for example, or words typically found in similar collocational contexts.

A teacher who tries to generate distractors through intuition and introspection may, therefore, encounter the following paradox: if the distractor is too distant from the key, in a semantic distribution sense, it is likely that the student will find the correct answer very easy to deduce; if the distance is too close, sentences incorporating the distractors may turn out to be infelicitously correct.

If a corpus is consulted when manually generating distractors, the teacher may well have access to the necessary distributional information. Nevertheless, the process is time-consuming and tedious, especially if large numbers of items are required, and the advantages of automation are apparent.

Automatic cloze generation

A growing amount of research has found that cloze can be effectively generated through automated systems. Hoshino and Nakagawa (2007) devised an NLP-based teacher's assistant, which first asks the user to supply a text. The system then suggests deletions that could be made, and helps the teacher to select appropriate distractors. Mostow et al (2004) generated cloze items of varying difficulty from children's stories. The items were presented to children via a voice interface, and the response data was used to assess comprehension. Both of these systems use longer texts, while Sumita et al (2005) describe the automatic generation of single sentence cloze exercises from the World Wide Web. Sumita et al obtain distractors from a thesaurus, and check to make sure that there are zero Google hits for hypothesized sentences in which the **key** (the correct answer) is replaced by distractors.

Our system is similar to that of Sumita et al, in that we select single sentences of authentic language to build our cloze exercises, and that we look for words with similar lexical distribution to the key to serve as distractors. However, we do not constrain our choice of distractors to synonyms, or even near synonyms; indeed, key and distractor could perfectly well be antonyms, as long as they can occur in the same contexts. Another difference between the two systems is that the Japanese team use a published resource to find distractors, and extract carrier sentences from the web. We use distributional information from a corpus for both of these purposes.

Our system is designed to work in a bottom-up fashion. The teacher/user is first invited to select the correct answer (the key); that is to say, the particular lexical item of which they want to check or reinforce the student's understanding. As far as we are aware, this type of architecture is unique. Other automated systems, by contrast, require the user to select a text, and offer assistance in deciding which word to *delete*. This is significant for two reasons: first, because when we are writing cloze exercises for our students, we often use a vocabulary item as a point of departure.

Secondly, our architecture is capable of generating large numbers of cloze items on a given topic ("Business", perhaps, or "Starting out at University"). In Smith, Sommers & Kilgarriff (2008) we reported how to extract corpora, on such topics, from the world-wide web, using WebBootCat (WBC; Baroni et al 2006). The corpora were then used to generate wordlists containing vocabulary salient to the topic. Such wordlists could be readily used as lists of keys to bootstrap collections of on-topic cloze exercises.

System architecture

The system works like this. First, the teacher specifies the key, or a list of keys to be processed. Assume, then, that the teacher/user wishes to teach or test the use of the adjective *sunny*, as used to describe personality. She would enter *sunny* into our system as her chosen key. The system will find words which have a similar lexical distribution to that of *sunny*, such as *rainy*, *windy* and so on. It will do this by establishing that these **potential distractors** (PDs) and the key are all found with some set of other words (**key and PD collocates**, KPDCs) such as *weather* and *climate*.

Next, the system looks in the corpus for a word which co-occurs with the key, but never with the PDs. This word is termed the **key only collocate** (KOC). In this example it could conceivably be *personality*, which co-occurs with *sunny* but no other weather adjectives. A sentence that includes the KOC *personality* along with the key *sunny* is then selected from the corpus. All that remains is to delete the key from the sentence, and supply key, distractors and sentence to the student in an appropriate format, as shown in the "Cloze generation system architecture" diagram.

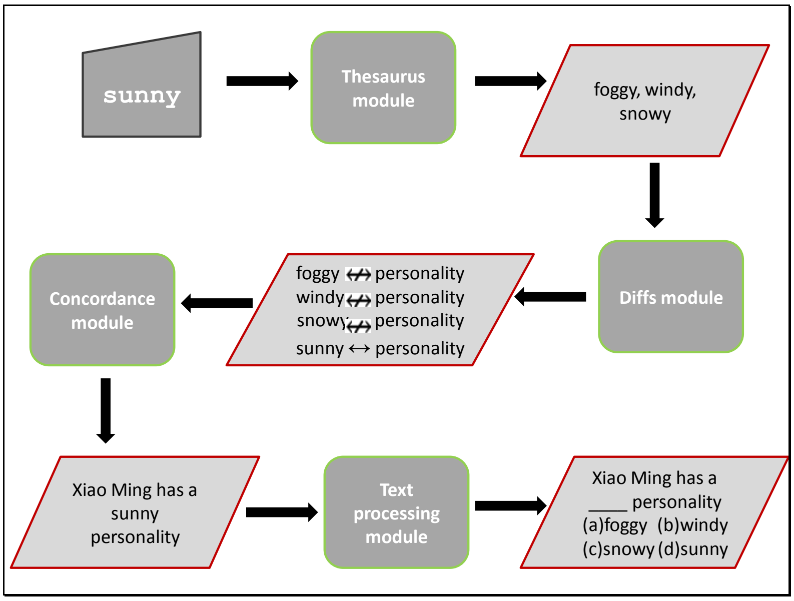


Figure 1: Cloze generation system architecture

Thus, the carrier sentence, the key and the three incorrect answers (distractors) are returned by the system. Subsequently, in the interactive mode, the teacher would be asked if they were satisfied with the item, whether they wanted to generate a new item using the same key, or whether they were happy with the sentence but would like to create a new set of distractors.

Here is an example of a cloze item actually generated by our system.

(1) *They have an enviable ____ of blue-chip clients.*

Ans: investment infrastructure asset portfolio

The learner is asked to complete the underscored gap with one of the four answers given.

The reader will agree that only the (key) answer *portfolio* is possible, and that if any of the three distractors were inserted, the sentence would become meaningless.

In this work, we make use of the Sketch Engine (SkE) suite of corpus query tools described by Kilgarriff et al (2004), and the ukWaC web corpus to which it provides access.

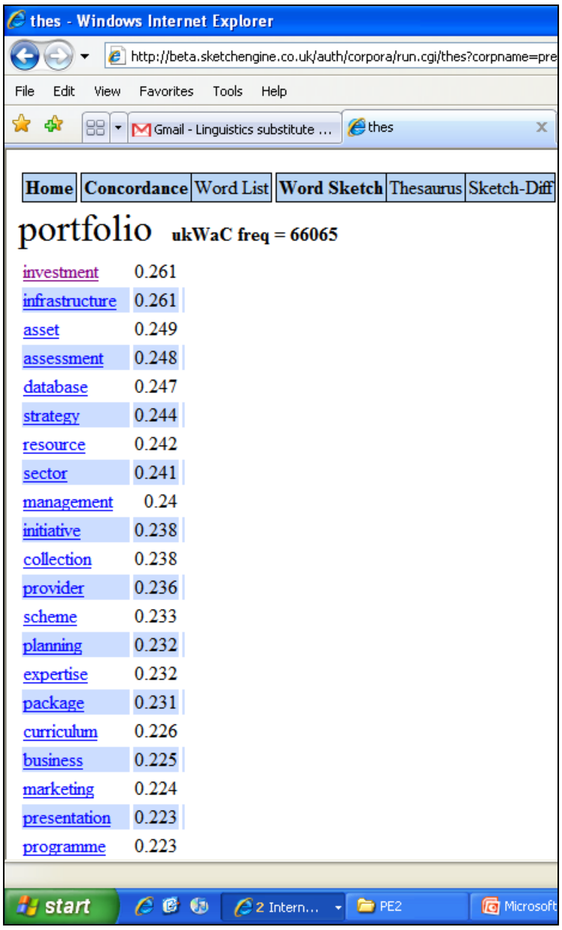
It needs to be made clear at this point that our system is not computationally implemented. The procedure for deriving the carrier sentences and distractors currently involves the manual implementation of rules which will be automated when we have the necessary time and resources available; we have taken care to set the system up in such a way that it can be readily programmed.

We now describe each step of the algorithm used for generating cloze items in detail.

Thesaurus Module

The Thesaurus module of SkE outputs words which typically occur in the same context as the search term. We show below the SkE Thesaurus output for *portfolio* (the key for the cloze item presented at (1) above). The screenshot reveals that most of the words with similar distribution to *portfolio* are in fact not synonyms or near synonyms: only *collection* and *package* qualify in that regard. A number of the words, as one might expect, have to do with business and the world of investment, with *investment* itself and *asset* ranking high on the list. The presence of the word *curriculum* on the list reflects the fact that the term *portfolio* is now widely used in the education domain.

The three top-ranking list members – *investment*, *infrastructure* and *asset* are noted and retained for use as PDs (potential distractors).


thes - Windows Internet Explorer
http://beta.sketchengine.co.uk/auth/corpora/run.cgi/thes?corpname=pre
File Edit View Favorites Tools Help
Gmail - Linguistics substitute ...
thes
Home
Concordance
Word List
Word Sketch
Thesaurus
Sketch-Diff
start
2 Intern...
PE2
Microsoft

portfolio ukWaC freq = 66065

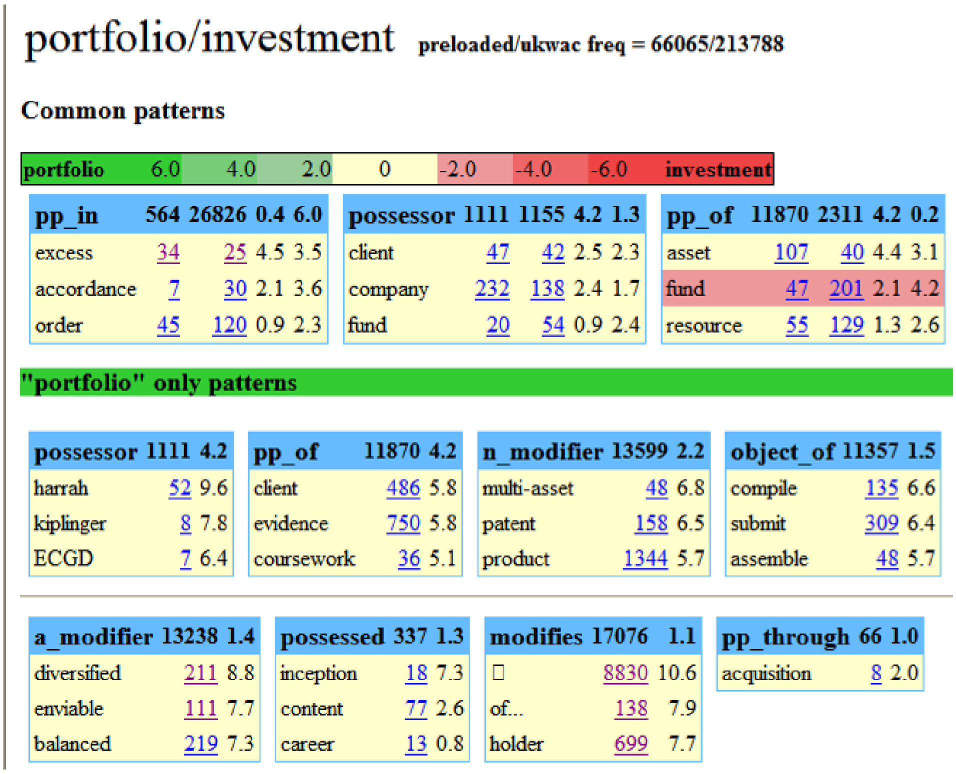
investment	0.261
infrastructure	0.261
asset	0.249
assessment	0.248
database	0.247
strategy	0.244
resource	0.242
sector	0.241
management	0.24
initiative	0.238
collection	0.238
provider	0.236
scheme	0.233
planning	0.232
expertise	0.232
package	0.231
curriculum	0.226
business	0.225
marketing	0.224
presentation	0.223
programme	0.223

Figure 2: SkE Thesaurus entry for *portfolio*

Sketch Differences Module

We next consult the Sketch Differences display. The screenshot below shows sketch differences for *portfolio* and *investment*, in contexts where either can occur in the ukWaC corpus. Notice how the display divides the output into grammatical relations between keyword and collocate. The screenshot shows us that *portfolio* occurs 34 times in a PP_IN relation with *excess*, while *investment* occurs in this collocation 25 times.

Typical contexts are "… an investment/ a portfolio in excess of *n* million dollars".



portfolio/investment **preloaded/ukwac freq = 66065/213788**

Common patterns

portfolio	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	investment

pp_in	564	26826	0.4	6.0
excess	34	25	4.5	3.5
accordance	7	30	2.1	3.6
order	45	120	0.9	2.3

possessor	1111	1155	4.2	1.3
client	47	42	2.5	2.3
company	232	138	2.4	1.7
fund	20	54	0.9	2.4

pp_of	11870	2311	4.2	0.2
asset	107	40	4.4	3.1
fund	47	201	2.1	4.2
resource	55	129	1.3	2.6

"portfolio" only patterns

possessor 1111	4.2
harrah	52 9.6
kiplinger	8 7.8
ECGD	7 6.4

pp_of	11870	4.2
client	486	5.8
evidence	750	5.8
coursework	36	5.1

n_modifier	13599	2.2
multi-asset	48	6.8
patent	158	6.5
product	1344	5.7

object_of	11357	1.5
compile	135	6.6
submit	309	6.4
assemble	48	5.7

a_modifier	13238	1.4
diversified	211	8.8
enviable	111	7.7
balanced	219	7.3

possessed	337	1.3
inception	18	7.3
content	77	2.6
career	13	0.8

modifies	17076	1.1
□	8830	10.6
of...	138	7.9
holder	699	7.7

pp_through	66	1.0
acquisition	8	2.0

Figure 3: Part of Sketch Differences entry for *portfolio* and *investment*

Of course, we are interested in situations where the two words do not share a collocate, and for this we glance down at the "portfolio only" patterns. Alongside each collocating word, in the Sketch Differences screenshot, is shown the frequency of the collocation (an underlined integer) and the **salience** (an index of the number of times *portfolio* occurs with the collocating word, as opposed to other words, given to one decimal place).

We now search for the collocate appearing only with *portfolio* (and never with *investment*) with the highest salience. We apply the condition that the collocate must be a correctly spelled English word, not a proper name. Thus, the non-alpha character □ with salience of 10.6 is rejected, as is *harrah*, a proper name (salience 9.6). The third-ranking in salience (8.8), *diversified*, is selected, and marked as a potential Key Only Collocate (KOC).

We next consider the second PD, *infrastructure*. The potential KOC *diversified* also does not occur in ukWaC in collocation with this PD, so it remains a candidate. However, when we move on to consider the third PD, *asset*, we find that *diversified assets* does indeed occur in the corpus. This means that *asset* cannot be used as a distractor for the key *portfolio* in the context *diversified portfolio*.

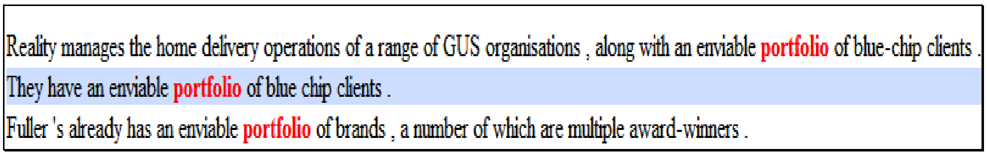
We therefore go on to consider the collocate appearing only with *portfolio* with the fourth highest salience: this turns out to be *enviable*. This time, we find that the potential KOC does not occur in collocation with any of the PDs, so it is adopted as KOC.

So far, we have decided on the key, as well as the three distractors. We have also established that we wish our carrier sentence to include the collocation *enviable portfolio.* The next step is to determine what the carrier sentence will be: we do this by consulting a concordance.

Concordance Module

The SkE concordancing software is equipped with a feature called GDEX (Husak et al, forthcoming) which favours sentences which are between 10 and 25 words long, containing only common words, and some other related constraints. GDEX sorts the order in which concordance sentences are presented, so that optimal sentences appear first. This means that the sentences which are most likely to be selected for dictionary examples or cloze exercises appear conveniently at the beginning of the concordance display.

From the concordance output from which the screenshot below is taken, we may now extract the sentence shown at (1) above. Note that if the user is dissatisfied with the first sentence, for any reason, they can be prompted to select the second or a subsequent sentence.



Reality manages the home delivery operations of a range of GUS organisations , along with an enviable portfolio of blue-chip clients .

They have an enviable portfolio of blue chip clients .

Fuller 's already has an enviable portfolio of brands , a number of which are multiple award-winners .

Figure 4: Part of SkE concordance entry for *portfolio* and *enviable*

BNC cloze example

In our experiments, we also generated (2), this time from the British National Corpus. Again, the correct answer choice is supposed to be *portfolio*.]

(2) *Albert E Sharp Fund Managers have launched AES European unit trust, which seeks long-term capital growth from a diversified _____ of European Securities.*

Ans: asset portfolio stock holding

Unlike ukWaC, the corpus used to generate (1), the BNC does not contain any examples of the adjective *diversified* modifying any of the PDs. However, the concept of a "diversified **holding** of European Securities" does seem quite plausible; given two apparent possible answers, it is unlikely that many teachers would find (2) an acceptable cloze exercise.

The way in which the BNC was compiled means that it consists mostly of clean text, and relatively little noise, while ukWaC contains a fair amount of duplication and non-textual data. This might be taken as a compelling argument for preferring the BNC as a source corpus. However, the GDEX software does a good job of ensuring that the most meaningful sentences from a ukWaC concordance are presented first. What is more, if we posit that certain collocations have a vanishingly small chance of occurring – and that is the claim that one makes when setting the distractors for a cloze exercise – we should be using the very largest corpus available. The larger the corpus, the more exhaustive the evidence; and the less likely the system will be to generate unwanted **correct** distractors, such as *holding* in (2) above.

Next steps

We have described an algorithm which is capable of generating a carrier sentence and distractors, given a user-supplied key (correct answer). We have shown how modules of the Sketch Engine corpus query tool can be used to generate these components.

As mentioned above, we will shortly prepare an implementation of the algorithm that will allow a user to supply a key at a computer, and be presented with a suggested cloze item. If the item is not satisfactory, the user will be able to run the program again and generate a new exercise.

Beyond straightforward programming, some work will be necessary to ensure that distractors match the key in terms of inflectional morphology (plural –s and the like). A review of any copyright issues involved will also be necessary.

Once implemented, this work can be put to good use immediately. Teachers who use the program will be able to generate authentic cloze items in very short order. As mentioned above, by supplying as input a list of vocabulary items pertinent to the topic of a unit or lesson, such as the “Business” or “Getting started at university” lists described in Smith et al (2008), it will be possible to produce a set of highly relevant cloze exercises. These exercises can be used for assessment, or simply as part of day to day teaching, making students aware of the collocational patterns in which the topic vocabulary commonly participates. The exercises can be used in class, in the lab, or at home, and could be incorporated into an interactive CALL interface, making students’ learning experience more enjoyable and fruitful.

References

Alderson, J. C. 1978. “A study of the cloze procedure with native and non-native speakers of English.” Doctoral dissertation, University of Edinburgh.

Baroni, M., Kilgarriff, A., **Pomikálek, J.** and **Rychlý, P.** 2006. “WebBootCaT: instant domain-specific corpora to support human translators.” Paper presented at *EAMT 2006*, Oslo, 247-252.

Hoshino, A. and **Nakagawa, H.** 2007. “Assisting cloze test making with a web application.” Paper presented at the *Society for Information Technology and Teacher Education International Conference 2007* (pp. 2807-2814). Chesapeake, VA: AACE.

Husak, M., Kilgarriff, A., McAdam, K., Rundell, M. and **Rychlý, P.** Forthcoming. “GDEX: Automatically finding good dictionary examples in a corpus.” Paper to be presented at *EURALEX*, Barcelona. July 2008.

Jonz, J. 1990. “Another turn in the conversation: What does cloze measure?” *TESOL Quarterly* 24(1): 61-77.

Kilgarriff, A., Rychlý, P., Smrž, P. and **Tugwell, D.** 2004. “The Sketch Engine.” Paper presented at *EURALEX*, Lorient, France. July 2004.

Mostow, J., Beck, J. E., Bey, J., Cuneo, A., Sison, J., Tobin, B. and **Valeri, J.** 2004. “Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions.” *Technology, Instruction, Cognition and Learning* 2: 97-134

Oller, J. W., Jr. 1973. “Cloze tests of second language proficiency and what they measure.” *Language Learning* 23: 105-8.

Smith, S., Sommers, S. and **Kilgarriff, A.** 2008. “Learning words right with the Sketch Engine and WebBootCat: Meaningful lexical acquisition from corpora and the web.” Paper presented at the 2008 *CamTESOL conference*, Phnom Penh.

Sumita, E., Sugaya, F. and **Yamamoto, S.** 2005. “Measuring Non-native Speakers’ Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-BlankQuestions.” Paper presented at the *2nd Workshop on Building Educational Applications using NLP*, Ann Arbor.

Taylor, W. L. 1953. “Cloze procedure: A new tool for measuring readability.” *Journalism Quarterly* 30: 415-433.

USING CORPORA AT SECONDARY SCHOOLS:
TEACHING LITERATURE AND LINGUISTIC KNOWLEDGE

Bettina Fischer-Starcke[212]

Abstract

Analysing literary texts by means of corpus linguistic techniques provides researchers with information on the text that cannot be seen with the naked eye, but which serve as the basis for the texts' literary interpretations. The analytic procedure contrasts with that by literary critics who rely on their intuitions and subjective perceptions of textual features for their interpretations.

As the discussion and interpretation of literary works form a major part of advanced language classes at secondary schools, introducing corpus linguistic techniques into the classroom allows for not only discussing literary meanings, but also for introducing learners to aspects of the organisation of language and for making it explicit. These include, for instance, the difference between lexical and grammatical words and their frequencies in language. Linking literary and linguistic topics in the classroom sharpens the learners' awareness of their correlation and of the interdependence between form and meaning in language.

After a theoretical introduction, this paper presents concrete ways of using corpus linguistic techniques in secondary school foreign language teaching. This includes the use of wordlists, keywords and concordance lines. This is followed by a discussion of benefits and challenges from this type of classroom activities for learners and teachers of a language, including practical requirements of the approach.

Keywords: Corpus linguistics, foreign language teaching, literature, language, learning

Introduction

Foreign language classes at an advanced level at secondary schools frequently analyse literary works in the target language for their literary meanings. This aims (1) at familiarizing learners with literary texts and developing literary interpretations of the text and (2) at consolidating and furthering the learners' language competences of the target language. While these two goals co-exist, classroom activities frequently make only the first of them explicit.

Using corpus linguistic data extracted from a literary text in foreign language teaching (FLT), facilitates and makes explicit knowledge on the specific text and on the target language in general. The latter includes for instance insights into linguistic basics such as the distinction between lexical and grammatical words and their frequencies in language. This furthers the learners' understanding of the language and supports both reception and, as a consequence, production of the language. Also, based on this awareness of general language patterns, specific patterns of the text in question can be identified and analysed for their functions. These form the basis for a literary interpretation of the text and for recognizing the correlation between form and meaning in language.

The approach to FLT suggested here fulfils the two main functions of language teaching classes: (1) making learners aware of linguistic patterns and furthering the learners' competences in the language and (2) providing a literary interpretation of a literary text. In addition, it furthers learners' technical competences by working with corpora and specialised software and motivates learners through independent and exploratory work with the data.

[212] Bettina Fischer-Starcke is Assistant Professor for English Business Communication at the Vienna University of Economics and Business Administration, Austria. She studied English and German for her MA and the German secondary school teaching diploma at Trier, Germany, and English at Palmerston North, New Zealand. In Palmerston North, she received a Postgraduate Diploma in English. In 2008, she completed her PhD on corpus stylistics in Trier. Her main research interests are corpus linguistics, co rpus stylistics, phraseology and the application of these fields to foreign language teaching. Her publications are articles on all of the above topics. The publication of a monograph on corpus stylistics with Continuum is scheduled for 2009.

In the following, the approach is illustrated by an analysis of Jane Austen's novel *Northanger Abbey* (1818, *NA* in the following) by means of a wordlist, an analysis of a list of keywords and of concordance lines. First, I will show how this data allows for conclusions on the linguistic organisation of the language and how linguistic patterns encode the text's literary meanings. Second, I will discuss benefits from and challenges posed by these classroom activities for learners and teachers. This includes the technical requirements.

Corpus linguistics and foreign language teaching

For most linguists, the greatest use of corpus linguistics for FLT seems to be its function as a source for teaching materials. Flowerdew (2001) for instance describes the design of corpus-based EAP teaching materials on the basis of insights into learner problems gained through learner corpora. Also the compilation of EFL dictionaries, as described by Gillard and Gadsby (1998), profits from understanding learners' needs through the analysis of learner corpora and the compilation of entries based on general corpora. The use of corpora for compiling a grammar for ELT is discussed by Mindt (2002).

Confronting students with corpora seems to be restricted to FLT at universities. For instance Mair (2002), Davies (2000) and Bernardini (2004) discuss their use of corpora in ELT at universities. They either confront learners directly with a corpus to allow them to extract their own data for analyses or they use data which has been extracted by the teacher. The direct access to linguistic data through corpora shows learners the connection between form, meaning and language use without any external teaching (Johns 1990). This furthers the learners' motivation and supports inductive learning.

The use of literary texts for corpus linguistic analyses at universities has been described by Louw (1997) and Jackson (1997). I discuss approaches for using them at secondary schools elsewhere (Starcke 2007).

Practical Approaches

Linguistic and literary insights into the target language and a specific literary text can be gained by using different corpus linguistic techniques. The most feasible ones for secondary school teaching are wordlists, keywords and concordance lines as using these techniques is relatively easy to learn for both teachers and learners. Also, the analysis of the data allows for results even from inexperienced analysts and for achieving different learning goals. This will be demonstrated in the following.

There are two main reasons for choosing Austen's *Northanger Abbey* as an example text. First, novels by Austen are frequently read and discussed in English classes. They are classics of English literature and their language is relatively easy to understand for learners of English. Second, access to *NA* is not restricted by copyright. It can be stored and analysed electronically – a necessity for any corpus analytic or corpus stylistic analysis.

Wordlists

The analysis of a wordlist for *NA* which is ordered according to frequency reveals that its first 62 words are grammatical words. The first lexical word, *said*, occurs as number 63 on the list. This can be detected by the pupils and be used in the classroom in order to discuss

- the organisation of language into two categories of words, lexical and grammatical words,
- their respective frequencies and
- the reasons for their frequencies in the text and in language in general.

This introduces learners to one of the basic principles of the lexical organisation of language.

Keywords

Keywords (cf. Scott 1999) are words which occur statistically more frequent in a text or corpus than in a reference corpus. This frequency is due to their importance for the content of the text as they indicate what a text is about (Scott 2002). For the present purpose, *NA* was compared with a corpus of literature contemporary to Austen (circa

4,370,000 tokens) and the resulting list of keywords serves as the basis for the following analysis (see Fischer-Starcke forthcoming for the keywords and for information on the corpus).

On the list of 130 positive keywords, i.e. keywords as defined above, there are four occurrences of words from the semantic field *emotions* (*feelings, engagement, attentions, admiration*). Two conclusions can be drawn from that. First, semantic fields are a second feature of lexical organisation. Apart from grammatical categories, also semantic features are used to classify words into groups. The fact that the same word can be classified into different groups depending on the criteria used, e.g. a lexical word is also part of a semantic field, highlights that language is a network which encodes meaning on different linguistic levels. Second, the occurrence of the semantic field confirms linguistically the literary position that emotions are a prominent topic in the novel (cf. for instance Bergmann 2002, Brooks & Watson 2000 and Litvak 1996). This knowledge serves as a starting point for the literary analysis of the text which is performed by analysing concordance lines of the keywords.

Concordance lines

For illustrative purposes and due to limitations of space, only a limited number of observations on the concordance lines will be discussed. The concordance lines were extracted for the lemmatized forms of the keywords.

*FEELING**

The lemma FEELING* occurs 66 times in the text with 39 instances describing and discussing the nature of somebody’s feelings and the strength of one’s feelings (14), e.g.:

ss Tilney, and poured forth her joyful **feelings**. It was doomed to be a day of

in a faltering voice. "Alas! For my **feelings** as a daughter, all that I know,

for her friend seemed rather the first **feeling** of her heart; but that at such

only replied, "I cannot wonder at your **feelings**. I will not importune you. I

aintance with her, softened down every **feeling** of awe, and left nothing but te

nature of FEELINGs

s just the same; he has amazing strong **feelings**. Good heavens! What a delight

hing to change them. But I believe my **feelings** are stronger than anybody's; I

illing words, and wound up Catherine's **feelings** to the highest point of ecstasy

y afford a return. The strength of her **feelings** she could not express; the natu

ith a curiosity so justly awakened, and **feelings** in every way so agitated, repo

strength of FEELINGs

Especially the novel’s female protagonists frequently discuss their own and other people’s feelings, which are often of a romantic nature, and their strength. These are intertextual references to sentimental novels, a genre satirized by *NA*.

The second genre to be satirized by *NA* is gothic novels. This can also be seen from collocations of FEELING*. While positive feelings are expressed in six instances, FEELING* also co-occurs with negatively connotated and denotated words and expressions in 15 instances, e.g.:

ry. She had no power to move. With a **feeling** of terror not very definable, sh

ter, oppose the connection, turned her **feelings** moreover with some alarm toward

ing whomsoever they chose, without any **feeling** of humanity or remorse; till a v

knew her beloved Catherine to have so **feeling** a heart, so sweet a temper, to b

ther, so pure and uncoquettish were her **feelings**, that, though they overtook an

erness, was a sight to awaken the best **feelings** of Catherine's heart; and in th

negative and positive FEELINGs

Collocates such as *terror, alarm* and *remorse* indicate the intertextual reference to gothic novels.

In both sentimental and gothic novels, the protagonists examine their own and other people's feelings and fears. These intertextual references to the genres can be extracted from the analysis of concordance lines and they can serve as starting points to a discussion of e.g. satire in *NA*. The findings can be supported by further evidence of satire gained from further corpus linguistic analyses or by findings from literary critics.

*ATTENTION**

The lemma ATTENTION* occurs 40 times in the novel and is used to explicitly describe the attention paid to women in 19 instances, e.g.:

astonishment. Isabella talked of his **attentions**; she had never been sensible
Yes, very sure." "Is it my brother's **attentions** to Miss Thorpe, or Miss Thorp
benevolence; thanking him for such an **attention** to her daughter, assuring him
been very remiss, madam, in the proper **attentions** of a partner here; I have noo her was that he paid her rather
more **attention** than usual. Catherine had nev

ATTENTION* to women

Descriptions of a woman paying attention to a man occur only four times in the data. The fact that it is mostly a man who pays attention to a woman allows for insights into courtship rituals at the time of the novel and their representation in sentimental and gothic novels.

*ADMIRATION**

While the previous analyses showed a focus on human relationships, the analysis of ADMIRATION* (19) also indicates the importance of material objects in the novel. While six instances of ADMIRATION* co-occur with people, five instances describe somebody's admiration for objects, e.g:

t be our heroine's opinion of him, his **admiration** of her was not of a very dan
. No man is offended by another man's **admiration** of the woman he loves; it is
t be our heroine's opinion of him, his **admiration** of her was not of a very dan
had ever been at, and looked with great **admiration** at every neat house above th
in Gothic ornaments, stood forward for **admiration**. The remainder was shut off
after Thorpe had procured Mrs. Allen's **admiration** of his gig; and then receiving

ADMIRATION* of people and objects
The nearly equal admiration of people and objects indicates criticism of Bath, the setting of that part of the novel in which mostly objects are admired. This can serve as a starting point for discussing the role of Bath for English society at Austen's time.

*ENGAGEMENT**

For many learners of English, the primary lexical meaning of *engagement* is that of the stage after accepting a marriage proposal and before the wedding. But upon analysing the concordance lines, learners will see that this is only one meaning of the word, with the other meaning, namely that of an appointment, being more prominent in 19th century English. Thirteen of the 30 instances of ENGAGEMENT* in the text describe an appointment, 12 refer to marriage plans, e.g.:

therine away, when he recollected this **engagement**," said Sarah, "but why not do

ld, "do not be so distressed. A second **engagement** must give way to a first. I

t you had just been reminded of a prior **engagement**, and must only beg to put of

wer of refusal; that in both, it is an **engagement** between man and woman, formed

can your brother mean? If he knows her **engagement**, what can he mean by his beh

. Once or twice indeed, since James's **engagement** had taught her what could be

ENGAGEMENT*

This finding can be used (1) for vocabulary training and (2) as an indicator of the importance of romantic and social relationships in the novel. While the first is a linguistic topic, the latter leads to literary interpretations of the novel and cultural information on 19th century England.

As shown above, findings from the analysis of concordance lines can be used for several purposes in FLT:

(1) to detect literary meanings, e.g. relationships between protagonists and the importance of material objects,

(2) to discuss intertextual references between the novel and the two genres it satirizes. This includes the role of emotions in the novel and their literary functions,

(3) to gain cultural knowledge on 19th century England and

(4) to gain linguistic knowledge, e.g. on word meaning.

Also biographical parallels to Austen's life, e.g. her time in Bath, can be drawn.

Benefits and Challenges

The analysis of literary texts by using corpus linguistic techniques in secondary school FLT has the double advantage of realising what not only many pupils but also school curricula expect. This is (1) the discussion of literary works and their meanings and (2) improving the learners' linguistic competences in the target language.

But there are further benefits resulting from the approach to FLT presented above. The close reading process and the search for patterns in the data trains analytic thinking, furthers reading competences and is likely to enlarge the learners' personal lexicon as they are likely to look up words in a dictionary. The activities proposed are highly motivating for learners as they involve creative, exploratory and independent work on the data. The use of modern technologies, i.e. the computer and software, is also likely to further their motivation and media competences.

While these are benefits of the approach, it is also highly demanding of learners and teachers. Their proficiency in the target language must be high in order to be able to recognize linguistic patterns in the data. In addition, they have to think creatively and analytically, skills that are only infrequently trained at secondary schools. Teachers also need knowledge on corpus linguistic theory and its analytic techniques so that they can answer questions and explain the use of the analyses to the learners.

Not only the learners, but also teachers benefit from the approach. The choice of data, of teaching aims and of the lessons' linguistic or literary focus are left to them. Furthermore, it is the teachers' choice whether they prepare the data so that learners will retrieve mostly expected results or whether the target is exploratory learning with learners extracting and analysing their own data. This choice leaves teachers freedom in their lesson design.

Practical requirements for this type of classroom activity are access to a computer, to the relevant software and to the data. Both data and software can be retrieved from the internet, in many cases free of charge (e.g. Project Gutenberg provides free literary texts and see Lee's website for software and data collections).

Conclusion

Analysing literary texts by means of corpus linguistic techniques has great potentials for FLT. It allows for gaining linguistic and literary insights into the data and the language system and it trains skills such as analytic and creative thinking, working independently and working in an exploratory way. While the proposed approach to FLT poses challenges for both learners and teachers, the benefits gained from it outweigh them. By treating both linguistic and literary issues, it overcomes the traditional dualism between them that is manifest in secondary school education. It demonstrates to the learners that form and meaning are interdependent in language, one of the basic linguistic principles. This provides insights into how meaning is created in language and challenges learners to question the distinction between literature and its interpretation from language learning as it is frequently presented in schools. This allows learners to discover potentials of language and its creative use in literature while at the same time furthering their language competences.

References

Austen, Jane. 1818. *ibiblio* *P2P*. [Northanger Abbey]. eBook #121. April, 1994. [last update: 4. August 2002]. http://www.gutenberg.net/etext/121 [access date 24/05/2008]

Bergmann, Jenna R. 2002. "Romantic anti-dualism and the blush in *Northanger Abbey*". *Wordsworth Circle* 33/1: 43-47.

Bernardini, Silvia. 2004. "Corpora in the classroom. An overview and some reflections on future developments." In *How to Use Corpora in Language Teaching*, John McH. Sinclair (ed.). Amsterdam, Philadelphia: John Benjamins, 15-36.

Brooks, Marilyn and **Watson, Nicola**. 2000. "*Northanger Abbey*: contexts." In *The Nineteenth-Century Novel: realism*, Delia da Sousa Correa, Dennis Walder, Stephen Regan (eds.). London: Routledge, 62-86.

Davies, Mark. 2000. "Using multi-million word corpora of historical and dialectal Spanish texts to teach advanced courses in Spanish linguistics." In *Rethinking Language Pedagogy from a Corpus Perspective*. *Papers from the third international conference on teaching and language corpora*, Lou Burnard and Tony McEnery (eds.). Frankfurt/Main etc: Peter Lang, 173-185.

Fischer-Starcke, Bettina. Forthcoming. *Corpus Linguistics and the Study of Literature*. London: Continuum.

Flowerdew, Lynne. 2001. "The exploitation of small learner corpora in EAP materials design." In *Small Corpus Studies and ELT. Theory and practice*, Mohsen Ghadessy, Alex Henry and Robert L. Roseberry (eds.). Amsterdam, Philadelphia: John Benjamins, 363-379.

Gillard, Partick and **Gadsby, Adam**. 1998. "Using a learners' corpus in compiling ELT dictionaries." In *Learner English on Computer*, Sylviane Granger (ed.). London, NY: Longman, 159-171.

Jackson, Howard. 1997. "Corpus and concordance: Finding out about style." In *Teaching and Language Corpora*, Anne Wichmann et al. (eds.). London, NY: Longman, 224-239.

Johns, Tim. 1990. "From printout to handout: grammar and vocabulary teaching in the context of data-driven learning." *CALL Austria* 10: 14-34.

Lee, David. Bookmarks for Corpus-Based Linguists. http://devoted.to/corpora [access date 24/05/2008]

Litvak, Josef. 1996. "Charming men, charming history." In *On your Left: The new historical materialism*, Ann Kibbey et al. (eds.). NY: NY UP, 248-274.

Louw, Bill. 1997. "The role of corpora in critical literary appreciation". In *Teaching and Language Corpora*, Anne Wichmann et al. (eds.). London, NY: Longman, 240-251.

Mair, Christian. 2002. “Empowering non-native speakers: the hidden surplus value of corpora in continental English departments.” *Teaching and Learning by Doing Corpus Analysis. Proceedings of the fourth international conference on teaching and language corpora, Graz 19-24 July, 2000*. Bernhard Kettemann and Georg Marko (eds.). Amsterdam, NY: Rodopi, 119-130.

Mindt, Dieter. 2002. “A corpus-based grammar for ELT.” In *Teaching and Learning by Doing Corpus Analysis. Proceedings of the fourth international conference on teaching and language corpora, Graz 19-24 July, 2000*. Bernhard Kettemann and Georg Marko (eds.). Amsterdam, NY: Rodopi, 91-105.

Projekt Gutenberg. http://www.gutenberg.org [access date 24/05/2008]

Scott, Mike. 2002. “Picturing the key words of a very large corpus and their lexical upshots or Getting at the *Guardian*’s view of the world.” *Teaching and Learning by Doing Corpus Analysis. Proceedings of the fourth international conference on teaching and language corpora, Graz 19 – 24 July, 2000*, Bernhard Kettemann and Georg Marko (eds.). Amsterdam, NY: Rodopi, 43-50.

Scott, Mike. 1999. *Wordsmith Tools*. 3.0. OUP. [Computer Software]

Starcke, Bettina. 2007. “Korpuslinguistische Daten als Grundlagen von Literaturrezeption im Unterricht.“ In *Fremdsprachenforschung heute – Interdisziplinäre Impulse, Methoden und Perspektiven*. Sabine Doff and Torben Schmidt (eds.). Frankfurt/Main: Peter Lang, 211-224.

LEARNERS PUT CONCEPTUAL METAPHOR THEORY TO THE TEST

Scott Staton[213]

Abstract

The cognitive approach to metaphor has been investigated for several decades but as yet has had relatively little impact on foreign language teaching and learning. This paper aims to show how learners can test the theory empirically using corpora while reaping numerous language learning benefits.

In the first phase, students are introduced to selected classics in contemporary metaphor theory. The readings illustrate the potential of the theory for learners but they also reveal the importance of native speaker introspection in the building of the theory. The second phase introduces recent work that tests the theory empirically using corpora. The studies presented in this phase exemplify clearly both the methodology used in corpus research and the structure of a research article. Next, students do their first hands-on work with corpora, proceeding from the familiar (Google) to the new (the BYU interface to the BNC). The forth phase addresses the gap that needs to be bridged between the typically abstract conceptual metaphors (MORE IS UP) and their concrete linguistic realizations ("prices have skyrocketed"). The fifth and final phase is the research project. For this project students choose a conceptual metaphor from the literature or posit one of their own. They then assemble and select a limited set of lexical items for testing and search their words in the BNC and/or in another corpus. The findings are then written up in a paper which has the structure and the character of a research article.

The project is undoubtedly ambitious. Certainly students do not profit in equal measure from each phase, but the overall impression is that on the whole learners meet the challenge.

Keywords: Conceptual metaphor, corpus linguistics, autonomous learning, English, metaphor

The backdrop

Metaphor has come to occupy an increasingly central role in the study of language and thought. Once deemed a rhetorical figure, a piloted deviation from "literal" meaning, metaphor is now recognized as constituting and shaping meaning and even thought. The contribution of cognitive linguistics has been crucial to the contemporary discussion of metaphor and provides the backdrop for this paper.

The paper reports on a ten-week course held with advanced students in the intercultural language and literature degree program at the University of Florence. The main aim of the course is to introduce students to conceptual metaphor theory and to provide them with the analytical and methodological tools to test the theory autonomously using empirical data. The mother tongue of most students is Italian, but there is a significant minority of students from other European countries. All students have studied at least one other language. Most have little or no background in linguistics.

In the first phase, students are introduced to selected "classics" in conceptual metaphor theory (Lakoff and Johnson 1980, Lakoff 1987). The readings illustrate the explanatory power of the theory and suggest how learners could benefit from investing effort in the study of language and language use from a cognitive perspective. A central claim of conceptual metaphor theory is that metaphor is not so much a rhetorical trope or a linguistic device as a way of conceptualising the world that surrounds us. For example, on this view the phrase "high prices"

[213] Scott Staton received a BA from Indiana University and an MA from the University of Chicago, the latter involving a two-year scholarship at the Goethe Universität in Frankfurt a.M. . His is currently lecturer at the Faculty of Letters and Philosophy of the University of Florence. He has also taught at the University of Bologna and the European University Institute, in San Domenico, Fiesole. He is member of the editorial board of a collection of practical volumes for university level language instruction entitled Lingue d'europa: materiali didattici per lo studio delle lingue straniere (published by Allinea). Moreover, he contributes to the organization of an annual conference on university level language teaching held in Florence called "Giornate di studio sui materiali didattici per l'insegnamento delle lingue". His current research interests include applications of cognitive linguistics to language learning and teaching, academic writing, and American roots music.

is not simply a high frequency collocation but rather a reflection of the underlying conceptual metaphor MORE IS UP. In other words, quantity (the target domain) is thought of in terms of verticality (the source domain). Similarly, the familiar expression "What emerges from all this" is a linguistic instantiation of SEEING IS UNDERSTANDING. The first step, then, is to motivate learners to make unexpected connections, to see the familiar from a new perspective.

The central text in this phase is Lakoff's (1987) case study "Anger". The study begins with an array of seeming unrelated expressions, including:

> He's just letting off steam.
>
> Don't get a hernia.
>
> Try to keep a grip on yourself.

Lakoff goes on to present "a common folk theory of the physiological effects of anger" (1987: 381) and then unfolds an analysis which reveals the interrelationship between of a number of conceptual metaphors, including THE BODY IS A CONTAINER FOR THE EMOTIONS, ANGER IS HEAT and ANGER IS A DANGEROUS ANIMAL. The case study "Anger" is suitable for several reasons. First of all, it constitutes an impressive illustration of the explanatory power of the theory. A great many often apparently idiomatic expressions are listed, discussed, and given a sense of order. Moreover, this wealth of language is in itself a payoff for learners, especially as it appears in ordered chunks. It is also worth noting that the topic of anger is experientially relevant to young learners, probably more so than the sister case studies on *over* and *there*-constructions. Finally, an early focus on the conceptual content of emotions paves the way for students' projects on other (more attractive) emotions.

While the readings at this point are certainly eye-opening for learners unfamiliar with conceptual metaphor theory, as non-native speakers they tend to be somewhat overwhelmed by the sheer quantity of unfamiliar language. After all, the theory does not deal with expressions like "That got me angry" but rather "That pissed me off" or "Mom's going to have a cow when she hears about this." The inability of non-native speakers to pull up language like this is all too evident. This issue is addressed in the second phase.

Phase 2: A corpus-linguistic approach

The second phase introduces learners to work that tests the theory empirically using corpora. The studies presented in this phase (esp. Deignan and Potter 2004) exemplify clearly both the methodology used in corpus research and the structure of a research article, which students can later use as a model for their own writing. In this approach a corpus linguistic perspective is applied to conceptual metaphor by investigating the source domain. If the source domain of many central metaphors is, as the theory claims, bodily experience, it follows that names of parts of the body (nose, mouth) are likely candidates for metaphorical sources domains. As the lexis pertaining to the human body is straightforward and limited, this is a natural starting point for a corpus search. Results are analysed manually and many intriguing problems concerning the confines between metaphor and metonymy surface quickly.

Phase 3: Hands-on work with a corpus

The next phase involves students in their first hands-on work with corpora. A series of lab sessions is run to allow them to experiment with searches of various kinds. We start with Google. Students begin by conducting a series of piloted searches aimed at revealing how the search engine can be used not only to find information but also to gather data about language forms and use. One such search seeks to collect data on the countability of nouns like 'energy' and 'knowledge.' It becomes quickly evident that the data base includes instances of the plural of both nouns that are either not explained or not even contemplated in learner's dictionaries. The aim of these initial activities is to show learners how they can discover on their own instantiations of linguistic phenomena that require explanation.

The Google search also produces numerous unnatural and ill-formed expressions, which are not always immediately obvious as such to learners, and reveals the weaknesses of the search engine as a corpus tool. The perceived need for a more reliable tool is answered by the British National Corpus. Two different interfaces are presented. The first, from the BNC homepage, allows the user to examine 50 random hits of any given word form. The advantages over Google are immediately obvious. Samples are given in complete sentences, and there is no risk of calling up unreliable data. The second, hosted by Brigham Young University (BYU), offers a wide range of search options (indeed, probably too many for most learners at this level). The activities that follow are aimed primarily at allowing learners to familiarize themselves with the mechanics and potential of the BYU interface. The exercises include repeating the search carried out with Google, trying out collocation searches, and testing Deignan and Potter's results from Phase 2.

Phase 4: Preparing learners for autonomous learning

The aim of the research project is to verify whether the conceptual metaphors cited in the literature, or those hypothesized by students, find empirical support in a corpus. There are various ways to search a corpus for manifestations of conceptual metaphor. (See Stefanowitsch 2006a for an overview.) One is to search for vocabulary directly connected with the target domain. Stefanowitsch (2006b) has demonstrated the effectiveness of this method in his challenge to the introspective method adopted by Lakoff and others. Thus, a search for "anger" will produce "direct/target anger at," "mounting anger," and so forth. Another approach is to search for lexis in the source domain. In either case it is useful for the student-researcher to have available a lexical set for investigation. The forth phase deals with the task of gathering a set of lexemes for the corpus search.

Desktop resources such as dictionaries and thesauruses are the obvious starting point. Monolingual learners' dictionaries have become familiar companions for practically all university students. Thesauruses designed for native speakers, on the other hand, can be overwhelming and frustrating, depending on the headword. The web also offers some innovative modes of revealing and organizing relationships between words and ideas. One that many learners enjoy using is the thinkmap Visualthesaurus. A clear advantage of this resource is that it does not simply list vocabulary items but groups them graphically according to sense. Another is that the number of items is generally limited and therefore manageable. A different approach to semantic relations is provided by WordNet. WordNet gives not only synonyms and antonyms but also hyponyms. Thus if we want to determine which flowers serve as a source domain to talk about people (a subclass of the metaphor PEOPLE ARE PLANTS), we can quickly generate a list of species of flowers and either search them manually or, using the BYU interface, automatically in the BNC.

Phase 5: The research project

The fifth and final phase is the research project. Students may choose a conceptual metaphor from the literature or posit one of their own. A good place to start is the Master Metaphor List. This is a long list of conceptual metaphors with a few examples of linguistic realizations for each posted by G. Lakoff on the Berkley cognitive science website. Since the examples given are on the whole quite limited, learners are faced with the problem of identifying and collecting suitable source domain vocabulary for empirical investigation, which is in itself a fruitful learning task. Throughout the research phase, it should be stressed, learners need support and access to consultation. This need regards choosing an appropriate metaphor, assembling a lexical set for examination, collecting data, and, above all, analysing the data.

It is not possible to foresee all the ways in which conceptual metaphors reveal themselves in the language, but a few guidelines can help. Once the source domain has been chosen, it is useful to consider several ways in which a conceptual metaphor can be realized linguistically. If we take PEOPLE ARE PLANTS as an example, an obvious place to start is with plant varieties. In the subclass of flowers, *wallflowers* and *pansies* come to mind, in that of fruit, *peaches* and *nuts*. Attributes are also a source of mappings onto the target domain. For instance *green*, *sour, withering* are all commonly used metaphorically. What is less obvious to learners is how the source domain entity "behaves" or interacts in the world. If a flower "blossoms" or "wilts," it changes state. Indeed, the change of state verbs, or ergative verbs, used to describe natural processes are a common source for mappings. The interaction of entities with other entities in the source domain may be quite straightforward. For example, *a hard nut to crack* rather clearly implies an agent who might attempt to crack a nut/convince a person. In expressions like *sour grapes* or *rotten apple,* on the other hand, it might on the surface appear that we are dealing with attributes, but actually the attributes motivate effects. In the first case, the mapping involves the attitude of a person in the target domain, and in the second the effect that a rotten apple has on other apples in the source domain, and by extension one person on other people in the target domain. Finally, it is also worth pointing out to learners that these three categories correspond rather neatly to nouns, modifiers, and thematic relations in argument structure.

Once the research has been carried out, the findings are written up in a paper which has the structure and the character of a research article. The choice of the paper, as opposed to the oral exam or oral presentation, is dictated mainly by the fact that the students who have taken the course over the last few years have had practically no experience in writing academic papers (in any language). The results are predictably varied. Those who reach the main goal of the project are able to demonstrate that a given conceptual metaphor is not merely the fruit of the fantasy of some theoretician but finds empirical confirmation in the corpus data. Som, however, do not achieve this aim. The main obstacles seem to be the unwillingness to abandon the familiar notion of metaphor as rhetorical figure, the difficulty in managing the mass of decontextualized language extractable from the corpus, and, most unfortunately, the sense of disorientation that can accompany autonomous learning.

There are, of course, students who become deeply motivated and produce thoughtful and original work. Here are a few theses that students have advanced and supported using corpus data.

- Food and food preparation is a far richer and far more commonly exploited source domain in French than in English. (Heather)
- The metaphor HUMAN RELATIONSHIPS ARE HANDCRAFTED OBJECTS is widely attested not only in the BNC but also in American sign language. (Manuela)
- The collocations connected with AFFECTION IS WARMTH suggest that affection is not conceptualised as simply a less intense state of love but rather qualitatively different, involving a stronger element of reciprocity. (Sonia)

Conclusions

The project is complex. It involves conceptual metaphor theory, corpus searches, autonomous learning, and the research paper, but it is not expected that all students profit in equal measure from each component. Indeed, some rise to the challenge in one area and sink, at least temporarily, into confusion in another. Few, however, remain unaffected by the experience.

References

Lakoff, G. 1987. *Women, Fire, and Dangerous Things: what categories reveal about the mind.* Chicago: University of Chicago Press.

Lakoff, G., Johnson, M. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Stefanowitsch, A. 2006a. “Corpus-based approaches to metaphor and metonymy.” In *Corpus-Based Approaches to Metaphor and Metonymy*, A. Stefanowitsch, S. Gries (eds.). Berlin: Mounton de Gruyter, 1-16.

Stefanowitsch, A. 2006b. "Words and their metaphors : a corpus-based approach.” In *Corpus-Based Approaches to Metaphor and Metonymy*, A. Stefanowitsch, S. Gries (eds.). Berlin: Mounton de Gruyter, 63-105.

Deignan, A., Potter, L. 2004. “A corpus study of metaphors and metonyms in English and Italian.” *Journal of Pragmatics* 36/7: 1231-1252.

CORPUS, HUMOR AND TRANSLATOR TRAINING

Stella E. O. Tagnin[214]

Abstract

This study intends to examine how trainee translators, especially novice ones, deal with conventional units in the language such as collocations and what the implications are of their translation solutions. Based on the evidence produced by a small corpus consisting of 65 translations of a children's very short story, it will be argued that students need explicit training in the translation of conventional units and that this can be satisfactorily achieved by translation tasks of humorous texts.

Keywords: trainee translator corpus, translator training, conventionality in language, collocations, humor

Introduction

The motivation for this article was a translation task given to 55 Specialization in Translation students at the University of São Paulo, Brazil in 2004. They were asked to translate a very short children's story (Freeman: s/d) – only 340 words long – from American English into Brazilian Portuguese (Appendix 1). Some students submitted more than one translation, which produced a corpus totaling 65 productions. The story revolves around the opposition of two collocations *the best of friends* vs. *the worst of enemies,* each of which appears six times in the text. The first line of the story is already telling: *Nate, Lily and Audrey were the best of friends and the worst of enemies*. The opposition consists of two pairs of antonyms: *best* vs. *worst* and *friends* vs. *enemies*, in a parallel structure *Det Adj Prep N*. Students produced the following solutions for the pairs (the numbers in parentheses refer to the frequency of occurrence in the corpus; no number indicates one occurrence. A literal translation is offered in brackets for the benefit of readers not familiar with the Portuguese language.):

1. **melhores amigos** e **piores inimigos** *[best friends and worst enemies]*
2. os **melhores amigos** do mundo e... os **piores inimigos** *[the best friends in the world and… worst enemies]*
3. os **melhores amigos** e os **piores inimigos** (39) *[the best friends and the worst enemies]*
4. os **melhores amigos** e também os **piores inimigos** *[the best friends and also the worst enemies]*
5. amigos de verdade e inimigos mortais *[true friends and mortal enemies]*
6. amigos inseparáveis e inimigos mortais *[inseparable friends and mortal enemies]*
7. amigos do peito e inimigos de guerra *[bosom friends and war enemies]*
8. grandes amigos como grandes inimigos *[great friends as well as great enemies]*
9. grandes amigos e grandes inimigos (2) *[great friends and great enemies]*
10. super amigos e super inimigos (2) *[super friends and super enemies]*
11. os melhores amigos e os maiores inimigos *[the best friends and the greatest enemies]*
12. os melhores amigos e os mais terríveis inimigos *[the best friends and the most terrible enemies]*
13. os melhores amigos ou ferozes inimigos *[the best friends or ferocious enemies]*
14. os melhores super amigos mas às vezes os piores inimigos *[the best super friends but sometimes the worst enemies]*
15. ótimos amigos e terríveis inimigos *[excellent friends and terrible enemies]*
16. grandes amigos mas às vezes viram inimigos *[great friends but sometimes become enemies]*

[214] Stella E. O. Tagnin is a retired professor of English Translation, Department of Modern Languages at the University of São Paulo, Brazil. She has taught Specialized Translation, Literary Translation and Translation of Humor at the Translation Program for over 25 years. She is the coordinator of the COMET Project – Multilingual Corpus for Teaching and Translation (www.fflch.usp.br/dlm/comet). She is the author of O Jeito que a Gente Diz (2005) and has edited three special issues of leading journals in Brazil focusing on corpora: Crop 10 (2004), TradTerm 10 (2004) and Cadernos de Tradução - Tradução e Corpora (2003). She has published several articles on Corpus Linguistics, Conventionality and Phraseology (in teaching and in translation), Specialized Translation and Terminology (construction of glossaries), Translation of Humor. She is currently working on a project of a bilingual glossary (Portuguese-English) of Brazilian cooking ingredients and dishes.

And two occurrences, by the same translator, disrupted the opposition completely by dividing the sentence into two:

> Às vezes, não existiam no mundo amigos melhores que o Nando, a Lily e a Déia. Às vezes, não existiam inimigos piores.

As can be seen, 42 renderings (lines 1 to 4) were literal translations (*melhores* = best, *piores* = worst; *amigos* = friends, *inimigos* = enemies) that preserved the original opposition. A few (lines 5 to 7) used two well-known collocations which, when combined, failed to maintain the lexical opposition: *amigos de verdade e inimigos mortais* (true friends and mortal enemies), *amigos do peito e inimigos de guerra* (bosom friends and war enemies), *amigos inseparáveis e inimigos mortais* (inseparable friends and mortal enemies). In lines 8 to 10 the translators repeated the adjective in both members of the pair, but only one member, again, could be said to be a well-know collocation: "grandes amigos" and "super amigos". A Google search for Brazilian pages only will highlight the difference:

Translation	**Google hits**
grandes amigos	547.000
grandes inimigos	17.900
super amigos	27.000
super inimigos	90

Most of the others seem not to have recognized the lexical opposition in the first place, producing pairs in which at best only one member was a collocation but with no conventional relation to the other member, like *ótimos amigos e terríveis inimigos* (excellent friends and terrible enemies), *melhores amigos ou ferozes inimigos* (best friends or ferocious enemies), *os melhores amigos e os mais terríveis inimigos* (the best friends and the most terrible enemies).

Other collocations (column 1 below) are also present in the text. Of special interest are *brand new* and *magic wand,* which occur in the following sentence: *On the day that Lily brought her brand new magic wand, both Nate and Audrey wanted to be her best friend.* In Portuguese the modifier may either precede the noun or follow it. In general when the modifier antecedes the noun, it conveys a subjective connotation, so that the most usual structure is *noun + modifier*. Figure 1 shows a summary of how these were translated by the students (column 2); their hits in Google (column 3) and how they were translated by professional translators according to their occurrences in an English-Portuguese parallel corpus Compara (Frankenberg-Garcia & Santos, 2001) (column 4). We have used *NP* to indicate the position of the noun phrase in the translation. So, "*NP* nova" indicates that the modifier follows the noun phrase while "nova *NP*" indicates that it precedes it.

Collocation	**Student translations**	**Google hits for .br**	**COMPARA (v. 10.0.3)[215] hits**
brand new	*NP* novinha em folha (25)	28.600 + 68.300 (novinho)	8 (various variants)
	NP nova em folha (3)	3.500 + 12.000 (novo)	1 (Brazilian original – novo)
	NP novinha (19)	no search performed because translations are not collocations	
	NP nova (7)		
	nova *NP* (7)		
	novíssima *NP* (1)		

[215] http:www.linguateca.pt/COMPARA [27-May-2008]. This corpus consists of literary samples of originals and translations in several Portuguese (Portugal, Brazil, Angola, Mozambique) and English (Britain, United States, South Africa) variants.

magic wand	varinha mágica (50)	21.400	1 (Brazilian translation)
	varinha de condão (14)	6.560	1 (Portuguese translation)
	vara de condão (1)	1.590	no hits

Figure 1: Comparative chart of occurrences for translations of some collocations in *The Best of Friends*

Only 28 out of the 65 translations used a corresponding collocation for *brand new*, namely “nova/novinha em folha”, always following the *NP.* All other options relied solely on the adjective “nova” or the more colloquial “novinha”, not reproducing the collocation. One even used the superlative “novíssima”.

If we look at the occurrences of *brand-new* in Compara we will see that “novinha/novinho em folha” is well represented as its equivalent: for 11 occurrences of *brand new* “novinha/novinho em folha” has been used 8 times, thus confirming that this is the preferred translation. However, as seen above, only 43% of the students used this collocation in the translation.

In opposition, *magic wand* was translated 50 times as “varinha mágica” and 14 as its synonym “varinha de condão”. The difference between both seems to be one of register, “varinha de condão” being slightly more formal. In fact, the Brazilian dictionary Aurélio registers “varinha de condão”, but not “varinha mágica”.

Another collocation which seems to have caused no trouble is seen below:

Collocation	**Student translations**	**Google hits for .br**	**COMPARA (v. 10.0.3) hits**
play house	brincar de casinha (61)	9.380	no hits
	brincar de mamãe, papai e filhinha (2)	no hits	no hits
	brincar dentro de casa (1)	[play inside] inadequate translation	

Figure 2: Comparative chart of occurrences for translations of “play house”

We do not intend to claim that students are always wrong. The above examples have shown that when it comes to very conventional combinations, like *magic wand* or *play house*, for example, and there is an equivalent conventional translation, students have no problem. However, when it comes to less obvious collocations students are not always able to identify them so that we posit that they should receive explicit training in translating them. As one of the problems is identifying these units in the first place, texts in which they are manipulated in order to create humor are excellent material to raise students’ awareness of their presence. First, because when students are told that they are dealing with humorous texts they will be motivated to find the humor in it – to discover how it was constructed – or else they will become frustrated for not understanding the joke! Whereas in a regular text a collocation or phraseology may go unnoticed, such is not the case with humorous texts whose humor is built around such units. In other words, humor relies on a double reading – or double entendre – of the phraseology: a literal meaning and a conventional meaning. If the conventional meaning is not identified by the reader/listener, the intended humor will go flat:

- What would you do if you were in my shoes?
- Polish them.

In this joke humor is based on the double reading of *to be in someone’s shoes*: the conventional meaning is “to be in someone’s situation” and the literal meaning is “to be wearing someone’s shoes”. Unless both readings are activated – and prove to be incongruous – there will be no humor.

Humor certainly does not rely solely on wordplay but most of the time it involves leading the reader down the garden path (Delabastita 1987, Yamaguchi 1988), that is, leading the reader to one interpretation, when(s)he suddenly finds that it is another one altogether: the reader is cheated! But (s)he is amused, and this is why (s)he laughs: “amusement [is] the enjoyment of something which clashes with our mental patterns and expectations” (Morreall, 1989:1). This breach of expectation – or *incongruity* – is one of the main strategies for the creation of humor (Morreall 1989). This can be found at various linguistic levels, from phonological to textual. Here are a few examples:

- *Phonological*: Two peanuts walk into a bar, and one was a salted.
- *Morphological*: Need a therapist? Try Therapist Finder at www.therapistfinder.com
- *Syntactic*: Time flies like an arrow. Fruit flies like a banana.
- *Lexico-semantic*: Two antennas met on a roof, fell in love and got married. The ceremony wasn't much, but the reception was excellent.
- *Textual*: Sorry, we can't e-mail your pizza as attachment.
- *Intertextual/Cultural*: Evidence has been found that William Tell and his family were avid bowlers. However, all the Swiss league records were unfortunately destroyed in a fire, and we will never know for whom the Tells bowled.

Translating humor

If understanding humor is hard enough, translating it is even harder. But it gives students a unique chance to put their creativity to work,

> The search for a translation solution is an exercise of creativity. It develops creative skills that are useful in the learning of a language: either to produce oral discourse or to write texts, the domination of a language has a creative component. (Laurian 1992:125)

And, as claimed above, it raises their awareness of specific aspects of language:

> "The translation of jokes leads students to a reflection on languages by means of contrastive analysis. Even when it is not explicit, such an analysis is necessary to describe where the difficulties lie. That leads to a stronger awareness of the manner in which each language works, that is, the way it approaches objects and ideas." (Laurian 1992:125)

Here is an example of the problems involved in the translation of linguistic humor based on a collocation: in a Frank and Ernest cartoon, the characters are looking at a plate of food but they don't seem to be enjoying it. The owner of the place says: "It is continental breakfast. We just didn't say which continent." Unless the translator knows that *continental breakfast* is a collocation which refers to a specific type of breakfast[216], (s)he will translate it literally as "café da manhã continental" in Portuguese (*continental* being a cognate), which is not "the way we say it" (Tagnin 2005b), because this is the usual type of breakfast in Brazil and is simply called "café da manhã" [*morning coffee*]. Therefore, the humor would be totally lost in a literal translation. In order to reproduce the humor using a similar strategy one would have to look for a collocation that would allow for this type of wordplay. A possible solution could be: "É café colonial, sim. Só não dissemos de que colônia.", in which the collocation "café colonial" refers to a type of mid-afternoon meal consisting of a huge variety of snacks, sweets, cakes, coffee, milk etc. It would roughly translate into English as "It is a colonial meal. We just didn't say which colony it comes from." Notice that though *colonial* is a cognate, the English translation is not funny because *colonial meal* (or even *colonial coffee*) is not a collocation in that language and therefore does create a wordplay.

An experiment

In order to verify how novice translators deal with translating linguistic humor we selected eight short humorous texts for a recent experiment with thirteen translation students at the University of São Paulo. They were asked a) to identify how humor was achieved; b) to explain how they went about discovering it in case they were not able to identify it immediately; and c) to translate the text into Portuguese. Let us look at each text to see how students handled the problem.

1. There's "instant coffee" that is instant, "fast food" that is fast… so what happened to "rush hour traffic"?

[216] A light breakfast consisting usually of coffee or tea and a roll, pastry, or other baked good. (The American Heritage Dictionary of the English Language, 4th edition, CD-ROM)

Students had no problem understanding the joke, but were not able to translate it satisfactorily, mainly because *rush* is also used in the Portuguese collocation “hora do rush” [*rush hour*] but users are unaware of the meaning of *rush* in isolation because they have learned “hora do rush” as a chunk, i.e., as one lexical unit. So, if they were to retain the Portuguese collocation, the meaning of *rush* would have to be explained – which a few students did thus “killing” the joke. Only two students recreated the joke:

- O “algodão doce” é doce, o “bom-bom” é bom, o que acontece com o “trem bala”? [*Sugar candy – literally sweet cotton – is sweet, bonbon is good (“bom” means good), what happens to the bullet train?*]. The ambiguity resides in “trem bala” where “bala” can mean *hard candy* or *bullet.*
- O trem expresso é rápido, o café expresso é rápido, o que aconteceu com as vias expressas? [*The express train is fast, expresso coffee is fast, what happened to expressways?*] The humor lies in the fact that expressways are usually slow due to heavy traffic.

2. If you are going to start cross-country skiing, start with a small country.

Most students were not familiar with this sport because there is no skiing in Brazil, so they produced a literal translation, which did not make much sense in Portuguese. Two students used a different type of sport and two came close to using ambiguity:

- Se você for correr o Rally dos Sertões, cuidado com o Lampião! [If you’re going to enter the Backland off-road rally, watch out for Lampião (a famous outlaw leader who terrorized the Brazilian Northeast early last century)]
- Quer ser bem sucedido na natação? Saiba que ou você nada ou nada. (Do you want to be successful at swimming? Well, then either you swim [= “nada”] or nothing happens [= “nada”])

3. *At a dating service agency:*

- What do you mean I failed the typing test?

- You don’t appear to be anybody’s type.

Although *type* and “tipo” are cognates, *typing* translates as “datilografia” in Portuguese, which makes a literal translation inadequate because it does not recover the ambiguity of *typing*. Some students used other kinds of tests (real or invented) but were unable to recreate the wordplay. Some students even left *typing test* in English! The best solution used the pair “tipografia-tipo” [*typography-type*]:

- Numa agência de casamento [At a wedding agency]:

  - O que você quer dizer com “você falhou no teste de tipografia”? [What do you mean “you failed the typography test?”]

  - Parece que você não faz o tipo de ninguém. [You don’t appear to be anybody’s type.]

4. *At Frank & Ernie´s Modern Studio Dance: We start with the basics: reeling, writhing and rhythmic tics.*

Only one student identified the original trinomial that underlies the wordplay: “reading, writing and ‘rithmetics”, but could not offer a translation. Another student traced the pun back to Lewis Carrol´s *Alice in Wonderland:* “Mock Turtle’s school teachers reeling, writhing and fainting in coils”, itself a pun on the three R’s above. But the translation suggested did not recover the pun. Both resorted to the Web to identify the humor. The other translations used vocabulary related to dancing without any kind of wordplay, which seems to indicate that they did not research at all.

5. *I´d like to paint the town red, but I´m married to old Mr. Turpentine here.*

Seven out of thirteen students recognized the idiomatic expression “paint the town red” and used the Portuguese equivalent “pintar o sete” [= “paint the seven”] as well as a Brazilian product similar to turpentine: “terebentina”, “aguarrás” and “thinner” [pronounced /tiner/ in Brazil]. Three did not get the joke and provided a literal translation while two other ones offered a semantic translation, denoting they understood the idiomatic expression, but without any wordplay:

- Gostaria de aprontar todas, mas o problema é que sou casada com o Sr. Reprimenda [*I'd like to whoop it up but I'm married to Mr. Reprimand.*]
- Queria tanto cair na farra, mas sou casada com o próprio Sr. Cândido! [*I'd love to have a ball but I´m married to old Mr. Innocent here.*]

6. *- Dear Mr. Buck, is there a simple explanation for the booming stock market?*

   - Money see, money do.

Three students did not even give the translation a try, although one identified the pun; six gave a literal translation and two of them recognized they did not get the joke, one left the saying in English and three produced a translation that either made no sense in Portuguese

- O que o dinheiro vê, o dinheiro sente [What money sees, money feels]
- Veja grana, faça grana [See money, make money]

or made sense but was not funny at all:

- O dinheiro imita o movimento da bolsa. [Money imitates the movement of the stock market.]

In short, out of thirteen students, only two understood the pun – they went to the web –, but none was able to suggest an acceptable translation. Granted, it would require a complete recreation in Portuguese.

7. *At the zoo: I can never remember – is it "feed a toad and starve a beaver" or "feed a beaver and starve a toad"?*

This one produced surprising results. Three students came up with a metaphoric reading of the animals: toad = "despicable person", beaver = "hard working person" and thus gave it a literal translation. Although these animals "might" – remotely – call up the same connotations in Portuguese, they are not immediately brought to mind – besides, there is no wordplay in this translation.

Two other students just produced a literal translation and four offered no translation at all. One student suggested

- No zoológico: sempre me confundo entre "alimente o veado e deixe a lebre passar fome" ou "alimente a lebre e deixe o veado passar fome". [*At the zoo: I´m always confused between "feed a deer and starve a hare" or "feed a hare and starve a deer."*]

The strategy involved using animals that rhymed with the translation of *cold* ["resfriado"] and *fever* ["febre"] in Portuguese, respectively "veado" [= *deer*] and "lebre" [= *hare*]. However, as the underlying saying – "Feed a cold and starve a fever" – has no equivalent in Portuguese, this rendering has no humorous effect whatsoever. Another suggestion was

- Na perfumaria: Nunca sei ao certo – é o creme que não compensa? [*I'm never sure – is it the cream that doesn't pay?*]

Another student understood the joke literally, that is, read it as actually referring to feeding animals and wrote: "Either I'm very wrong or this is the most direct of all 8 items. It doesn't look like a wordplay to me, rather it seems like an interpretation of whether or not to feed animals." Only two students were able to identify the saying, but one of them got it twisted "feed a fever and starve a cold". Interestingly, a Google search for *feed starve* returns the correct saying, which seems to imply that most students did not even go to the trouble of searching for it.

8. A headline in Newsweek: Do you, Sir, take this man…?

This seems to have been the easiest to understand and translate as the formula exists in both languages: eleven out of thirteen students got it right. Two seem not to be familiar with it and produced literal translations for "take": "pega" [= *catch*] and "leva" [= *take away*]. One student produced a very creative solution using another part of the wedding ceremony:

- Eu vos declaro marido e marido. [*I declare you husband and husband.*]

It must be said that students enjoyed this task – some even wrote a remark to that effect on their task sheets – and some expressed their enjoyment in class, asking for more as they realized the challenge involved in this kind of exercise. As one of them put it, "Understanding is OK, but translating… that's another story!"

However, as we have seen, "understanding" is not always OK!

Concluding remarks

This paper has shown that the translation of humorous texts can be very effective in raising students' awareness to conventional lexical units, which are pervasive in language but go mostly unnoticed, especially because they do not pose a comprehension problem. However, these units are quite often manipulated in order to create humor and if the translator does not identify the conventional unit on which it is based, not only will (s)he not "get" the humor as (s)he will not be able to translate it adequately, that is, preserving the humor.

To that effect we carried out an experiment with novice translators, in which they were asked to translate eight short humorous texts. The results revealed that a) most of them were not familiar with the conventional units used to create humor and therefore were unable to identify them; b) did no research to try to identify the basis of humor, mainly because, again, they were not able to identify the lexical units on which humor was based, c) were unable to produce a satisfactory translation, mainly because as yet they have had no training in the strategies of translating humor – which is the topic of another discipline in its own right. Nevertheless, the experiment was successful in raising students' awareness of the problem of translating linguistic humor. For this reason, we believe that regular translation exercises with a variety of humorous texts, ranging from jokes and cartoons to rhymes and poems will enhance students' knowledge of the language as well as their creative skills in manipulating language.

And where does corpus come into the picture? First, we used a trainee translator corpus (our 65 translations) to identify students' difficulties. This corpus is currently being expanded with other student productions. Second, we used Compara, an English-Portuguese parallel corpus to check for possible translations. Though this works for collocations and phraseologies – which are recurrent in the language –, it will hardly do so for humor – how can one identify wordplay in a corpus? A better bet would be the Web, at least to identify the basis of the humor. In fact, what kind of corpus would be helpful to translate humor? A parallel corpus of humorous texts and their respective translations? A comparable corpus of original humorous texts in both languages? The latter may be a better choice, not to find adequate translations, but to study each language's strategies to create humor, which might eventually provide the translator with adequate mechanisms to recreate humor in the target language.

References

Delabastita, Dirk. 1987. "Translating pun. Possibilities and Restraints" *New Comparison* 3:143-159

Frankenberg-Garcia, A. & **Santos, D.** 2001. "COMPARA, um corpus paralelo de português e inglês na Web", *Cadernos de Tradução IX*. Universidade Federal de Santa Catarina, Brazil, 61-79.

Freeman, Rachel Santema**.** s/d *The Best of Friends*. Unpublished.

Laurian, Anne-Marie. 1992. "Possible/impossible translation of jokes." *Humor* 5-1/2: 111-127.

Leibold, Anne. 1989. "The translation of humor: who says it can't be done?" *Meta*, XXXIV,1: 109-111.

Morreall, John. 1989. "Enjoying incongruity." *Humor* 2.1: 1-18.

Novo Dicionário Aurélio. Positivo Informática. CD-ROM

Tagnin, S.E.O. 2005a. “O humor como quebra da convencionalidade” *Revista Brasileira de Lingüística Aplicada* v. 5, n. 1: 247-257.

Tagnin, S.E.O. 2005b. *O Jeito que a Gente Diz*. São Paulo: Disal.

Yamaguchi, Haruhiko. 1988. “How to pull strings with words – Deceptive Violations in the Garden-Path Joke.” *Journal of Pragmatics* 12: 323-337.

Appendix

The Best of Friends

Written by Rachel Santema Freeman

SYNOPSIS
We all have friend trouble at one time or another. This story tells of the ups and downs in the friendship of Nate, Lily, and Audrey, and more importantly how much their friendship really means to them.

Nate, Lily, and Audrey were the best of friends and the worst of enemies.

The first thing in the morning they would hug and whisper and teeheehee. They were the best of friends.

But then Nate would decide he was tired of girls' secrets and would holler that he thought girls were silly. Then they were the worst of enemies.

When the three decided to play house and pretend that the carpet was alligator water, they were the best of friends.

But when Audrey announced that she got to be the baby and they had to pretend it was her birthday, Nate and Lily pushed Audrey into the alligator water. Then they were the worst of enemies.

When they decided to color beautiful pictures for each other, they were the best of friends.

But when Nate decided it would be more fun to dump the colors on the floor, he made Lily and Audrey very mad. And when he tried to draw scribbles on their pictures, the girls said they were the worst of enemies.

On the day that Lily brought her brand new magic wand, both Nate and Audrey wanted to be her best friend.

But when Lily refused to share and told them she was going to turn them into toads, they were the worst of enemies.

It sometimes happened that just before going home, Nate, Lily, and Audrey would have a tussle, or a ruckus, or a brawl, and they would part the worst of enemies.

But at night Nate would say, and Lily would say, and Audrey would say, that they missed the others and could not wait until morning so that they could see their best friends.

Nate, Lily, and Audrey ... always the best of friends.

THE PHRASEOLOGICAL ERRORS OF FRENCH-, GERMAN- AND SPANISH-SPEAKING EFL LEARNERS: EVIDENCE FROM AN ERROR-TAGGED LEARNER CORPUS

Jennifer Thewissen[217]

Abstract

The value of learner corpora in the field of learner phraseology has been convincingly illustrated in a number of corpus studies (Granger 1998, Nesselhauf 2003, 2005, Paquot 2007). In addition, the growing attention paid to phraseological errors (Nesselhauf 2003, Osborne 2008, Wang and Shaw 2008) shows that phraseology still very much remains a linguistic "bête noire" even for the more advanced learners. In this study we look at several types of phraseological errors committed by three learner populations, viz. French- German- and Spanish- EFL learners. We do so by using a learner corpus which has been (a) fully error tagged, (b) divided into mother tongue backgrounds, (c) stratified into proficiency levels. This paper reports on two main analyses: (1) we provide an overview of several types of phraseological errors in the three learner populations by basing ourselves on the typology of phrasemes recently developed by Granger and Paquot (forthcoming 2008), (2) we then carry out an analysis of phraseological errors in terms of grammaticality vs acceptability errors (James 1998). The TaLC presentation itself will additionally look at the phraseological errors in the corpus (a) from the point of view of potential L1 influence, i.e. we determine how many phraseological errors in the three populations can be traced back to the learners' L1, and (b) from the point of view of language proficiency, i.e. we investigate whether the number and type of phraseological errors differ according to the proficiency level.

Keywords: learner corpora, error tagging, phraseological errors, mother tongue backgrounds, proficiency levels

Introduction

The current phraseological boom is evidenced by a series of new publications, especially the phraseology volumes by Granger and Meunier (forthcoming 2008) and Meunier and Granger (2008). These volumes are testimony to the upsurge of academic interest in the field of phraseology but also reflect the widening of the scope of phraseology itself which is now seen to encompass multi-word units that would previously not have been considered as phraseological. Learner phraseology in particular has been arousing keen interest among researchers. To this day, many aspects of learner phraseological use have been studied: collocations involving high-frequency verbs have been the focus of studies such as that by Altenberg and Granger (2001) who investigated learners' phraseological use of *make*; phrasal verbs were, among others, studied by Hulstijn and Marchena (1989) and Laufer and Eliasson (1993); recurrent word combinations were the focus of a thorough analysis by De Cock (2003). While these studies nicely show learners' patterns of over- and underuse in phraseology, they do not yet provide us with a general overview of the wide range of phraseological errors committed by EFL learners. The present paper addresses this issue by drawing a larger picture of the types of phraseological errors committed by three EFL populations, viz. French-, German- and Spanish- speaking learners. This paper is written within the larger context of my PhD project, the aim of which is, among others, to analyse learner phraseological errors across mother tongue backgrounds and proficiency levels by using a fully-error tagged learner corpus. This paper thus constitutes a first exploratory investigation of learner phraseological errors in the wider sense. This study is subdivided into two main parts: (1) we give a general breakdown of the phraseological error types in the corpus by basing ourselves on Granger and Paquot's (forthcoming 2008) classification of phrasemes, and (2) we look at the phraseological errors in terms of grammaticality and acceptability (2008) errors. The TaLC presentation will, in addition, interpret the phraseological errors in terms of both potential L1 influence and proficiency levels.

[217] Jennifer Thewissen is an English language assistant at the Université catholique de Louvain, Belgium. She works at the Centre for English Corpus Linguistics where she is currently working on her PhD project. Her doctoral research focuses on the importance of the construct of linguistic accuracy both in SLA and in the field of language testing. Her research is based on an error-tagged sample of the International Corpus of Learner English.

Data and methodology

The learner corpus used here is the *International Corpus of Learner English* (ICLE) which consists of essays by learners from as many as 16 mother tongue backgrounds (Granger et al. 2002). Three learner populations are the object of this study, viz. French-, German-, and Spanish-speaking EFL learners (henceforth FR, GE and SP). As shown in Table 1, a total number of 223 learner essays were used in our analysis. Each learner text was submitted to a rigorous rating procedure: the texts were given to two professional raters who were asked to give each essay a Common European Framework grade (CEF) (Council of Europe 2001) ranging from threshold level B1 to mastery level C2. In cases where the first two raters disagreed by more than one band score, a third rater was called in to rerate the problematic texts. The mean CEF score was calculated for each L1 subcorpus and is presented in Table 1. These results show that while ICLE can generally be said to represent advanced learner writing, it also contains texts that represent lower proficiency levels. Our FR and GE samples were both rated at the advanced C1 level while the SP sample was found to display B1 proficiency overall.

In addition to being independently rated by two, and when necessary three, professional raters, each text was error tagged by a native-speaker linguist, i.e. each text was manually annotated for errors. A total of about 50 000 tokens per subcorpus were error tagged following the guidelines in the Louvain Error Tagging Manual 1.2. (Dagneaux et al. 2005). Following the manual, each error in the corpus is preceded by a descriptive tag which explains the nature of the error and is followed by a possible correction in between dollar signs, as in the following sentence where the error was tagged LP for lexical phrase: *(...) this type of evasion is* ***(LP) at everybody's hand*** *$at everybody's disposal$* (FR). Table 1 describes the number of error-tagged essays and tokens in each subcorpus as well as the mean CEF score for each mother tongue background.

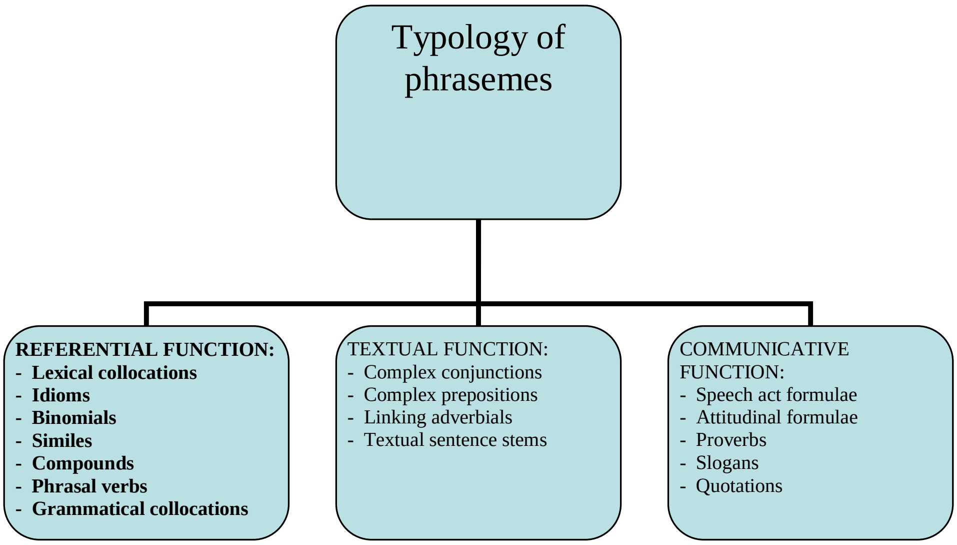
Subcorpora	**Number of essays**	**Overall tokens**	**Mean CEF score**
FR	74	50 558	C1
GE	71	49 945	C1
SP	78	51 860	B1
Total	**223**	**152 363**	

Table 1: Data description

Two of the 56 error tags of the Louvain Error Tagging Manual (Dagneaux et al 2005) will be analysed here: the LP tag which refers to lexical phrase errors and the X*PR tag, which refers to one specific subcategory of lexico-grammatical errors, viz. dependent preposition errors. Lexical phrase errors are lexical errors that affect word combinations, viz. compounds, idioms, phrasal verbs and some types of lexical collocations. The second tag, i.e. X*PR, targets dependent preposition errors. It is subcategorized according to the grammatical category of the word the preposition is attached to: verb for XVPR, noun for XNPR and adjective for XADJPR.

A caveat of this study is that the LP and X*PR categories do not represent all the phraseological errors in the corpus. A number of other error categories also include phraseological errors and will be the subject of future research. Among the other tags that contain phraseological errors, we especially have LS, i.e. lexical single errors, which target errors in (a) isolated single words as in *he* ***(LS) affirms*** *$claims$ he is innocent* where the student confused two existing words, and in (b) lexical collocations where a single lexical word is erroneous as in ***(LS) high*** *$heavy$ responsibilities.* LS will thus need a considerable amount of weeding out to isolate the errors which are collocational in nature from those which affect words in isolation. In the meantime, because the LP and X*PR tags only represent part of the picture, the results presented here should be seen as exploratory.

All the LP and X*PR concordances in the FR, GE and SP subcorpora were extracted with WordSmith Tools 4 (Scott 2004) and were classified phraseologically following the typology of phrasemes recently developed by Granger and Paquot (forthcoming 2008). As the authors point out, this typology purposely adopts "a much wider perspective and includes many word combinations that would traditionally be considered to fall outside the scope of phraseology". Granger and Paquot's typology is presented in Graph 1 below, with the types of phrasemes found in the LP and X*PR categories highlighted in bold:



Graph 1: Typology of phrasemes (Granger & Paquot 2008)

It was nevertheless necessary to adapt the classification proposed in Graph 1 in order to classify the LP and X*PR errors into the different referential phraseme categories. The adaptation is explained below along with the resulting breakdown of the phraseological errors in the three subcorpora.

Breakdown of phraseological errors

I analysed grammatical collocations, idiom-like phrases and phrasal verbs separately in so far as they constitute clearly identifiable entities:

1. **Grammatical collocation errors** correspond to X*PR errors and concern errors on dependent prepositions, i.e. cases where the dependent preposition in N/V/ADJ + preposition combinations is erroneous, e.g. *marriage may not* ***(XVPR) appeal*** *$appeal to$ people* (GE); *In my view there is no* ***(XNPR) justification in*** *$justification for$ capital punishment* (GE); she is ***(XADJPR) hard to*** *$hard on$ her son* (GE).

2. **Errors in idiom-like phrases** concern lexically opaque, i.e. non-compositional phrases, where the overall meaning cannot be deduced from the sum of the parts, e.g. *This book gives you* ***(LP) food for the mind*** *$food for thought$* (FR); ***(LP) to turn over a new leave*** *$to turn over a new leaf$ (GE).*

3. **Phrasal verb errors** exclusively include errors on verb + adverbial particle combinations, e.g. *when I* ***(LP) stand up $get up$*** *at 9.30* (GE); *I hope to be able to* ***(LP) get my point through*** *$get my point across$* (SP); *people marry and* ***(LP) set up*** *$start$ a family* (GE)[218].

However, I grouped compound errors, binomials, similes and lexical collocation errors in the same category referred to broadly as the lexical collocation category. While the task of distinguishing between these types of word combination in the English of native speakers already constitutes a challenge to say the least (Cowie 1998, Howarth 1998, Nesselhauf 2005, Paquot 2007), it becomes even more arduous when dealing with learner errors. For instance, are the following examples instances of lexical collocation, compound or free combination errors?

[218] Errors on prepositional verbs, i.e. verb + dependent preposition combinations, are classified in the grammatical collocation category.

- *The people who died in the war were **(LP) civil people** $civilians$* (SP)
- *I decided to make a last attempt **(LP) to get my stomache filled** $to satisfy my hunger$* (GE)

In their study of collocational errors by advanced EFL learners, Wang and Shaw (2008: 209) also emphasise the problem of distinguishing between errors that affect lexical collocations and free combinations: "when the collocations were produced or misused by the learners, it is very difficult to say which category, namely free ones or restricted ones, they belong to". The following examples illustrate the errors in the lexical collocation category: ***(LP) to run out of hand*** *$to get out of hand$* (FR); ***(LP) tendencies of consumption*** *$consumer habits$* (SP); *I am **(LP) the only child of my parents** $an only child$* (GE); *they are the **(LP) supporters of their families** $breadwinners$* (SP); *things like satellites or computers didn't even **(LP) come to their minds** $exist$ at that time* (GE); ***(LP) daughters and sons*** *$sons and daughters$* (GE).

The breakdown of errors in the lexical collocation, grammatical collocation, idiom-like phrases, and phrasal verb categories is presented below for each mother tongue background.

Type of phraseological error	**FR**	**GE**	**SP**	**Total**
Lexical collocation category	97 (44%)	108 (41,5%)	**213 (44,5%)*****	418 (43,5%)
Free combinations				
Lexical collocations				
Compounds				
Grammatical collocations	69 (31%)	82 (31,5%)	**208 (43,5%)*****	359 (37,5)
Idiom-like phrases	20 (9%)	21 (8%)	18 (4%)	59 (6%)
Phrasal verbs	35 (16%)	48 (18,5%)	37 (8%)	120 (12,5%)
Total	**221 (100%)**	**259 (100%)**	**476 (100%)*****	**956 (100%)**

Table 2: Breakdown of LP and X*PR phraseological errors

No significant difference was highlighted in the total number of phraseological errors between the FR and GE groups[219]. In fact, Table 2 shows that the phraseological profiles for the FR and GE groups are very similar, with no significant difference in the number of errors across the four phraseological error subcategories. However, a highly significant difference was found between the total number of errors in the SP and the FR and the SP and GE data ($p \leq 0.0001$ each time). The difference between the SP group and its FR and GE counterparts is mainly due to the significantly higher number of errors in the lexical and grammatical collocation categories in the SP data (with $p \leq 0.0001$ for FR and SP and GE and SP). Concerning grammatical collocations, the majority of X*PR errors affect verb + dependent preposition combinations. This applied to the three subcorpora: for the FR group 69,5% of all X*PR errors affect verbs; for the GE and SP groups the proportions reach 62% and 72%, respectively. Examples of XVPR errors include:

- *We use the word "religion" and say "television is the opium of the masses" in order to **(XVPR) refer** $refer to$ society at the end of the 20th century* (SP)
- *They are **(XVPR) dying for** $dying from$ starvation or lack of medicines* (SP)

The higher number of lexical and grammatical collocation errors in the SP subcorpus is to be related to the lower level of proficiency displayed by the SP sample.

Grammaticality vs acceptability errors

It is generally agreed that distinguishing between correct and erroneous instances is more straightforward for certain linguistic categories than for others. This is the case for article errors, for instance, (Tomiyana 1980, Ekiert 2004, Díez Bedmar and Papp 2005) as well as certain syntactic and morphological errors (Bardovi-Harlig & Bofman 1989) where errors usually "occur in a patterned, rule-governed way" (Ferris 1999: 6) and are therefore more easily detectable. Lexis, however, is a domain where the distinction between right and wrong becomes much more blurred. Ferris (1999: 6) calls lexical errors "untreatable" errors, i.e. errors for which "there is no handbook

[219] This study uses the chi-square test to detect statistically significant differences.

or set of rules students can consult to avoid or fix those types of problems". A multi-word unit may indeed be erroneous in one of two ways: it may be (1) formally wrong or (2) formally correct but inappropriately used in context. To make this distinction James (1998) differentiates between grammaticality errors which correspond to formally inexistent multi-word units, and acceptability errors which are formally existing sequences but which are inappropriately used in context. Formally inexistent multi-word units include cases such as *they should **(LP) keep close watch** $keep a close watch$ on them* (FR), which are near-hits, viz. the combination is a very close approximation of an existing multi-word unit, and cases such as *I would like to **(LP) make a vindication of** $stress$ the importance of literature* (SP), which bear no resemblance to any existing multi-word unit. On the other hand, acceptability errors include cases such as (in reference to mental hospitals) *such a method could not work with those who are **(LP) out of their mind** $mentally ill$* (FR), which contain existing but contextually inappropriate combinations.

The British National Corpus (http://corpus.byu.edu/bnc/) as well as a number of native and learner dictionaries and my own native speaker intuition were used to establish the existing vs inexistent nature of the errors in the three subcorpora. The results are presented below:

	FR	**GE**	**SP**	**Total**
Acceptability errors	125 (56,5%)	131 (50,5%)	190 (40%)	446 (46,5%)
Grammaticalityerrors	96 (43,5%)	128 (49,5%)	286 (60%)	510 (53,5%)
Total	221 (100%)	259 (100%)	476 (100%)	956 (100%)

Table 3: Proportion of grammaticality vs acceptability word combinations

No significant difference was found in the number of grammaticality and acceptability errors between the FR and GE subcorpora. The SP subcorpus was found to include significantly more acceptability ($p\leq 0.01$) and grammaticality ($p\leq 0.02$) errors than the FR subcorpus, but no significant difference was highlighted between the GE and the SP data ($p\leq 0.08$ for acceptability and $p\leq 0.1$ for grammaticality errors).

The error proportions within each sample show that the FR subcorpus includes slightly more acceptability than grammaticality errors while the GE group almost displays a 50-50 distribution between the number of grammaticality and acceptability errors; the SP group, on the other hand, clearly committed a much higher proportion of grammaticality than acceptability errors (60% vs 40%). This may again perhaps be related to the fact that the SP population displays a lower level of language proficiency than its FR and GE counterparts. Thus, whereas the higher level of proficiency in GE and SP samples allows these learners to use a high number of existing multi-word units but inappropriately in context, the SP group, because of more limited language command, tends to use more inexistent combinations in the first place.

Potentially transfer-related errors

The TaLC presentation will give the results of the number of potentially transfer-related errors in the three language groups. The process used here to detect potential transfer errors is back translation, a method advocated, among others, by Granger (2008) as one way of assessing the potential influence of the learners' L1 phrasicon on L2 performance[220]. Word combination errors were translated back into the learners' L1 and in cases where an L1 equivalent to the error could be found, the word combination was categorised as a potential transfer error. An example of this is *everyone has to work hard at school, at university, and later on in **(LP) the active life** $at work$ in order to find a job* (FR) where "the active life" is a direct calque of the French "la vie active", meaning "at work". As will be shown during the TaLC presentation, the results for the more advanced samples, i.e. FR and GE, seem to confirm Kellerman's (1984: 121) claims that the "hoary old chestnut" according to which transfer does not affect the more advanced learner "should finally be squashed underfoot as an unwarranted generalization based on very limited evidence" (for more on the link between transfer and language proficiency see also Kellerman 1977, 1978, 1979, Wode 1976, Zobl 1980).

Our in-depth analysis of phraseological errors will also show that the causes of errors at the more advanced levels are by no means clear-cut but rather, as Granger (2004: 135) puts it, "advanced interlanguage is the result of a

[220] We refer to « potentially » transfer-related errors in so far as, although the error bears trace of possible L1influence, it is not certain that the learner did indeed resort to the L1 when the error was committed.

very complex interplay of factors: developmental, teaching-induced, and transfer-related”. This complex interplay will be illustrated across the three languages.

Conclusion

Although our analysis only takes two error tags into account and therefore yields a relatively patchy picture of the phraseological errors found in the learner corpus investigated here, it nevertheless goes a long way to showing the value of adopting a computer-aided-error-analysis approach to the study of learner phraseological errors. Our analysis of the LP and X*PR errors in the French, German and Spanish components of ICLE has revealed a number of interesting findings: first, the Spanish subcorpus contains significantly more phraseological errors than its French and German counterparts. This is mainly due to the significantly higher number of errors in the lexical and grammatical collocation categories. Second, in terms of grammaticality and acceptability errors, the Spanish subcorpus also stands out as including more grammaticality than acceptability errors. As suggested, both these findings should be related to the Spanish subcorpus displaying a lower level of language proficiency, which leads these learners to make (a) more overall phraseological errors and (b) of a different kind (more grammaticality errors).

References

Altenberg B. and **Granger S.** 2001. The grammatical and lexical patterning of make in native and non-native student writing. *Applied Linguistics* 22 (2), 173-194.

Bardovi-Harlig K. and **Bofman T.** 1989. Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11: 17-34.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Cowie A.P. (ed.) 1998. *Phraseolgy: Theory, Analysis and Applications*. Oxford: Oxford University Press.

Dagneaux E., Denness S., Granger S., Meunier F., Neff J. and **Thewissen J.** 2005. *Error Tagging Manual Version 1.2*. Centre for English Corpus Linguistics. Université Catholique de Louvain, Louvain-la-Neuve. Unpublished manual.

De Cock S. 2003. *Recurrent sequences of Words in Native Speaker and Advanced Learner Spoken and Written English*. Unpublished doctoral dissertation. Université catholique de Louvain. Centre for English Corpus Linguistics.

Díez Bedmar M.B. and **Papp S.** 2005. The usage of central articles by Spanish and Chinese learners of English at University level. Paper presented at the workshop held in conjunction with the 4th International Contrastive Linguistics Conference, September 20-23 2005, Santiago de Compostela, Spain.

Ekiert M. 2004. Acquisition of the English Article System by Speakers of Polish in EFL and ESL settings. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, 4 (1), 1-23.

Ellis R. 2003. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Ferris D. 1999. The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing* 8, 1-10.

Granger S. 1998. Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In Cowie, A. (ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 145-160.

Granger S. 2004. Computer learner corpus research: current status and future prospects.

In Connor U. and Upton T? (eds.) *Applied Corpus Linguistics: A Multidimensional*

Perspective. Amsterdam & Atlanta: Rodopi. 123-145.

Granger S. 2008. Some major challenges for theoretical and applied phraseological research. Paper to be presented at the 3rd International Postgraduate Conference on Formulaic Language (FLaRN), Nottingham, 19-20 June 2008.

Granger S., Dagneaux E. and **Meunier F.** 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain. Available from http://www.i6doc.com.

Granger S. and **Meunier F.** (eds). 2008. *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: Benjamins.

Granger S. and **Paquot M.** 2008. Disentangling the phraseological web. In Granger S. and F. Meunier (eds) *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia : Benjamins.

Howarth P. 1998. The Phraseology of Learners' Academic Writing. In Cowie A.P. (ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: OUP, 161-186.

Hulstijn J. and **Marchena E.** 1989. Avoidance: grammatical or semantic causes. *Studies*

in Second Language Acquisition 11: 242-255.

James C. 1998. *Errors in Language Learning and Use*. London and New York: Longman.

Kellerman E. 1977. Towards a characterization of the strategy of transfer in second language learning. *Interlanguage Studies Bulletin* 2 (1): 58-145.

Kellerman E. 1978. Giving learners a break: native language intuitions as a source of

predictions about transferability. *Working Papers on Bilingualism* 15, 59-92.

Kellerman E. 1979. Transfer and non-transfer: where are we now? *Studies in Second*

Language Acquisition, 37-57.

Kellerman E. 1984. The empirical evidence for the influence of the L1 in interlanguage. In Davies A., Criper C. and Howatt A. (eds.) *Interlanguage*. Edinburgh: Edinburgh

University Press, 98-122.

Laufer B. and **Eliasson S.** 1993. What causes avoidance in L2 learning: L1/L2 difference, L1/L2 similarity, or L2 complexity? *Studies in Second Language Acquisition* 15 (1), 35-48.

Meunier F. and **Granger S**. (eds). 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam and Philadelphia: Benjamins.

Nesselhauf N. 2003. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics* 24 (2), 223-242.

Nesselhauf N. 2005. *Collocations in a Learner Corpus*. Amsterdam: Benjamins.

Osborne J. 2008. Phraseology effects as a trigger for errors in L2 English: The case of more advanced learners. In Meunier F. and Granger S. (eds) *Phraseology in Foreign Language Learning and Teaching*, Amsterdam/ Philadelphia: Benjamins, 67–83.

Paquot M. 2007. *EAP vocabulary in EFL learner writing: from extraction to analysis: A phraseology-oriented approach*. Unpublished doctoral dissertation. Université catholique

de Louvain, Centre for English Corpus Linguistics.

Scott M. 2004. *WordSmith Tools 4*. Oxford: Oxford University Press.

Tomiyana M. 1980. Grammatical errors and communication breakdown. *TESOL Quarterly* 14, 71-79.

Wang Y. and **Shaw P.** 2008. Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME Journal* 32, 201-228.

Wode H. 1976. Developmental sequences in naturalistic L2 acquisition. *Working Papers on Bilingualism* 11, 1-13.

Zobl H. 1980. The formal and developmental selectivity of L1 influence on L2 acquisition. *Language Learning* 30, 43-57.

IN THIS PRESENT PAPER… SOME EMERGING NORMS IN LINGUA FRANCA ENGLISH WRITING IN THE SCIENCES?

Christopher Tribble[221]

Abstract

In the work of a number of influential authors (e.g. Ammon 2000, Jenkins 2000), there is an assumption of a native speaker hegemony over academic communication in the English language. Such a position is in line with accounts which have expressed concern at the growing dominance of the English language in international scientific communication since the 1980s (e.g. Pennycook 1994, 2001), critical perspectives implicit in the linguiscism of Skutnabb-Kangas (1988), and arguments around linguistic imperialism (e.g. Phillipson 1992) which have focused on the controlling and privileged role of native speaker gatekeepers.

In this paper, I suggest that this account is neither empirically grounded, nor currently true, and argue that in professional and academic writing, both authorship and gate keeping authority have shifted, and that the production and evaluation of texts it is no longer a native speaker monopoly. What was once a foreign language for many writers in academic and professional settings, is now accepted by many practitioners as a convenient lingua franca, a resource no longer within the control of a single mother tongue speech community. I go on to argue that teachers also need to recognise this reality and that they will be better served by using the notion of expertise (c.f. as the starting point for their identification of relevant exemplars, rather than the notion of the native-speaker.

In the paper I present a diachronic study of a 1 million word corpus of articles taken Acta Tropica[222] *, and consider how the authors, reviewers and editors function as a written Lingua Franca discourse community. In the course of this study I use the academic components of the BNC as a reference and show how studies of keywords, lexical bundles, and discourse structure indicate what may be emerging English Lingua Franca (ELF) norms. I shall also report on an ongoing but problematic attempt to re-situate the corpus based findings in communities of practice. At the end of the paper I call for the selection of exemplars for writing instruction on the basis of the expertise of their authors (Rampton 1990, rather than on their mother tongue status.*

Keywords: writing, lingua franca, academic, science, genre

Linguistic politics and L2 writing

Drawing on Kachru 1985, Ammon has argued:

> "in spite of the majority of non-native speakers or the non-inner-circle countries, many of whom use the language actively and regularly in institutional frameworks, the native speakers of the inner-circle countries retain the hold to the yardstick of linguistic correctness." (Ammon 2000: 112)

Such a position is in line with accounts which have expressed concern at the growing dominance of the English language in international scientific communication since the 1980s (e.g. Pennycook 1994, 2001), critical perspectives implicit in the *linguiscism* of Skutnabb-Kangas (1988), and arguments around *linguistic imperialism* (e.g. Philllipson 1992) which have focused on the controlling and privileged role of native speaker gatekeepers.

While I accept that at the time these authors did their most influential work the world was largely constructed along the lines that they outlined, I would suggest that this situation no longer holds true in all cases – at least not in such direct and simple ways. For English language users and teachers in the twenty-first century, things have changed. Arguments against the ever-increasing dominance of a small number of global languages still matter – especially when the spread of large *spoken* languages like English, Spanish or Mandarin Chinese comes to be

[221] Dr Christopher Tribble is a lecturer at King's College, London University, where he teaches courses in English for Academic Purposes and Managing and Evaluating Innovation on the MA in ELT and Applied Linguisitcs programme, and Text and Corpus Analysis on the BA in English Language and Communication. He has published and presented widely on the teaching of writing and on corpus applications in language education (most recently with Mike Scott 2006, Textual Patterns: key words and corpus analysis in language education in the Benjamin's Studies in Corpus Linguistics Series) and is a member of the TALC organising committee. Apart from this academic work, Chris Tribble is a consultant and trainer in project management and project and programme evaluation, and a documentary photographer specialising in work with development organisations and in theatre and performance. email:christopher.tribble@kcl.ac.uk. home page: www.ctribble.co.uk

[222](http://www.elsevier.com/wps/find/journaldescription.cws_home/506043/description?navopenmenu=1 accessed February 7, 2008

associated with the death of smaller languages (e.g. Skutnabb-Kangas and Phillipson 1994). I would contend, however, that in professional and academic *writing*, both authorship and gate keeping authority have shifted and the production and evaluation of these texts it is no longer a native speaker monopoly. What was once a foreign language for many writers in academic and professional settings, is now accepted by many practitioners as a convenient *lingua franca*, a resource no longer within the control of a single mother tongue speech community. Tardy (2004) discusses the impact of this change in attitudes amongst pluri-lingual research students. My purpose in this paper, is to argue that teachers also need to recognise this reality and that they will be better served by using the notion of expertise as the starting point for their identification of relevant exemplars, rather than the notion of the native-speaker.

In developing this argument I will focus on the example of a well respected research journal which is written and edited by scholars who are for the greater part *not* first language users of English but who do have expertise as *lingua franca* writers in specific domains. From my perspective, the fact that it is now relatively easy to have access to this kind of data is of revolutionary importance because the identification of appropriate text exemplars for use in English for Academic Purposes writing programmes is one of my central professional concerns – especially where a genre informed approach is being implemented. Without access to a *range* of relevant exemplars it is not possible to put in place such a teaching programme. Moreover, if teachers feel obliged to restrict themselves to the production of "native speakers" they can be accused of perpetuating asymmetries of power, or more seriously, fail to provide exemplars that are genuinely relevant to the their students' *lingua franca* writing needs. However, if teachers can choose relevant exemplars on the basis of the writers' expertise rather than on the basis of the accidental criterion of mother tongue status, Ammon's concern about fairness and unfairness becomes irrelevant, and students get the educational programmes that they need. Win, win.

If I am correct, *native speaker* ceases to be a useful criterion for the selection of exemplar texts. The critical thing is the extent to which a text is likely to be acceptable in the eyes of peers in the discourse community in which an expert writer already acts, or which they wish to enter. If the text is published in a respected peer-reviewed journal, it's an expert text. The L1 status of the writer has become irrelevant.

The value of expertise as a criterion for text selection in writing instruction

As an example of how things have changed, in the first part of the paper, I summarise work undertaken with a small set of 8 recent articles published an international research journal – *Acta Tropica*. This journal is published by Elsevier – one of the leading publishers of academic journals (and, what is more, a joint Anglo-Dutch company), is edited by Swedish scholars, and has an editorial board of 23, only 9 of whom are based in countries where English is a first language (see appendices for a list). Eight full text articles were accessed via www.sciencenow.com[223] from *Acta Tropica* 92 (2004). Of the thirty-six authors involved in the production of the eight articles, twenty nine are from countries where English is neither the first language nor an official language. Four come from the "inner-circle" countries of the USA or Australia, and three come from India – where English is an official language. This author information is summarised in the table below.

5. Authors	6. Country	7. Authors	8. Country	9. Authors	10. Country
11. 6	12. Brazil.	13. 2	14. USA	15. 3	16. India
17. 4	18. Argentina	19. 2	20. Australia		
21. 3	22. Kenya				
23. 3	24. Central African Republic				
25. 3	26. Cameroon				
27. 3	28. Venezuela				
29. 1	30. Argentina				
31. 1	32. France				
33. 2	34. China				
35. 2	36. Germany				

[223] a resource available to staff and students in higher education institutions with subscriptions to electronic journals

37. 1	38. Switzerland

I do not have information on the membership of the journal's peer review panel, and only have country information and family names for the editorial board and the article authors. However, given the information to hand, it seems reasonable to assume that in the case of the *Acta Tropica*, expertise is not the monopoly of those who have English as a first language. Again, although we have no information on whether the authors of these articles had any linguistic support when preparing for publication, the collection of texts, nevertheless, appears to offer a glimpse of a community of researchers, expert authorities and editors which is sufficiently coherent to represent a discourse community (Swales 1990) for whom English is a *Lingua franca.*

Expert exemplars vs. native speaker exemplars

This preliminary study has two main purposes. The first is to problematise the notion of "native speaker" as a criterion for the selection of pedagogic examples in writing instruction, with a second being to flag the need for language teachers to have a critical awareness of how best to select from potential examplar texts when developing pedagogic accounts of writing practices in a specific discourse community.

In subsequent sections of the paper, I comment on features of a larger collection of *Acta Tropica* papers which might indicate the existence of emerging Lingua Franca norms, and in the final section comment on some of the challenges I have been facing in building a more ethnographic account of the writing practices of contributors to *Acta Tropica*.

References

Ammon, U., (2000) "Towards more fairness in international English: linguistic rights of non-native speakers?" in Phillipson, R. (ed.) *Rights in language*. London: Lawrence and Erlbaum : 111-116

Jenkins J (2000) *The phonology of English as an international language* Oxford: Oxford University Press

Kachru, B.J. (1985) "Standards, codification, and sociolinguistic realism: The English language in the outer circle" In R. Quirk, & H. Widdowson (Eds.) *English in the world: Teaching and learning the language and literatures,*Cambridge: Cambridge University Press pp.11–30

Pennycook, A. (1994) *The cultural politics of English as an international language,*New York Longman

Pennycook, A. (2001) *Critical applied linguistics: A critical introduction,*Mahwah, NJ: Lawrence Earlbaum Associates

Phillipson, R.,(1992). *Linguistic imperialism* Oxford: Oxford University Press.

Seidlhofer, B., (2001) "Closing a conceptual gap: the case for a description and pedagogy of English as a lingua franca" , *International Journal of Applied Linguistics* 11/2:133-157

Skutnabb-Kangas, T., (1988) "Multilingualism and the education of minority children" In T.Skutnabb-Kangas, & J.Cummins (Eds.) *Minority education: From shame to struggle,*Clevedon: Multilingual Matters pps: 9–44

Skutnabb-Kangas, Tove, and Robert Phillipson, (Eds) (1994). *Linguistic human rights: overcoming linguistic discrimination.* Berlin: Mouton de Gruyter.

Swales, J.M., (1990) *Genre Analysis,*Cambridge: Cambridge University Press

Tardy, C. (2004) "The role of English in scientific communication:*lingua franca* or Tyrannosaurus rex?" , Journal of English for Academic Purposes 3 (2004):247–269

POLITENESS IN ACADEMIC SETTINGS: THE CASE OF MICASE

Josta van Rij-Heyligers[224]

Abstract

The Michigan Corpus of Academic Spoken English (MICASE) project is a specialised corpus containing speech events of academic and research interactions. In my work as a language, learning and research advisor, MICASE represents a body of valuable texts for examining the markers of politeness in interactions between advisers and students. However, as I argued in a previous TALC presentation, a corpus linguistics method to analysing data (product orientation) is incomplete without a discourse approach to meaning creation (process orientation). Accordingly, the discourse of politeness can be analysed from a genre and/or critical perspective, which represent a top-down, deductive framework to test certain assumptions on politeness and positions of power, but these assumptions can be supported by a bottom-up, inductive corpus linguistics approach. Such an approach tends to start from a whole corpus reading involving sample texts of a specific corpus, followed by a corpus-based analysis of all the texts that either verifies or disproves the findings observed from the small sample. Alternatively, the process of analysis can be reversed or combined. The present paper introduces theories of politeness and power, advances an eclectic approach to examine politeness in specific contexts, briefly reviews corpus findings of the literature on MICASE, presents two discursive events from MICASE for closer analysis, and relates these observations to the theories presented. The implications of the use of a specialised corpus and eclectic approach to the study of politeness are then discussed.

Keywords: Politeness theories, critical/genre theory, corpus linguistics, MICASE, eclectic approach

Introduction

In advisory sessions, acts of politeness can facilitate as well as hinder communication. But what do these acts consist of and what makes them (in)effective? Traditional theories of politeness are founded on the premise that speech acts represent rational behaviour (Terkourafi 2005), and that utterances can be analysed on the basis of stable meaning (semantics) and form (syntax). However, in interactions between academic advisor and student(s), factors like position of power, degree of co-membership or social and cultural similarity, and nature of the task are likely to influence the communicative events taking place. These factors may well unhinge the stable meaning and form of politeness behaviour. Rational frameworks may thus not suffice and an eclectic approach to the study of politeness may well be needed. As Candlin and Hyland observed (1999: 2) texts are multidimensional constructs requiring multiple perspectives for their understanding. This paper advances such an eclectic approach to the study of politeness and corpus analysis. It briefly reviews the politeness theories and highlights the critical and genre perspectives of politeness. It also advances the use of corpus-based approaches and specialised corpora for analysing politeness behaviour in context. In particular, the Michigan Corpus of Academic Spoken English (MICASE) corpus (Simpson, Briggs, Ovens, and Swales 2002**)** provides a helpful resource to deepen our understanding of politeness behaviour in academic settings. One setting relevant to my work is highlighted: advisory sessions.

Politeness theories: from rational to critical

Although no unified theory to the study of politeness exists, traditional approaches generally consider interlocutors as rational, face-saving agents and politeness as a linguistic device or strategy to which universal rules can be applied. For example, the earlier studies of politeness interpret politeness according to illocutionary (speaker's intention) factors; and conversational maxims (see Paltridge 2000), such as Lakoff's three rules of politeness: "not to impose, to give options, and to be friendly" (Terkourafi 2005: 239). The best known, and most influential of these theories, however, is that of Brown and Levinson (1987), who presented a hierarchical taxonomy of politeness concerned with face threatening acts (FTAs) that range from 'on record' (direct) strategies, consisting of bald on record (without redress), and positive and negative politeness (with redress); to

[224] Josta van Rij-Heyligers (EdD) is an academic learning advisor and coordinator of EAP courses at the Student Learning Centre of the Centre for Academic Development, University of Auckland, New Zealand. Her interest in corpus linguistics is related to he r work of teaching academic English in a multicultural setting and her formal studies in Applied Linguistics. She has published two papers she presented at previous TaLC conferences (Granada, 2004, published in 2007, and Paris, 2006, in press).

‘off-record’ (indirect, e.g. hints) strategies. In this framework, negative politeness deals with the individual’s negative face such as claim to territory and self-preservation, whilst positive politeness sees to a person’s positive face: claim for self, self-image, to be liked or approved of. FTAs concerned with negative face consist of warnings and threats, directives, request and suggestions; FTAs relating to positive face consist of, for example, disapprovals, accusations, criticisms, complaints, contradictions and challenges (Harris 2003).

The above theories have in common the view that politeness constitutes socially appropriate behaviour and face-risk minimisation. Later theories, however, emphasise that politeness norms across and within cultures can be challenged. Politeness is not just a linguistic tool or strategy; rather it is negotiated at the micro-level by the participants involved. The focus on politeness that comes into being through situated exchanges means a priori predictions based on universal rules cannot be made. Researchers in Japan and China have shown, for example, that Brown and Levinson’s politeness frame does not always hold in their respective settings (see Paltridge 2000). Also, acts that violate politeness rules do not necessarily constitute inappropriate behaviour (Harris 2003). They could be seen as unintended (due to such factors as cultural difference and level of discourse competency) and intended, when a deliberate challenge is made to dominant power or norms. Ide’s notions of discernment politeness and volitional politeness are of relevance here; the former pertains to speakers following social norms and conventions and the latter relates to the speakers’ intentional use of politeness strategies (Haugh 2003). These alternative theories of politeness indicate that polite behaviour can be normative (what is socially appropriate/constraint) as well as strategically chosen (Watts 2003). Clearly, what constitute politeness (and impoliteness) varies across (and within) cultures and is often related to issues of power (an element not entirely negated in Brown and Levinson’s work).

Politeness in institutional settings

Polite behaviour for the production and maintenance of society’s power structure may feature more highly in institutional and academic settings than in every day conversations. As Scollon and Scollon (1995) noted, a hierarchical system of politeness typifies institutional discourse as asymmetrical distribution of power exists. Those in power generally employ, using Brown and Levison’s terms, positive politeness or involvement strategies, whereas those with less power tend to use negative politeness or independence strategies not only to ‘save’ negative face, but also to establish or maintain social distance or a level of formality (Harris 2003). This distancing reduces the risk of conflict. In fact, negative politeness strategies can serve both institutional (instrumental) and interpersonal (social-integrative) goals, whilst also reifying existing power relations (Harris 2003; Youmans 2001).

In addition, the degree of co-membership influences politeness behaviour. Erickson and Shultz (cited in Bardovi-Harlig and Hartford 1993) defined co-membership as shared attributes of social identity. However, in institutional settings, there is besides *social* co-membership also *role* co-membership, established by expertise and institutional status and familiarity. This aspect of co-membership complicates communication as the need to ascertain interlocutors’ status through role-preserving strategies is enhanced. Politeness theory assumes that the use of politeness markers (linguistic elaboration) increases with increased social distance. Yet, this notion may not be applicable to interactions taking place in institutions. For example, Bardovi-Harlig and Hartford (1993) suggested that in academic advice sessions a non-linear relationship exists. That is, undergraduate students’ low co-membership with advisors leads to less use of linguistic devices to mitigate a threatened status balance (or FTA) as this balance is generally stable and rights and obligations well-defined. In contrast, high co-membership involves more ambiguity in rights and obligations on the part of graduate students since acting within their status boundaries is difficult to assess. Accordingly, more status-preserving strategies (negative politeness) through the use of mitigators may be used by these students in advisory sessions, but once completed they, unlike undergraduates, may initiate small talk, which strengthens co-membership (positive politeness).

Analysing politeness in academic settings: an eclectic approach

The above theories points to the need for an eclectic approach to the study of politeness. Critical and genre theories extend frameworks proposed by theorists such as Brown and Levinson: a critical perspective relates language use in different contexts to issues of power (see Paltridge 2000); a genre approach examines utterances in their specific context or discourse community that makes up the basic rules for ‘appropriate’ speech, and focuses on exploring structural and linguistic regularities of a particular ‘type’ of speech or text. Like the traditional theories on politeness, these approaches represent deductive frameworks to test certain assumptions but, as stated, incorporate contextual and power factors into their analysis. An eclectic approach further makes use of corpus-based methods as regularities can be observed and assumptions arising from theory can be substantiated by the bottom-up, inductive method of corpus linguistics. As the tools developed by corpus linguists enable the examination of a large corpus of text according to word frequency and co-occurrence of words and phrases, they can, also because of sheer text volume, provide “new insights into language usage” and “avenues for further investigation into aspects of culture and society” (Barlow 2004). A corpus which provides access to individual texts further lends itself to deeper analysis of the discourse.

But not all corpora are equally useful. Lee (2001) observed that aspects of genre can be better examined in small specialised corpora as a high level of homogeneity across texts can be assumed. Mega corpora like the British National Corpus (BNC) are of little use to investigate specific discourse, genres or settings because even though they contain many texts and contexts they often have few samples of each (Gómez 2004). Fortunately, researchers have recently turned to developing more specialised corpora that can be of assistance to the study of politeness in institutional settings. The Michigan Corpus of Academic Spoken English (MICASE) project is one example of a specialised corpus. This corpus contains texts that can give insights into acts of politeness within specific academic settings; and as individual texts and settings are given, it also allows for a closer examination of genre strategies and power issues.

The case of MICASE

Literature on MICASE

Since its launch on the internet in 1999, the MICASE corpus has yielded countless studies (see http://lw.lsa.umich.edu/eli/micase/publications.htm): from investigations into sentence-initial ellipses, which were found to be hardly used in research speech (Swales 2002), to formulaic expressions used by professors and students (Simpson, 2002), and numerous other topics. Many of these studies have used multiple approaches for examining the corpus; an example is the publication *English as a GloCalization Phenomenon* (Pérez-Llantada and Ferguson 2006), which contains articles on MICASE as a ‘linguistic microcosm’ based on various theories: genre theory and pragmatics, corpus linguistics, discourse analysis and sociolinguistics (Gómez 2007). Similarly, techniques for examining the corpus have ranged from quantitative corpus-based methods of word frequency and collocations to qualitative analyses of selected texts. Some of the findings relate, either directly or indirectly, to politeness ‘speech’ in academic settings. For example, using a corpus-based approach, Mendis (2002) found that in the MICASE corpus instructions were not associated with imperatives but with a variety of forms, e.g. hedges, indicative of politeness strategies to mitigate FTAs. Ann Mauranen (2002) observed that in MICASE direct disagreements or criticisms (positive FTAs) were hard to find. But she also pointed out that the markers of these - in contrast to those of praise - are difficult to detect and interpret in word lists. As negative evaluations tend to be highly specific and context-dependent, they would first call for the reading of whole texts before corpus-based approaches can be applied to the patterns discerned (Mauranen 2002).

The literature suggests that an eclectic approach to the study of politeness is warranted and that a closer reading of specific texts can assist in finding patterns that support or refute the assumptions politeness theories propose. The following section presents an examination of two texts in MICASE.

Examination of two MI-CASES

MICASE provides several texts pertaining to advising sessions which give information about interlocators’ institutional membership and power status. The two cases selected are hereafter referred to as MICASE 1 and MICASE 2. At the level of genre, such discourse involves three overlapping scripts: problem formulation, resolution, and explanation (Decitre, Grossi, Jullien and Solvay 1987). These scripts consist of eliciting the needs/capabilities of the person seeking advice (diagnosis period), making a decision on a possible solution to the problem (directive period), and discussing agreed actions in support of the solution (report-writing period) (see Agar, cited in Bardovi-Harlig and Hartford 1993: 234).

In MICASE 1 (file ID: ADV355SU094) the session takes place in the office of a senior faculty member in the field of Linguistics (S1: male, age 51 and over), who advises a senior undergraduate student (S2: female, age 17-23) on matters of independent study. In MICASE 2 (file ID: ADV285SG135) the session is between a senior faculty member in the field of Education (S3: male, over 51 years of age) and a senior graduate student of that department (S4: male, between 31-50 years of age). All are native speakers of English. Role co-membership in MICASE 1 is asymmetrical and defined since S1’s institutional status is higher, reinforced by the meeting taking place in his office (territorial space). Socially, they may share some commonalities, but they differ in age and gender. The advisor, S1, takes on his expected role and initiates the session with “well first thing is I found that article”, which suggests the student needs advice on locating relevant readings. But as the conversation develops, more issues emerge. Most advice is forthcoming in the middle to last part of the conversation, with agreed actions not explicitly stated. In MICASE 2 both social and role co-membership are less defined, as both participants have expertise, are familiar with the field, male and closer age-wise. Their somewhat symmetrical status is reinforced by the location - a small conference room (shared/neutral space). The graduate student (S4) ‘starts’ the session with “mkay”, taken up by the faculty advisor “okay, Jeff … bringing us up to date”. The latter sets the tone for the dialogue, opening the floor to S4, who is given individuality by being called by name (positive face). This dialogue also resembles the advice scenario as problem exploration is elicited by S3, explored by S4 (Jeff), and then linked to a ‘deeper’ problem by S3, who in the end provides a suggestion for action to the problem Jeff had earlier signalled (a mild negative FTA).

How does the differences in status and situation influence politeness behaviour? A closer look at MICASE 1 indicates that S1 employs linguistic ‘power’ markers congruent with his role, such as asking questions and giving (indirect) directives and suggestions. At times, S1 also asks direct questions that request information (often a marker of male speech), whereas S2 hardly poses any questions, and if she does it is mainly for clarification. As most of S1’s questions do not invite S2 to elaborate, her positive face needs are not fully attended to (Mullany 1999). A frequency analysis, using Michael Barlow’s MonoConc Pro (2003), reveals nine occurrences of “you can”, six of “you could”, four “would/let me recommend”, and seven “if you …”. The transcript shows that all are made by S1 and none by S2. It indicates that S1 employs a relative high number of indirect directives (suggestion and advice). On the surface, S1’s use of these strategies is to minimise imposition (threats to negative face) and empower S2’s independence needs. However, from a critical perspective, S1 may actually tell her what to do, giving her direct responsibility to act on the ‘advice’, whilst being, himself, “absolved of responsibility for the consequences” (Youmans 2001: 68). These politeness strategies typify Anglo (middle-class) speech (Youmans 2001) and match the ideal of the autonomous learner. Here, S2 is indeed the one who promises to take independent actions to further her research, indicated by four incidences of I’ll: such as “I’ll start talking to …/contacting …”, “I’ll do that/find those books/be able to get this out”. Generally, the impression is S1 does not want to impose his advice, and having the higher status employs more discernment politeness strategies.

In MICASE 2, S3’s use of power markers is less noticeable, and the interaction ‘feels’ more like a shared exchange between equals. S3’s questions invite S4 to elaborate. Also, S3 interprets what S4 is saying and checks this with him. S3 hardly talks, but listens, especially in the beginning. He supports S4’s continuation by using “mhm” (37 of 40 occurrences) frequently. Only halfway the session does he take longer turns when he directs Jeff’s ‘poem’ problem to his own thesis process (return to expected role). Little or no “you + modal verb” or other variations of advice/suggestions mark S3’s speech (5 occurrences, 2 from S3). Instead, more “we + modal verb” (27 occurrences) is used, which is indicative of shared experiences and common identity. Indeed, *we* is often used by both (128 occurrences), reflecting high co-membership and addressing of each other’s positive face needs. There is also high use, especially by S4, of “you know, well, sort of, I think, I mean” (121, 46, 22, 21, 19 occurrences respectively), which mark informal conversations and are largely pathic (Bazzanella 1990). Moreover, the transcript suggests that both have meta-cognitive awareness of power in speech events. An example is that S4 clearly states a facilitating preference in his dealings with mentors.

S1’s control of the interaction in MICASE 1 is further reinforced through acts like interruptions, turn taking, and length of turn. Whereas S3 and S4 hardly interrupt each other and S3 gives S4 ample space to elaborate, S1 interrupts S2 more frequently than S2 does, and breaches appropriate turn-taking rules as S2 is blocked from completing her sentences. As S1 is mostly successful in claiming his turn, his interruptions are not attending to S2 negative face needs (Mullany 1999). Furthermore, when S1 has his turn he tends to keep it for a slightly longer time than S2, again indicating that he tries to control the session. Yet, S2 asserts her ‘control’ by keeping social distance (putting S1 in place) and remaining on-task (status-preserving strategies) when S1 uses positive politeness strategies. For example, S1 uses *we* 23 times (S2 never does), as a marker of asserting common ground. It should be noted, however, that the use of *we* is ambiguous and points to the need that utterances need to be analysed in context and over longer instances (and that corpus analysis is incomplete on the basis of frequency data). *We* can be seen as referring to S1 and the institution and may not necessarily include S2. However, S1 does use *we* to be inclusive on some occasions, indicating that he is attending to S2’s positive face. He also deviates from the task of giving advice by sharing opinions or experiences. To these strategies S2 responds, however, by guiding the conversation back to her study or topic of discussion or ignoring some of S1’s remarks. In doing so, S2 performs FTAs to S1 positive face needs. There are frequent instances that she is quite direct, stating what she wants to do. An example is when she says: “so the task is …. they will be … but I just left them like that…”. To which S1 respond: “sure, okay, that’s fine”. S2 further uses adverbs like definitely to make clear her thoughts and action. Yet, she also employs ‘mitigators’ such as, “I am not sure/wondering”, “I could do …/I would like …”, but is again pretty direct towards the end of the session. In contrast, S4 pays more attention to S3’s face needs. He is less direct as his use of phatic speech markers such as “you know” (18 of 18 occurrences) indicate, gives compliments and offers apologies at times, and attends to FTAs by using permission markers. Hence, the MICASE transcripts tend to indicate that social closeness leads to increased politeness, contrary to Brown and

Levinson's prediction, and that more politeness strategies are used by graduate students as roles are less clearly defined, which is congruent with Bardovi-Harlig and Hartford's suggestion.

References

Bardovi-Harlig, K., and **Hartford, B.S.** 1993. "The language of comembership." *Research in Language and Social Interaction* 26/3: 227-257.

Barlow, M. 2003. *Concordancing and corpus analysis using MP 2.2*. Houston: Athelstan [pdf file].

Barlow, M. 2004. "Software for corpus access." In *How to use corpora in language teaching*. J.M. Sinclair (ed). Amsterdam: John Benjamins, 205-221.

Bazzanella, C. 1990. "Phatic connectives as interactional cues in contemporary spoken Italian." *Journal of Pragmatics* 4: 629- 647.

Bou-Franch, P. 2002. "Misunderstandings and unofficial knowledge in institutional discourse." In *Culture and power: Ac(unoffially)knowledging Cultural Studies in Spain*, D. Walton & D. Scheu (eds). Bern: Peter Lang, 323-341.

Brown, P., and **Levinson, S.** 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Candlin, C.N., and **Hyland, K.** 1999. *Writing: Texts, processes and practices*. Essex: Addison Wesley Longman Limited.

Decitre, P., Grossi, T., Jullien, C., and **Solvay, J.-P.** 1987. "Planning for problem formulation in advice-giving dialogue." *EACL*: 186-190 http://acl.ldc.upenn.edu/E/E87/E87-1031.pdf [Access date 09/09/2006]

Gómez, I.F. 2004. "I think: opinion, uncertainty or politeness in academic spoken English?" *RAEL: Revista Electrónica de Lingüística Aplicada* 3: 63-84.

Gómez, I.F. 2007. "Bookreview: English as a gloCalization phenomenon. Observations from a linguistic microcosm*." IBÉRICA* 14: 167-192.

Harris, S. 2003. "Politeness and power: Making and responding to 'requests' in institutional settings." *Text* 23/1: 27-52.

Haugh, M. 2003. "Anticipated versus inferred politeness." *Multilingua* 22: 397-413.

Lee, D. 2001. "Genres, registers, text, types, domains and styles: Clarifying the concepts and navigating a path through the BNC Jungle." *Language Learning & Technology* 5/3: 37–72.

Mauranen, A. 2002. *'One thing I like to clarify …' Observations of academic speaking.* http://www.eng.helsinki.fi/hes/Corpora/one_thing.htm [Access date 07/02/2008]

Mendis, D. 2002. "How do you give instructions when instructing? Evidence from a corpus of academic speech." Paper presented at the *4th North American Symposium on Corpus Linguistics and Language Teaching*, Indianapolis, November 1-3, 2002.

Mullany, L. 1999. "Linguistic politeness and sex differences in BBC Radio 4 broadcast interviews." *Leeds Working Papers in Linguistics and Phonetics* 7:119-142.

Paltridge, B. 2000. *Making sense of discourse analysis.* Gold Coast: Antipodean Educational Enterprises.

Pérez-Llantada, C., and **Ferguson, G.R. (Eds.)** 2006. *English as a gloCalization phenomenon. Observations from a linguistic microcosm.* Valencia: Publicacions de la Universitat de València (PUV), English in the World Series, 3.

Scollon, R., and **Scollon, S.W.** 1995. *Intercultural communication: A discourse analysis*. Oxford, England: Blackwell.

Simpson, R. 2002. "A corpus-based study comparing students' and professors' use of formulaic expressions." Paper presented at the *4th North American Symposium on Corpus Linguistics and Language Teaching*, Indianapolis, November 1-3, 2002.

Simpson, R.C., Briggs, S.L., Ovens J., and **Swales, J.M.** 2002. *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.

Swales, J.M. 2002. “Any last minute thoughts on this particular search? The occurrence of sentence-initial ellipsis (SIE) in research speech.” Paper presented at the *4th North American Symposium on Corpus Linguistics and Language Teaching*, Indianapolis, November 1-3, 2002.

Terkourafi, M. 2005. “Beyond the micro-level in politeness research.” *Journal of Politeness Research* 1: 237-262.

Watts, R.J. 2003. *Politeness.* Cambridge: Cambridge University Press**.**

Youmans, M. 2001. “Cross-cultural differences in polite epistemic modal use in American English.” *Journal of Multilingual and Multicultural Development* 22/1: 57-73.

A DATA-DRIVEN APPROACH TO LEARNING AND TEACHING PHRASEOLOGY

Martin Warren[225]

Abstract

The phraseological tendency, or what Sinclair (1987) terms 'the idiom principle', refers to the way in which words are co-selected by speakers and writers. To fully describe the meaning and use of language we need to be able to identify and describe these word co-selections. However, phraseology is rarely foregrounded in learning and teaching activities in language classrooms. This paper outlines a new methodology to uncover word co-occurrences in a corpus and explains how to use the results to enable students to attain a better understanding of English phraseology.

*This paper presents a new way of searching for and describing phraseological variation. Using new software, ConcGram©, it is possible to find fully automatically all word co-occurrences that have constituency variation (e.g. A*B, A**B) and/or positional variation (e.g. BA, B*A), in addition to n-grams. The search results are termed 'concgrams' (Cheng, Greaves and Warren, 2006).*

The learning and teaching activities described in the paper are based on analysing the concgrams in a engineering text, then a specialised corpus of engineering English, and finally comparing them with those of a general corpus of English.

Keywords: phraseology, concgram, Engineering English, English for Specific Purposes, data-driven learning

Introduction

In recent years, there has been an increasing awareness of the importance of phraseology in English language description. In this paper the term 'phraseology' is used broadly and refers to the more-or-less fixed co-occurrence of linguistic elements (Hunston, 1995). Corpus linguists examining co-occurrences found in linguistic patterns have contributed to our understanding of the fact that when we speak and write, on most occasions, we select words in combination. This is termed the 'the idiom principle' (Sinclair, 2004a: 29), i.e. the phraseological tendency, whereby words are co-selected rather than being selected separately constrained only by grammar. These co-selections are now starting to be given space in new corpus-based grammars of the English language (see, for example, Biber et al., 1999; Carter and McCarthy, 2006), but have yet to be foregrounded, especially in language learning and teaching. Exceptions to this general observation are recent textbooks on phraseology and collocation (see for example, McCarthy, 2005; Sinclair, 2003 and Stubbs, 2002). Up until now most attention has been on the most frequently occurring contiguous word associations, n-grams, which are also termed 'clusters' or 'bundles'. This paper argues that greater emphasis should be placed on the learning and teaching of phraseology and applies a new computer-mediated research methodology to introduce and promote the learning and teaching of phraseology.

The paper describes learning and teaching activities which enable learners and teachers to raise their awareness of patterns of phraseology. This study therefore builds on the work of others who have advocated the use of corpora and corpus linguistics in language learning in general (see for example, Aston, 1997; Bernadini, 2002; Braun, 2005; Kennedy and Miceli, 2002 and Sinclair, 2004b) and the use of concordancing in particular (see for example, Bernadini, 2000; Cobb, 1997; Gaskell and Cobb, 2004; Johns, 1991; Sinclair, 2003 and Stevens, 1991).

Concgrams

Current searches for n-grams generate phrases made up of contiguous word associations, such as 'different people', but miss instances of the same phraseological pattern when it is realised in instances such as 'different

[225] Works in the Department of English at the Hong Kong Polytechnic University and teaches and conducts research in the areas of discourse analysis, discourse intonation, corpus linguistics, intercultural communication, lexical studies (especially phrase ology), and pragmatics.

kinds of people' or 'different types of people'. In other words, n-gram searches are only helpful in finding instances of co-selection that are strictly contiguous in sequence. The result is that many instances of word associations may be overlooked, and phrases that typically, or on occasion, occur in non-contiguous sequences risk going undiscovered. These limitations of n-gram searches have led to the recent development of searches for gapped n-grams or 'skipgrams' (see Wilks, 2005). Skipgrams include a certain amount of constituency variation (i.e. AB and A*B) of up to three intervening words.

Cheng, Greaves and Warren (2006) describe new software, ConcGram©[226], which is able to extract recurrent concgrams (i.e. sets of between 2 and 5 co-occurring words) fully automatically, within a wide span (up to 12 words on either side of the origin[227]), and which include all of a concgram's configurations irrespective of any constituency (e.g. AB and A*B) and positional variation (e.g. AB and BA) present. As a result, the associated words of a concgram may be the source of a number of patterns. The software has been designed to perform fully automated concgram searches, but the user can enter between one to five words as a user-nominated concgram search query. The fully automated capability of the search engine further increases the possibility that the searches will help to uncover not only a more extensive description of known patterns of co-selection and their meanings, but also new patterns.

1 next five years. We also estimate that operating **expenditure** will **increase** moderately, at a rate commensurate

2 in the Consolidated Account for 2008-09. Public **expenditure** as a proportion of GDP will **increase** from 15.9

3 two and a half times the present population. The **expenditure** on the Old Age Allowance will **increase**

4 Since the planning of these projects takes time, **expenditure** on infrastructure is unlikely to **increase**

5 health care system were to remain unchanged, **expenditure** on public health care services would z **increase**

6 and social development. We will **increase** **expenditure** on social services and welfare and return part of

7 lead to a decrease in revenue and an **increase** in **expenditure** in 2008-09. I have also earmarked $50 billion to

8 it is expected that the **increase** in overall **expenditure** on health care services will, on average, be two

9 is revenue by $33.5 billion and **increase** operating **expenditure** by $41.5 billion in 2008-09. The latter figure

10 measure will **increase** government **expenditure** on CSSA payments when inflation rises. 145

11 be sustainable. If we **increase** recurrent public **expenditure** or reduce recurrent public revenue, we must be

12 Government cannot **increase** public health care **expenditure** indefinitely, we hope that supplementary

13 we will **increase** the share of public health care **expenditure** to 17 per cent of government recurrent

Text 1: Sample concordance lines of the two-word concgram 'expenditure/increase'

[226] ConcGram© is written and developed by Chris Greaves, Senior Project Fellow, English Department, The Hong Kong Polytechnic University.

The above example of a two-word concgram illustrates the extent of the phraseological variation uncovered by the software. Both constituency and positional variation are clearly shown for the user to then analyse.

Data

The data used in this study consist of an engineering research article, a specialised corpus of engineering English and a general reference corpus. The engineering research article (Xu, Ng, Chen and Qu 2003) is taken from an engineering journal, Transactions of the Hong Kong Institution of Engineers, and contains 4,810 words. The article is part of the Hong Kong Engineering Corpus (HKEC) which contains approximately one million words of English Engineering texts collected in Hong Kong. The British National Corpus (BNC), comprising 100 million words, is used as the general reference corpus.

Learning and teaching activities

The paper presents replicable language learning activities that raise learners' awareness of the prevalence and importance of phraseology. The activities help to develop in learners the skills needed to conduct an initial study of the phraselogical profile of a text or corpus. Before doing the activities, the students need to be introduced to the broad notion of phraseology and trained to use the Concgram© software.

The suggested activities all work towards an initial determination of the aboutness of a text from its phraseological profile and also the identification of genre-specific phraseology. An outline of the activities is given below.

- Compile a list of the ten most frequent lexical words in the text.
- Compile a list of the ten most frequent lexically-rich two word concgrams in the text.
- Monitor and record the frequency with which the most frequent lexical words and concgrams found in the text are also found in the Hong Kong Engineering Corpus and the BNC.
- Discuss your findings.

The above activities are influenced by data-driven learning, DDL, (Johns, 1991) and the work on keyness by Scott and Tribble (2006). The concgramming of the texts, or corpora, should take place outside of the classroom and it is best to ask the students to form small groups to make the activities more interactive and collaborative.

1 (or computer) modelling of a building, framing **design**, **structural** analysis, component **design**, design

2 process framing plan (d) represents the final **design** of **structural** framing. One may further carry out

3 and cost-effectiveness. The **structural design** of a tall building involves several rather

4 instance, considering the preliminary **structural design** of a building, after the structure model is

5 engineers to achieve not only a safe **structural design**, but also a cost-effective design in terns of

6 has actually been applied to the **structural design** of more than 25 building projects with the

7 this model, one may perform **structural** framing **design** with assigned initial dimensions for all the

8 objectives of the preliminary **structural** framing **design** are: 1) To find which partitions are efficient to

9 reduced occupied **structural** space, and shorter **design** time, have been realised. Acknowledgements T

10 **structural** analysis, optimisation, automated **design** check, and cost analysis, one may easily

Text 2: Sample concordance lines for the two-word concgram 'design/structural'

One of key issues for the students is to be able to decide whether or not the words in a concgram are meaningfully associated or simply co-occur. In the above concordance of 'design/structural', lines 1, 9 and 10 are examples of 'design and 'structural' not being meaningfully associated with each other, unlike the instances in lines 2-8 which are meaningfully associated in a 'meaning shift unit' (Sinclair 2007), which Sinclair previously termed a 'lexical item' (1996 and 1998). The need to distinguish between co-occurrence and association is important because students need to exclude instances of co-occurrence and only count those which are associated.

Concgrams which represent the aboutness of a specific text or genre are termed 'aboutgrams' (Sinclair, personal communication, 2006). While the phraseological profile of a text is arrived at by identifying all of the word associations in it, the aboutness of a text is determined by a process in which the most frequently occurring lexical concgrams are placed on a provisional aboutgram list. This list is then referred to a specialised corpus of engineering texts. Those which are found to occur more frequently in the specialised corpus, are removed from the provisional aboutgram list. The same process is then repeated using a general corpus. The result is a list of aboutgrams which represent the aboutness of the text. The most frequent aboutgrams in the text are listed below.

Aboutgram
MR damper(s)
building(s)/storey
control/damper(s)
logic control
semi-active/control
control/MR
control algorithm
control/passive
building/response(s)
semi-active logic

Most frequent two-word aboutgrams in the engineering text

Those concgrams which are found to be frequent not only in the article but also in the HKEC are provisionally assigned as genre-specific aboutgrams. These concgrams are then searched for in the BNC to determine whether this is the case. Examples of engineering-specific, rather than text-specific, aboutgrams are listed below.

Aboutgram
design(s)/structural
structural model(s)
building(s)/design
architectural/model(s)
structural analysis
design/tall
data capture
structural optimisation
analysis/design
form/structural

Examples of two-word aboutgrams in the HKEC

Conclusions

There are three main areas in which concgramming could be used when using language corpora in learning and teaching. First, as a tool for textual analysis, for example, in terms of determining the aboutness of a text. Second, to help raise learners' awareness of the importance of phraseology. The third area, which leads on from the second, is the use of concgramming to help language learners to acquire the phraseology of specialised fields and their specific genres.

References

Aston, G. 1997. Small and large corpora in language learning. In B. Lewandowska-Tomaszczyk and J. P. Melia (eds.), *Practical applications in language corpora*. Łodz: Łodz University Press, 51-62.

Bernardini, S. 2000. "Systematising serendipity: Proposals for concordancing large corpora with language learners". In L. Burnard & T. McEnery, T. (Eds.), *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang, 225-234.

Bernardini, S. 2002. "Exploring new directions for discovery learning". In B. Kettemann & G. Marko (eds.), *Teaching and learning by doing corpus analysis*. New York: The Edwin Mellen Press, 165-182.

Braun, S. 2005. "From pedagogically relevant corpora to authentic language learning contents". *ReCALL 17(1)*: 47-64.

Cobb, T. 1997. "Is there any measurable learning from hands on concordancing?" *System, 25 (3)*: 301-315.

Biber, D., Susan C., Edward F., Johansson, S. and Leech, G. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Carter R. and McCarthy, M. 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Cheng, W., Greaves, C. and Warren, M. 2006. "From n-gram to skipgram to concgram". *International Journal of Corpus Linguistics* 11 (4): 411-433.

Gaskell, D. & Cobb, T. 2004. "Can learners use concordance feedback for writing errors?". *System, 32 (3)*: 301-319.

Hunston, S. 1995. "A corpus study of some English verbs of attribution". *Functions of Language*, 2/2, 133-158.

Johns, T, 1991. "Should you be persuaded: two samples of data-driven learning materials". In T. Johns and P. King (eds.) *Classroom Concordancing*. English Language Research: Birmingham University, 1-16.

Kennedy, C. and Miceli, T. 2002. "The *CWIC* project: Developing and using a corpus for intermediate Italian students". In B. Kettemann & G. Marko (eds.), *Teaching and learning by doing corpus analysis*. New York: The Edwin Mellen Press, 183-192.

McCarthy, M. 2005. *English Collocations in Use*. Cambridge: Cambridge University Press.

Scott, M. and Tribble, C. 2006. *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Sinclair, J. McH. 1987. "Collocation: a progress report". In R. Steele and T. Threadgold (eds.) *Language Topics: An International Collection of Papers by Colleagues, Students and Admirers of Professor Michael Halliday to Honour him on his Retirement*, 319-333. Volume III. Amsterdam: John Benjamins.

Sinclair, J. McH. 1996. "The search for units of meaning". *Textus* 9 (1): 75-106.

Sinclair, J. McH. 1998. "The Lexical Item". In E. Weigand (ed.) *Contrastive Lexical Semantics*, 1-24. Amsterdam: John Benjamins.

Sinclair, J. McH. 2003. *Reading Concordances*. London: Longman.

Sinclair, J. McH. 2004a. *English Collocation Studies*. London: Continuum.

Sinclair, J. McH. 2004b. *Trust the Text*. London: Routledge.

Sinclair, J. McH. 2005. Document Relativity. (manuscript), Tuscan Word Centre, Italy.

Sinclair, J. McH. 2007. Collocation Reviewed. (manuscript), Tuscan Word Centre, Italy.

Stubbs, M. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford:

Blackwell.

Wilks, Y. 2005. "REVEAL: the notion of anomalous texts in a very large corpus." Tuscan Word Centre International Workshop. Certosa di Pontignano, Tuscany, Italy, 31 June – 3 July 2005.

Xu, Y., Ng, C.L., Chen, J. and Qu, W. 2003. "Innovative Technology for Seismic Response Reduction of Tall Buildings with Podium Structures". *Transactions of the Hong Kong Institution of Engineers* Vol. 10/4: 88-94.

THE SACODEYL SEARCH TOOL – EXPLOITING CORPORA FOR LANGUAGE LEARNING PURPOSES

***Johannes Widmann*[228]**
***Kurt Kohn*[229]**
***Ramon Ziai*[230]**

Abstract

SACODEYL is an ongoing EU project that has created pedagogically motivated spoken language corpora in seven European languages: English, French, German, Italian, Lithuanian, Romanian, and Spanish. The corpora consist of structured video interviews of 13-17 year-old secondary school pupils. After the transcription phase, the corpora have been annotated with the needs of language learners in mind. Then the video recordings have been aligned with the transcript sections. Enrichment materials, such as learning packages, have been produced and they have also been aligned with the relevant sections.

This paper will focus on the development and the features of the SACODEYL search tool. This is one of the four tools that have been developed in the project. After a development phase of one year, the search tool is being used in a piloting and validation phase at the moment. The paper will present the design of the search tool, and show how the tool enables the teacher or language learner to retrieve all the annotations and materials in a pedagogically useful way. It will also discuss its innovative features for corpus access and exploitation and it will show how these features can be used in language learning scenarios of secondary schools to create a fruitful, authentic, and varied learning environment.

Keywords: pedagogic corpora, authentic language learning, autonomous language learning, search tool development, spoken corpora

Introduction

The EU-funded SACODEYL project has created pedagogically motivated spoken language corpora in seven European languages: English, French, German, Italian, Lithuanian, Romanian, and Spanish. The corpora consist of video-recorded interviews with teenagers, and these interviews are all equally structured, and based on (roughly) the same topics. All corpora have been annotated with the needs of language learners in mind, and the video recordings have been aligned with the transcript sections. In addition, enrichment materials, such as learning packages, have been produced and have been attached to the relevant sections. All this has been done to help learners and teachers to exploit the corpora in various ways and from various angles. From a theoretical language learning perspective, the aim has been to support the authentication process of the learners (cf. Widdowson 2003) and to support a learning environment based on constructivist principles (cf. Rüschoff 1999).

In this paper, we will first discuss the design principles of the search tool and briefly review some technical decisions that we took and the reasons for them. In the second section we will introduce all the features of the search tool and explain what their characteristics and relative strengths are. The next section will be about the kind of usage scenarios that we intend the tool to be used in. The paper will finish with a conclusion that will put this work-in-progress report in a larger perspective of what is desirable for future developments.

Search tool design principles

[228] Johannes Widmann is a junior researcher who is working on his PhD in Applied English Linguistics. He teaches linguistics at the University of Tübingen with a focus on teacher education. He has contributed to the Socrates-Minerva SACODEYL project and to the EU LdV EntecNet and CIP projects (see http://www.ael.uni-tuebingen.de/projects/index.html for details).
His interests are in computer-aided language learning (CALL), using LMSs such as Moodle to support blended learning scenarios, and corpus-based materials production. He has been working as a freelance language teacher for English and French for many years.

[229] Kurt Kohn is professor of Applied English Linguistics at the University of Tübingen; he is a member of the steering committee of "Sprachen & Beruf" Düsseldorf and of the advisory board of the International Journal of Applied Linguistics (INJAL). His research interests include theoretical and empirical issues of second/foreign language learning and teaching, technology-enhanced language learning and teaching (CALL, multimedia, e-learning), pedagogic corpora, content and language integrated learning (CLIL), English as a Lingua Franca, processes and strategies of translation. Publications include books and articles in these areas. Since the early 1990s, he has been involved in numerous EU-funded projects both as co-ordinator and partner.

[230] Ramon Ziai has a BA in computer linguistics and is an expert in JAVA server-based programming applications. He has a special interest in corpus-based language learning and web-based search tools. With this expertise he has been involved as lead programmer in the EU Minerva project SACODEYL.

A survey of the existing search tool software made it clear that almost all existing solutions started with a focus on concordance-based searches. The reason for this is clear when you look at the development of corpus linguistics and the origin of concordance application in the area of lexicography. The tools were designed for professional linguistic research and the production of dictionaries and reference books. Their impressive functionality can be quite daunting for the novice user and it usually takes some time to come to grips with the software. However, for use in the classroom these tools did not seem an ideal starting point to us (see Braun 2007 for a classroom study that supports this hypothesis). Most teachers do not feel comfortable teaching by just providing concordance lines or frequency lists, and students may have serious problems with understanding KWIC concordances. Both groups need additional ways to access the corpus. Thus, we decided to design the search tool from scratch for two reasons. First, we wanted to have a tool whose functions were motivated by the needs of language learners and teachers. Second, we wanted to have a tool that requires little or no installation and that is easily available on all computers without administrator rights, since ease of access is one of the issues for most teachers when they decide whether or not to use an eLearning tool (cf. Kohn/Glombitza/Helbich 2008).

Pedagogical design decisions

In order to achieve the first aim of pedagogical usability we added browse and view functions to the tool that do not rely on “vertical” reading, such as is the case with concordance lines, but rather support the “horizontal” reading mode that students and teachers are used to. In order to support this browse and view mode of corpus access efficiently, all the corpus transcripts have been given summaries in their metadata. Moreover, the transcripts have been structured into topic-based sections that were given titles in the annotation process. In the browse mode, these titles are displayed as well as the interview summaries. In all other search modes, the search can also be limited to certain sections only, based on topic filters and other annotation filters. These functions for corpus exploitation are more in line with the "horizontal" reading that teachers and students are used to. They also allow for an easier construction of a discourse context in which the individual search results might make sense. Furthermore, we wanted the search tool to be able to also display all the learning packages or other resources that were developed to enrich a specific corpus. This kind of corpus enrichment and the various access modes to the corpus make the process of searching and exploring it a more multi-modal activity that offers much more than just concordances and word lists.

Search tool - technical decisions

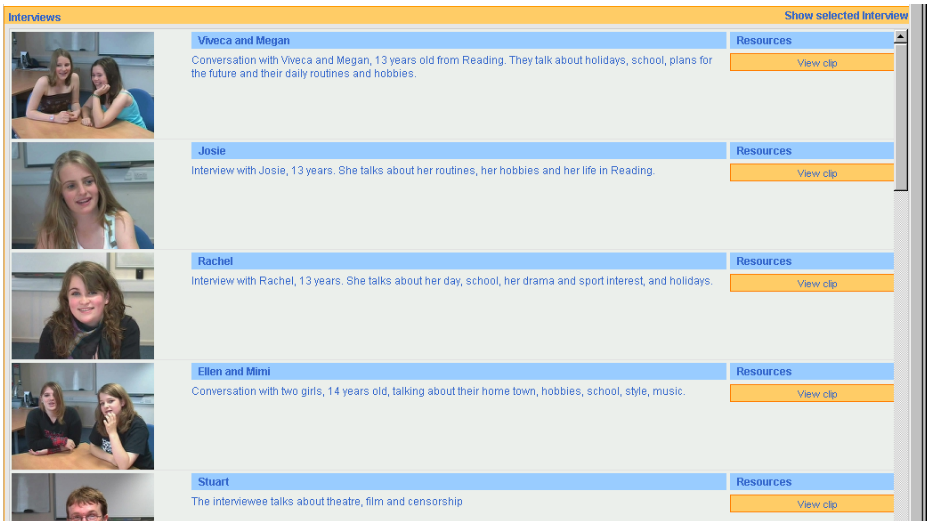
In order to achieve the second aim, the search tool runs online in all browsers that support JavaScript. This important feature makes it very easy to use, as there is no software installation needed on the client side, i.e. on the computer of the user. All you need is a web browser and an internet connection. All seven corpora of the project can be accessed freely. On the server end, it runs with standard Java Enterprise technology and can be set up on any standard server in a Java Servlet Container. All the tools that are being developed in the project are made available as open source tools, either for download or as web tools[231]. The data of the corpus files are stored using XML technology, and the coding format follows the TEI P5 guidelines. We chose this format in the hope of having a data format that is highly versatile and reusable in different contexts (cf. Ward 2002).
The video clips have been saved in RealMedia format which is a popular streaming format[232]. RealMedia uses the XML-based Synchronized Multimedia Integration Language (SMIL) technology that enables the search tool to jump to pre-defined points within the video. This is a very useful feature when you have a particular section in your search results and you want to show exactly this section to your students.

Search tool features

Based on the design decisions, the search tool offers four different modes of corpus access that can be classified according to increasing granularity. The first and broadest way of corpus access is the browse mode. In this mode, the users get an overview of all interviews that are included in the corpus. The users can browse through the summary descriptions that the annotator has given to each interview. They can also view the full text of each interview where the section titles are displayed as well.

[231] All the software that has been developed in the SACODEYL project is open source under the GPL and can be downloaded from www.um.es/sacodeyl. The search tool can be reached at www.purl.org/sacodeyl/searchtool

[232] The RealPlayer can be downloaded for free at http://germany.real.com/player/win/.



Interviews

Show selected Interview

Viveca and Megan

Conversation with Viveca and Megan, 13 years old from Reading. They talk about holidays, school, plans for the future and their daily routines and hobbies.

Resources

View clip

Josie

Interview with Josie, 13 years. She talks about her routines, her hobbies and her life in Reading.

Resources

View clip

Rachel

Interview with Rachel, 13 years. She talks about her day, school, her drama and sport interest, and holidays.

Resources

View clip

Ellen and Mimi

Conversation with two girls, 14 years old, talking about their home town, hobbies, school, style, music.

Resources

View clip

Stuart

The interviewee talks about theatre, film and censorship

Resources

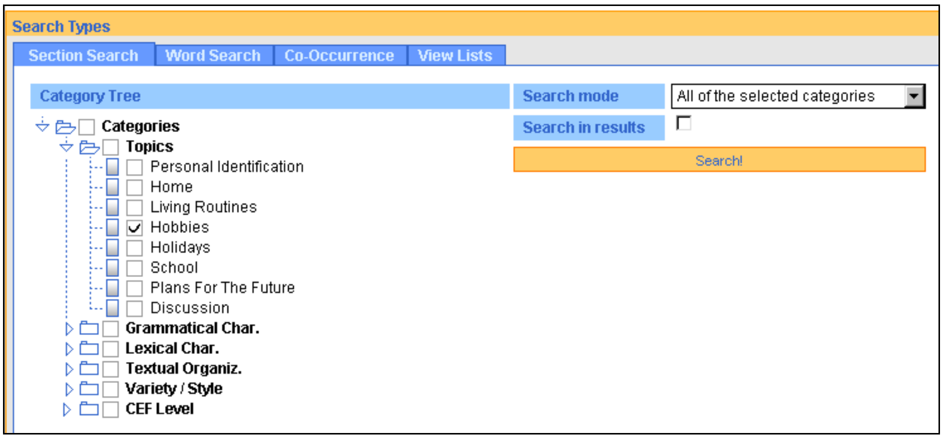
View clip

Fig.1: The browse mode view of the English corpus with a picture and an interview description.

Furthermore, the full video clips for each interview can be viewed in this mode. They open in a separate window. Thus, the teachers can easily zoom the clip and if they want to show the clip to their class they can show it in full screen mode. If they are done with the video, they simply go back to the web browser where they can choose further clips.

The section search mode

Next to the browse mode, there are three other search modes. All of these modes can be accessed simply via changing tabs (see Fig. 2). The first of these modes is the section search mode. This mode allows users to search for certain sections based on what has been annotated. All the categories that the annotators included during the annotation process can be freely combined as search filters. The users can select whether all of the selected categories must be included in the search results or whether they want all sections with at least one of the selected categories. Based on a search result, they can also do follow-up searches with new filters to further specify the sections that they get. In SACODEYL, we defined categories based on topics, grammatical characteristics, lexical characteristics, textual organization, variety/style, and Common European Framework (CEF) level. The topic and the CEF categories are applied at section level only, while the other categories are applied to sections, but they can also be applied more specifically to individual words or phrases, if desired.



Search Types

Section Search | Word Search | Co-Occurrence | View Lists

Category Tree

- □ Categories
  - □ Topics
    - □ Personal Identification
    - □ Home
    - □ Living Routines
    - ☑ Hobbies
    - □ Holidays
    - □ School
    - □ Plans For The Future
    - □ Discussion
  - □ Grammatical Char.
  - □ Lexical Char.
  - □ Textual Organiz.
  - □ Variety / Style
  - □ CEF Level

Search mode: All of the selected categories

Search in results: □

Search!

Fig. 2: The different search modes are arranged on tabs in the browser window

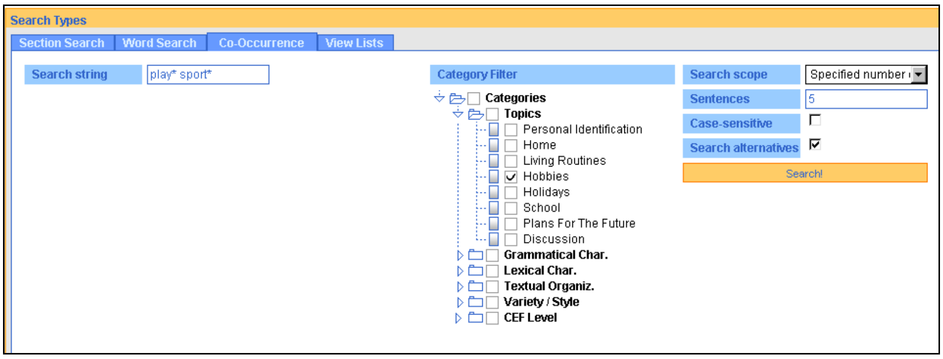
The search results are displayed as full sections with the section titles, as in the browse mode. Moreover, the search tool displays the categories that are returned for each section in the search results and it also highlights the specific words or phrases where the annotations have been applied within a section. Table 1 shows some examples of the categories that the annotators of SACODEYL created during the annotation process. These categories are not hard-wired into the search tool. Corpus annotators are free to create or delete their own categories. The search tool will display all categories that the annotators have created during the annotation process.

Category type	**Examples**
Topics	personal identification hobbies plans for the future discussion topics
Grammatical characteristics	tenses passives modality conditionals determiners/quantifiers adjective/adverb comparison
Lexical characteristics	topic specific terminology typical collocations Idiomatic expressions
Textual organization	basic cohesive ties pragmatic markers prosodic markers
Variety/Style	casual/informal language typical of spoken language teenage/youth language taboo/vulgar language
CEF level	A1, A2, B1, B2, C1[233]

Tab. 1: The annotation categories of SACODEYL project

The co-occurrence mode

The second search option is the co-occurrence search. In this mode you can enter two or more words and search for the places where they appear in the corpus. You can freely define the span in which the search words are supposed to appear (search scope). This search scope can range from 1 sentence to the whole interview. The mode is not a collocation search in a statistical sense because in our tool it is the users who define the search scope based on the familiar notions of "whole interview", "one section", and "number of sentences". As in the section search, it is also possible to apply category filters to restrict the co-occurrence search to certain sections only. In the results of the search, the search tool will display the entire text fragments as defined in the search scope.



Search Types

Section Search | Word Search | Co-Occurrence | View Lists

Search string: play* sport*

Category Filter

- □ Categories
  - □ Topics
    - □ Personal Identification
    - □ Home
    - □ Living Routines
    - ☑ Hobbies
    - □ Holidays
    - □ School
    - □ Plans For The Future
    - □ Discussion
  - □ Grammatical Char.
  - □ Lexical Char.
  - □ Textual Organiz.
  - □ Variety / Style
  - □ CEF Level

Search scope: Specified number

Sentences: 5

Case-sensitive: □

Search alternatives: ☑

Search!

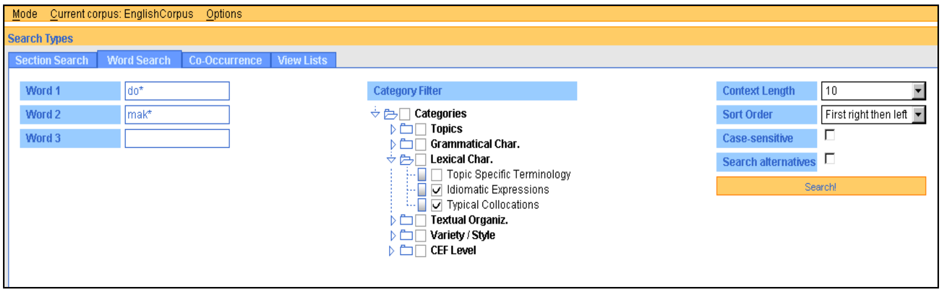
Fig. 3: Specification of a co-occurrence search for [play*] and [sport*].

[233] The project consortium did not create C2 materials.

For example, if you carry out the search in Fig. 3, it would yield a pretty good impression of all the different kinds of sports that the teenagers talk about in the corpus. As all search results range over 5 sentences, it is still possible to read a bit what the section is actually about and to be able to understand the context of the search words. This supports the authentication and the discourse comprehension process enormously. It can also serve as a preparation for concordance-based exercises because in this mode you already get a feeling for what it means to look at non-consecutive text fragments, but you still have more co-text for orientation than in pure concordance lines. In this search mode as well as in the word search mode that is described in the next section you can use familiar wild cards, such as the asterisk (*) to replace any number of characters between two spaces and the question mark (?) to replace one single character.

The word search mode

This third search option is closest to the traditional KWIC concordancers. While our word search mode is much less powerful than most other linguistic concordancing software, it offers some carefully selected features that users can apply to filter and sort the search results: First of all, it is possible to limit the search to certain sections by e.g. using the topic filters, just as in the section and co-occurrence search modes. Secondly, you can define the context length between 5 and 15 words to the left or right. The alphabetical sorting of the results can be done to the right or to the left of the KWIC. You can then decide whether the search is supposed to be case sensitive or not, which is very important for languages such as German where capitalization can result in a meaning distinction or where capitalization can be used to distinguish nouns from verbs. Finally, you can decide whether to search for spelling variations that were entered in the transcription phase, such as the spelling variation "going to/gonna", "because/'cos". This is important because we decided to include a standard spelling variant to all words where the transcribers decided to use a transcription variant that is phonetically closer to what you hear in the video recording. All of these options have default values, so novice users have to take no decisions except to type in their search words. In subsequent searches, they can always change the default values with their own values. Fig. 3 shows an example of a word search that focuses on the problem of using "do/make" idiomatically in English. For this the search words [do*] and [mak*] are entered and the category filters "Lexical Characteristics - Idiomatic Expressions" + "Typical collocations" are used. In order to spot typical patterns easily, the sort order "first right then left" is chosen, so the primary sort is to the right of the KWIC.



Word 1: do*
Word 2: mak*
Word 3:

Categories
- Topics □
- Grammatical Char. □
- Lexical Char. □
  - □ Topic Specific Terminology
  - ☑ Idiomatic Expressions
  - ☑ Typical Collocations
- Textual Organiz. □
- Variety / Style □
- CEF Level □

Context Length: 10
Sort Order: First right then left
Case-sensitive: □
Search alternatives: □

Fig. 4: Word search for [do*] and [mak*] with lexical filters and right-sort turned on.

While the other 3 search modes are more useful for zooming in on thematic topics and annotation categories, the word search mode is most useful to detect lexical patterns and the use of certain words. So the 4 approaches are complementary to each other with each one having its specific advantages. Table 2 gives a quick overview of the features of the individual tools:

Mode	***Advantages***
Browse mode	- Gives a quick overview of the available texts and allows for a whole text access to the corpus ("horizontal" reading) - Gives a quick summary of each interview in the corpus - Allows watching the full videos of all interviews
section search mode	- Allows to zoom in on topic-specific sections and on all the other annotation categories that are included in the corpus - Allows for follow-up searches to limit the section results with further

	categories - Returns section-based results with titles that still allow for some "horizontal" reading

co-occurrence search mode	- Allows searching for several words within a certain span - Returns span-based results that still allow for some "horizontal" reading - Gives a good overview of the lexical content of the results - If needed, the span can always be expanded to the full section
word search mode	- Allows for word pattern searches - Limit word pattern searches to certain topics or combine them with other lexical, grammatical, or register categories - Search for phrases with wild cards - If needed, the KWIC can always be expanded to the full section

Tab. 2: The different search tool modes and their advantages.

SACODEYL usage scenarios

The SACODEYL approach is geared to the needs of secondary school education. To this end, the interviews were structured according to those topics that feature in secondary school curricula. So all interviews have roughly the same topic progression. The interviews start with the interviewees presenting themselves, their families, and their homes. They go on with talking about their hobbies or their favorite past time activities. Next, they talk about their everyday routines and their school experiences. These are all typical topics that feature in regular school curricula of beginners to intermediate level. Almost all interviews include sections where students talk of their past holidays or of their plans for the future. Both of these topics are very useful for studying the use of tenses. At the end of the interviews there are often discussion questions where the interviewees are asked to voice their opinions and where the answers are open. These topics are more suitable for advanced learners. So all of the topics can be seamlessly integrated into existing regular secondary school curricula. The teachers do not have to do extra "SACODEYL lessons" but they can use the materials to teach the classes they would have been teaching anyway. The only change is that they will have a more varied approach to teaching and they have authentic materials available for their thematic topics. SACODEYL does not, however, prescribe one particular way of using the corpora and the enrichment materials. This is still the decision of the teachers.

SACODEYL with its focus on spoken language is of course most useful in lessons and exercises where the focus is on listening comprehension or on speaking and communicating. To this end, a range of exploratory and communicative exercises are being developed in the project that can serve as a guide to teachers for using the SACODEYL materials. The video interviews feature different accents and different speech rates, making it an ideal resource for listening comprehension activities at different levels of proficiency.

However, the corpora can be useful for writing activities as well. The search tool will return all the learning packages that have been developed for the individual interviews. These learning packages include various exercises for written production where students have to paraphrase typical spoken language utterances or where they have to write their own opinions on certain topics. So even though all transcripts are based on spoken language, it is possible to use SACODEYL for reading and writing tasks.

Although SACODEYL as a project has had its focus on secondary school education, the approach itself is open to be used in any other kind of topic domain or learning scenario. As mentioned above, the annotation categories can be defined freely, so any other kind of annotation could be applied to the corpus transcripts. With the SACODEYL tools, it is also possible to develop new corpora that are based on different topics altogether. The search tool can process all corpora that comply with the TEI XML format. An interesting example of a similar topic-based approach is the ELISA project, a forerunner of the SACODEYL project.[234] This project has developed comparable interviews that are based on topics of professional life. These interviews are more appealing to adult education learning scenarios and they could be relatively easily converted into SACODEYL corpora.

Conclusion

[234] See www.uni-tuebingen.de/elisa

The SACODEYL search tool is a first step in the direction of a pedagogically viable type of corpus exploitation that does not require too much technical expertise on the one hand while on the other hand it draws systematically on the insights of corpus research and tries to mediate and apply them to language teaching based on pedagogical principles. With its focus on teenager language it matches squarely the requirements of an authentic classroom where students are engaged in language and tasks they are really interested in.

The concept of topic-based sectioning of the transcripts enables teachers and learners to focus on specific topics more easily than it has been the case with professional concordancing software. The SACODEYL search modes allow for a balanced focus between the search for topics and the search for relevant words and phrases for these topics. The topic-based approach is also helpful for a communicatively oriented teaching approach where teaching structures is embedded in topic-driven tasks. At the moment, the exploratory topic-driven concept of the exercises is tested in a piloting phase where the different components of the search tool and the learning packages are validated in schools in four European countries (Germany, Lithuania, Romania, and Spain).

References

Braun, S. 2007. "Integrating corpus work into secondary education: from data-driven learning to needs-driven corpora." *ReCALL* 197/3: 307-328.

Kohn, K., Glombitza, A., Helbig, G. (2008). "Perceived Potential of ICT in European Schools - A Survey Report." 2nd Report of the EU Comenius Network EcoMedia. http://www.ael.uni-tuebingen.de/downloads/index.html [Access date: 25/05/2008]

Rüschoff, B. 1999. "Construction of knowledge as the basis of foreign language learning" In *The construction of knowledge, learner autonomy and related issues*, B. Mißler/U. Multhaup (eds.).Tübingen: Stauffenberg, 79-88.

Ward, M. 2002. "Reusable XML technologies and the development of language learning materials" *ReCALL* 14/2: 285-294.

Widdowson, H. 2003. *Defining Issues in English Language Teaching*. Oxford: Oxford UP.

TO BUILD A PRESCHOOLER'S ORAL CORPUS IN SINGAPORE: IMPLEMENTATION, APPLICATION AND IMPLICATION*[235]

Shouhui Zhao[236]

Yongbing Liu[237]

Hock Huan Goh[238]

Abstract

This paper reports some important issues arising from implementing a nearly completed public-funded educational research project entitled "An Investigation of Chinese Oral Competence of Singaporean Preschoolers: A Corpus-driven Study". In this paper, we will report on our experiences in building an education oriented oral corpus as well as the application significance of the Corpus. Following a brief description about the general research background of the project, drawing upon practical experience in the compilation and application process of the Corpus, a range of methodological issues involving data collection, transcription/cleaning, annotation and corpus construction will be explored and reflected, including a succinct discussion of the most difficult part – word segmentation and parts-of-speech identification of child's oral output. To further a critical reflection of the conventional application of computer corpus as a descriptive tool for linguistic purposes, the paper closes with an elaboration of the implications of building such an education-oriented linguistic computer corpus for wider potential applying areas in a multilingual society like Singapore.

Keywords: Oral/spoken corpus, Singapore, Chinese, home/family language, language planning (LP)

Background and Objectives

Sociolinguistic and Educational Context

Singapore has developed and implemented a policy of a "English-knowing" bilingualism through which English has grown as *de facto* national language, meanwhile Chinese, Malay and Tamil are defined in specific Singaporean terms as official mother tongue maintained and taught as school subjects with the key objective of enabling direct access to cultural traditions and the related values of the Singaporean ethnic communities. However, given the fact that Singapore is an English yet dominantly Chinese society in a Malay world in which Chinese language (CL) serves as *lingual franca* for 78% of the community, concern has arisen for the increasing imbalance in the bilingual ability among Chinese children, due to a shift in the use of home language (HL) from Mandarin to English. Therefore, over recent years, the Singaporean bilingual policy in language-in-education planning revolves largely around "the CL problem", with milestone review reports that mark the trajectory of CL development in Singapore's educational system.

[235] This paper makes use of data from the research project "An Investigation of Chinese Singaporean Children's Oral Competence in Mandarin - To construct at aural oral corpus and to generate a wordlist" (CRP 11/05 LYB), funded by the Centre for Research in Pedagogy and Practice, National Institute of Education, Singapore (http://www.crpp.nie.edu.sg). The authors wish to express their gratitude to the project team members. The views expressed in this paper are the author's and do not necessarily represent the views of the Centre or the Institute.

[236] Shouhui Zhao (PhD, University of Sydney) is an Assistant Professor at the National Institute of Education (NIE), Singapore. His academic interests are language education and planning. Zhao has been teaching and researching at 7 universities in five countries around Asia-Pacific region for 20 years since he completed his postgraduate study at Renmin University of China (Beijing).

[237] Yongbing Liu (PhD, university of Queensland) is the Direction of the Depart. of English Language and the Vice-Dean of the School of Foreign Studies at the Northeast Normal University (China). His academic interests are English language education and discourse analysis.

[238] Hock Huan Goh is a Phd candidate at the National Institute of Education (NIE), Singapore. His research interests include language competence and bilingualism. He graduated from the National University of Singapore with an MA and is currently a Research Associate in the Center for Research in Pedagogy and Practice (NIE).

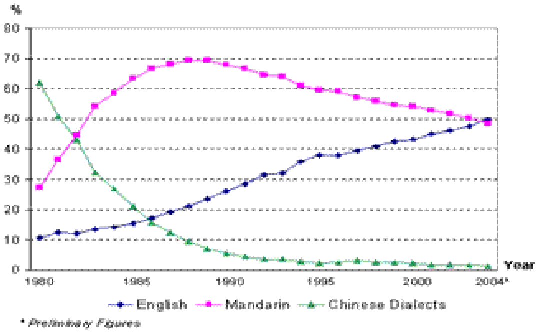


Figure 1: Dominant HL of Chinese P1 Students: 1980 to 2004 (MOE, 2004)

The most recent review report, released in 2004, which provides the charter for CL teaching in the foreseeable future, gives even greater flexibility and choice to the less proficient students. Government statistics has shown a sharp increase in the number Chinese Singaporeans who prefer English as their HL. The above figure shows the statistics obtained from the information provided by parents at Primary 1 registration. The review Committee recognized the existence of two major distinct groups of children from two different linguistic backgrounds, starting to learn CL in Singaporean schools. Based on this recognition, the Committee recommended that a new “Modular Approach” to the Chinese curriculum and pedagogy should be developed and implemented. This approach consists of four level differentiated modules (Core, Bridging/Reinforcement, Enrichment and School-based) to accommodate the different linguistic backgrounds of children who predominantly speak English at home (ESF) and those who predominantly speak Chinese at home (CSF).

Educational Objectives

To find out the extent of children from two family language backgrounds are being different from each other at the time of entering Primary 1, in order to provide baseline data for developing the CL (as mother tongue) curriculum for primary schools in educational innovation, a large-scale research project investigating Chinese Singaporean children’s oral competence in Mandarin, led by the authors of this paper, was launched. So far a multimodal spoken corpus of Singaporean Chinese preschoolers is partially completed (http://score.crpp.nie.edu.sg/score/) and a Wordlist on Preschooler’s Oral Mandarin (by both frequency and part-of-speech) has been generated. The children’s vocabulary attainments have also been analysed by employing quantitative measurements.

Methodological Issues: Data Collection

Sampling

Sampling was based on assumption that the linguistic representations of the population are closely tied to social and demographic factors. Apart from demographic balance (such as age, gender, etc.), two factors were emphasized in population sampling: the governance or funding sources of kindergartens and childcare centres and their geographical distribution and coverage.

Currently, preschool services in Singapore are run in the form of tripartite model according to the funding sources: public, church-affiliated and private. Therefore a random sample obtained from these three types of kindergartens in Singapore should broadly reflect the sociolinguistic diversity of Singapore, and proportional to the size of the sector relative to the total number of kindergartens/childcare centres across the island. In the current study, our sample of 1,200 participants is made up of an equal number of boys and girls aged at 5 and 6 years old from 20 public childcare centres and kindergartens, 10 church-affiliated ones and 6 private ones were selected.

Parents' Survey

For the purpose of sociolinguistic analysis, the parents of potential subjects were surveyed using a bilingual self-report questionnaire. The questionnaire consisted of 22 sociodemographic questions with focuses on the family language input and socioeconomic variables such as parents' education level, occupation and type of housing. The information obtained from the questionnaire not only allowed the researchers to check that the attributes of the drawn sample are representative of the population provided a second chance allowing the researchers to make the necessary adjustment in order to make the sampled population more desirable, but the rich background information also facilitated the interview/elicitation process. A total of 953 valid survey forms were processed.

Data Collection

Three research instruments for data collection were employed in this study, namely, students' interview, picture elicitation (narrative) and classroom observation. Further elaboration on data collection means, which is an important part of oral corpus construction, and data validity concerns is not possible due to length constraint, interested readers are referred to Zhao, et al (2007). Briefly, the data collection instruments were designed in accordance to the following principles:

- able to elicit utterances to the maximum;
- operationally viable and easy to use/manage;
- consistent, with minimal interference by researchers stable and duplicable, likely to get similar results by other researchers;
- suitable and effective for different groups of children with equal reliability and validity.

Data Handling and Corpus Compilation: Implementation

600 cases of valid high quality interview/picture elicitation (300 hours) and 24 cases of video-taped classroom observation (12 hours) were transcribed. About 320 hours orthographic transcription was automatically processed by the software specially developed by Corpus Construction Research Team at the Center. Although high accuracy of linguistic segmentation was achieved, some human intervention was still necessary to assure the quality, considering the characteristics of Singaporean CL and children's oral expressions. The corpus was then both automatically and manually tagged and processed with software (e.g. TACT, Concordance) to generate the oral wordlist which includes the coverage and frequency of the vocabulary in the children's Mandarin repertoire.

As the ultimate task of our project is to capture and analyze the Mandarin of Singapore preschool children, we faced the issue of how to process the data them for scientific and systematic analysis. Some major phases and the difficulties we encountered are briefly reported as follows:

Text Segmentation Standards and Word Identification

Word determination is considered as the greatest challenge in corpus linguistics building (e.g., Gardner 2007). Overcoming the lack of linguistics device and a universally agreed-upon standard for processing the oral language is a paramount task as CL is a non-inflectional language and it has been notorious for the contentiousness among Chinese linguists on even the most basic composite units. To establish an operational method of breaking down dividing transcribed spoken data into basic units is the first step of corpus construction. Identifying word is a relatively easy task in the written language, but is often problematic in oral speech, where secondary language factors and non-language factors may hinder the system by giving uncertain clues. Moreover, the data we are dealing with is the spoken speech by young children in a linguistically pluralistic community, thus making the segmentation and identification task even more complex.

The word segmentation standard used for the present corpus is Processing Standard for Modern Text Corpus Segmentation and Annotation-973 (http://www.chineseldc.org/

EN/doc/CLDC-LAC-2003-003/label.htm). Words in the corpus were segmented into twenty broad categories of parts of speech which were defined by Standard-973. It should be noted that while strictly adhering to the prescribed standards, necessary modifications were made during text cleaning which is a major undertaking in addition to transcription cleaning as discussed below.

Following word segmentation, word classification also proved to be another challenging task. Determining the parts-of-speech in Chinese language has long been one of most debatable topics in Chinese linguistics. Three criterion to identify the parts-of-speech including lexical meaning, syntactic function (through replacement) and morphology (through both overt and covert 'grammatical markers' in broad sense) were employed in this study.

Protocol and Conventions for Data Collection, Management and Transcription

To best capture the children's utterances as well to ease the tracking of data with demographic details, we stipulated a set of protocol and conventions for managing the audio- and video-recordings with close reference to the interview procedure and classroom settings. In order to ensure smooth processes in building the corpus and generating analyses, a set of customized transcription conventions to further standardize the orthographic features of Mandarin and specific transcription contents is also very essential.

Transcription of collected data is the fundamental and most important step in the entire corpus building process. With reference to the CRPP Transcription Process Volume 2, we drew up a set of standards for the transcription of CL which defines in detail on areas such as denotation of speakers, segmentation of audio output of spoken text, presenting instances of overlapping speech between interviewer and children, and symbols to denote common speech events like 'laughter', 'background noise', 'ungotten talk' and quasi-lexical vocalizations.

We also drew up a set of Transcription Conventions based on the CRPP Transcription Process Volume 3 which mainly covers two areas: i.e. language-specific conventions and context-specific conventions. The former involves standards for transcribing numerals, misspelled and localized words and terms, and using of Pinyin (Chinese alphabetic transliteration system), etc. and the latter deals with spacing between Chinese characters and English words when they co-exist, standardized comments for some non-audible expressions (i.e. silence after Interviewer's questions, possible physical expressions in silent turns, etc.), and standard procedures to best recover utterances due to bad recording, etc.

Data Cleaning and Verification

Besides having a well-defined transcription standards and conventions to guide transcribers to ensure the accuracy of transcriptions, the selection of a transcription tool that is easy to use is also important. For this reason, we adopted the CRPP standard transcription software with customized event settings standardized by the CRPP Corpus Team according to the above-mentioned transcription standards and conventions. In the transcription process, transcribers use this software to convert the entire audio recording (of an interview) into a collection of short audio segments separated by breakpoints, whereby synchronized text is input with the corresponding waveform of each segment. This proved to be a time-consuming and labor-intensive phase of this project, as transcribers have to literally segment each speaker's utterance, and input (or copy) them word-for-word. And errors like typo-errors and missing utterances were bound to occur in the transcription process. Therefore, to reconcile such issues, team members of this project had to check through every single transcription during the reviewing and data cleaning phase. This cleaning procedure is undertaken by team members (who collected the data) to make sure that the transcription accurately reflects the contents of the interviews and observations.

After the cleaning process, an automated segmentation and tagging application was then run to break down the utterances into words and tag the words with the relevant parts-of-speech. As our corpus merely consists of speech data, some speech features like stuttering or incomplete speech had poised some technical challenges in the segmentation and tagging process. To counter these problems, the project team ran several rounds of verification exercises on the database using the Concordance software to identify mis-tagged items. The identified errors are then fed back to the Corpus Team to manually correct the tags and refined their application.

Lastly, with the machine-readable database tagged and verified, the Corpus Team then ran various query applications to distill the children's utterances from the interviewers' utterances, to generate the required wordlist based on frequency count and parts-of-speech. The following is the workflow for building this corpus:

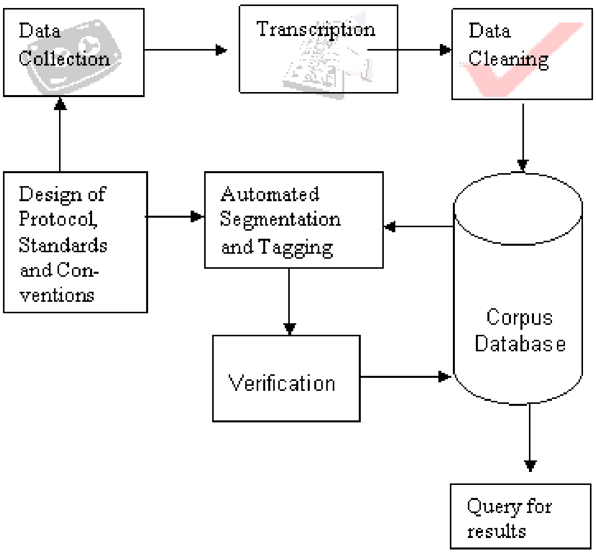


Figure 2: Process for Corpus Building (Adapted from Liu, et al, 2007)

Data Analysis: Applications

In what follows, we will briefly examine the correlations between child's linguistic competence and his/her family language background through investigation into three dimensions of language use.

Home Language Definition

As described in the introduction, of all the factors cited in the previous review reports, the streaming of HL, particularly between ESF and CSF, has the strongest impact on the curriculum development during the recent CL education reform. The key differentiating factors in defining the HL are the parents' self-reports of their own language preferences when speaking to their children, such as Ministry of Education's annual survey conducted during the Primary 1 Registration Exercise.

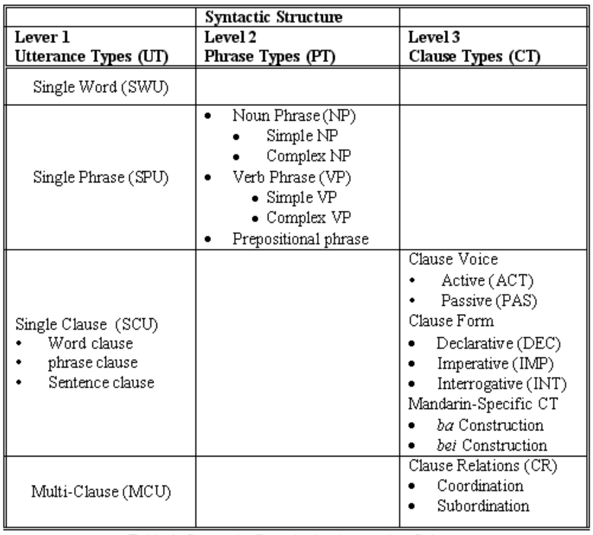
In this study, a new HL categorization approach which we believe can assess a child more accurately was developed. Instead of solely depending on parents' preferred daily languages reported in the survey, this study takes a holistic stance by using a multiple factor approach to define the preschoolers' HL use. Based on the assumption that various forms of ambient influence have different levels of impact in shaping a child's language inclination and verbal ability, the weighting of each of the four major factors was determined by the coefficient correlation between the child's oral productivity and various determining variables in the survey questionnaire. This was done through the calculation of the coefficient of determination which is the square of the correlation coefficient (Pearson *r*), an index showing the degree to which one can predicate one variable from the other in percentage terms (for details about differentiation steps and computation, see Liu & Zhao 2007). By calculating the scores of these four factors, children's language use at home was categorized into ESF, English-Chinese Speaking Family (ECSF) and CSF.

Vocabulary Competence

The initial results of the 600 children's oral production generated by the corpus shows that CSF children achieved the highest, followed by ECSF and ESF. On the average, children from CSF used 283 Chinese types in a half-an-hour interview cum picture elicitation, while ESF children used only 151 words. If the talkativeness (i.e., volume of oral production) is considered an indicator of children's language ability, we can see from the token count that the difference among the two groups is equally remarkable – CSF children (1256) are also much productive than their ESF counterparts (747). Sentence ratio and MLU (Mean Length of Utterance) are also computed in this study to investigate assess the children's linguistic skills in conversational fluency and comprehension ability (for details, see Zhao et al. 2007). These two forms of measurement yield differences that are not statistically significant enough to support the HL categorization in terms of statistical differences did not support HL categorization to a very high degree. The moderate variations within each group and across three groups seem to suggest that CL competence gap among these groups of children may not turn out to be as drastic as we previously predicated.

Syntactic Complexity

Syntactic complexity was examined using the following adapted annotation scheme by Yaruss (1999):



	Syntactic Structure	
Lever 1 **Utterance Types (UT)**	**Level 2** **Phrase Types (PT)**	**Level 3** **Clause Types (CT)**
Single Word (SWU)		
Single Phrase (SPU)	• Noun Phrase (NP) • Simple NP • Complex NP • Verb Phrase (VP) • Simple VP • Complex VP • Prepositional phrase	
Single Clause (SCU) • Word clause • phrase clause • Sentence clause		Clause Voice • Active (ACT) • Passive (PAS) Clause Form • Declarative (DEC) • Imperative (IMP) • Interrogative (INT) Mandarin-Specific CT • *ba* Construction • *bei* Construction
Multi-Clause (MCU)		Clause Relations (CR) • Coordination • Subordination

Table 1: Syntactic Complexity Annotation Scheme

Goh and Liu’s (2008a) corpus-driven analysis shows that the syntactic complexity of children among the three groups did not vary much in terms of the utterance length of single clauses. However, in terms of multi-clauses, children from ESF did show obvious limitations in producing longer utterances, as compared to the other two groups of children.

From the above results, their study concludes that the relation between language competence and language exposure is not as simplistic or clear-cut as hypothesized previously. Further syntactic annotation showed that children from all three groups are not differentiable in their competence in terms of phrase types, clause voice and clause form. They generally lack advancement in complex noun phrases, passive voice contractions and imperative and interrogative clause forms. This finding will be an important reference for curriculum designers to look into for further development of language skills for children of all language-speaking groups. Apart from these common areas for further reinforcement in the curriculum, the observed differences among children of different groups will have to be noted by educators, so that assistance can be targeted at their specific needs like increasing the utterance length of multi-clause among ESF children; and developing heir chunky coordinate multi-clauses into more fluent coordinate or even subordinate complex constructions.

The analyses of syntactic complexity found that children from differentiated HL background showed some differences in utterance types, but did not vary much in phrasal and clausal types, suggesting that Mandarin curriculum developers and educators will have to look into developing these lacking phrasal and clausal types as noted in the findings.

Code-switching (CS)

Though CS is widely understood as a common sociolinguistic feature of bilingual societies, such phenomenon in Mandarin still troubles educators as issues, concerns or even barriers in the learning of Mandarin in formal education. Besides reiterating the negative effect of CS, educators rarely address this phenomenon systematically or academically. Another study by Goh and Liu (2008b) use conversational turns as the measurement unit to describe CS in the corpus data. With the pool of data, the transcript of each child is annotated with the following three-level annotation scheme which is created from CS categories adapted from Poplack (1980) and Muysken (1997) and others as showed in the table.

Lever 1	Level 2	Level 3
Intra-Utterance CS (Intra)	Alternation (Alt) Insertion (Ins) Congruent Lexicalization (Cong)	Noun (N) Verb (V) Adjective (Adj) Conjunction (Conj) Preposition (Prep)
Inter-Utterance CS (Inter)		

Table 2: Description of CS Annotation Scheme

For the CS Intensity, the findings confirm the correlation between HL background and CS intensity. When it comes to the types of CS, generally speaking, their study finds that children from each HL group do not vary much in the distinction of types of utterance and types of intra-utterance CS; about the Common linguistic content of CS, the results seem to show that there may be a general lack of Content Words (especially nouns) in the children's Mandarin lexicon, and ESF and CSF children also show a more prominent lack of conjoining words (conjunction).

They thereby believed that educators need not be overly concerned with CS in spoken Mandarin, as its overall intensity is relatively not too high. Furthermore, the use of alternative language-code is a common communicative tactic/strategy among bilinguals. Educators probably need to give ESF children more support in view of their higher CS intensity. And they will have to look into ways to expand the children's repertoires for nouns and verbs, with an emphasis on conjunction for ESF and CSF children. The study found that this phenomenon showed close relation with HL background. The study concludes that such CS provides a resource for children less competent in CL, and it needs to be considered for CL education.

Concluding Remarks: Implications

Computer corpus application is a long-standing tradition in linguistic study as a descriptive tool. In this study, we have endeavoured to show an emerging phenomenon in language-in-education (one of four components in LP theories) and general education as a whole. For these two areas, education-oriented (linguistic) corpus usually involves for the purposes of informing. One level of informing is that of informing instruction for pedagogical purpose, a second level is identification of trend or eligibility of decision made at policy level.

The major thrust of this project was originally intended for curriculum development, i.e. to provide evidence showing that the new module-based curriculum innovation is justifiable in terms of its designated goals to serve the respective students with different language proficiency stemmed from their exposure to different linguistic environment at home. The corpus-based linguistic inquiry in Syntactic Complexity and Code-switching shows that higher levels of application can be achieved, thus can contribute towards providing empirical guidelines (via in-depth analysis) for pedagogical practice in the classrooms; more significantly, it goes further from curriculum and pedagogy to policy level.

The rich background information tagged with children's language ability for the first time affords a holistic stance in judging HL. The empirically critical analysis on more influential factors impinging children's language ability with reference to their actual linguistic ability appears to convince the authors that the complexity of HL has been oversimplified and the language differences in preschool years is over emphasized, or exaggerated. The findings seem to suggest there is wide-stretched continuum of CL competence and there are different cutting points in this continuum in terms of the relationship between linguistic ambience at home and students' actual level of language use. But the simplistic differentiation of ESF and CSF as a binary dichotomy, as Zhao et al (2007) argued, "overlooks the permeable boundaries of rather complex situations and the developmental nature of the language ability of young children".

One of the remarkable features of LP over recent years has been the use of empirical evidence derived from either large scale sociolinguistic survey or statistic data (e.g., Baldauf 2002; Baker 2006). However, partly because of its complex design and extreme high costs, the generation and utilization of Chinese spoken corpus has lagged far behind the written corpus (Gu 2004). Our literature review reveals that over the last two decades, a number of paper-and-pen based oral Mandarin wordlist in China and Singapore (e.g., Shi 1993; Ong 2002) or database in HK (Tse 2006) have been compiled, but the application of paper data has been severely limited by the lack of reliability and repeatability of previous works. The present Corpus will not only be used as an important researchable reference for designing assessment guide and evaluation benchmarks of curriculum development by educational authorities, but also to inform decision-makers in the area of LP (language-in-education) on the whole. In addition to that, this is the first time in the Chinese-speaking world that a children's spoken corpus of this scale is being constructed. We hope the description of our experience in corpus building and data analysis will contribute to the knowledge base from which recommendations of best practices in spoken corpus compilation may emerge.

References

Baldauf, R. B. Jr. 2002. "Methodology for policy and planning". In *The Oxford Handbook of Applied Linguistics*, R. Kaplan (ed.). New York: Oxford University Press, 391-403.

Baker, C. 2006. "Psycho-sociological analysis in language planning". In *An Introduction to Language Policy*, T. Ricento (ed.). Malden, MA. USA: Blackwell Publishing, 210-228.

Gardner, D. 2007.Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics* 28/2: 241–265.

Goh, H. H. and **Liu, Y. B.** 2008. Spoken Mandarin Competence of Chinese Children from Different Language-speaking Homes: Implications for Mandarin education from corpus-based analysis of Syntactic Complexity. Paper presented at the *American Educational Research Association (AERA) 2008 Annual Meeting*, New York, March 24-28, 2008.

Goh, H. H. and **Liu, Y. B.** 2008. Code-switching in Mandarin of Chinese Preschoolers – Implications to Mandarin Teaching for bilingual Chinese children. Paper presented at the *American Association for Applied Linguistics (AAAL) 2008 Conference*, Washington, D.C., March 19 - April 1, 2008.

Gu, Y. G. 2007. Sampling situated discourse for spoken Chinese corpus. http://www.ddyyx.com/guyueguo/paper/sampling_situated_discourse.pdf (Access date 16/04/2008)

Liu, Y. B., Goh, H. H. and **Zhao, S. H.** 2007. *Transcribing Chinese Language Classroom Talk: To Build a Computer Corpus*. Technical Report, CRPP, NIE, Nangyang Technological University.

Muysken, P. 1997. Code-switching processes: Alternation, insertion, congruent lexicalization. In *Language Choices: Conditions, Constraints, and Consequences*, M. Platz (ed.). Amsterdam; Philadelphia: J. Benjamins, 261-280.

Ong Y. P. 2002. Zhonghua yuyan wenhua yu jiaoxue [Chinese Language, Culture and Teaching]. Singapore: Xishan Wenyi Zhongxin.

Poplack, S. 2000. 'Sometimes I'll start a sentence in English y termino en espanol'. In *The Bilingualism Reader*, W. Li (ed.). London, New York: Routledge, 221-256.

Shi, H. Z. 1990. 3-6 sui ertong de yuyan fazhan yu jiaoyu [The development and education for children aged between 3-6 years old] in *Zhongguo Ertong Qingshaonian Xinli Fazhan Yu Jiaoyu* [Psychological Development and Education of Children, Adolescence and Youth in China], Z. X. Zhu (ed.). Beijing: Zhongguo zhuoyue Chubanshe [China Zhuoyue Press], 94-127.

Tse, S. K. 2006. *Xianggang Youer Kouyu Fazhan* [Child Language Development in Hong Kong]. Hong Kong: Xianggang Daxue Chubanshe [Hong Kong University Press]

Yaruss, J. S. 1999. Utterance length, syntactic complexity, and childhood stuttering. *Journal of Speech, Language, and Hearing Research* 42, 329-344.

Zhao, S. H. and **Liu, Y. B.** 2007. The home language shift and its implications for language planning in Singapore: From the perspective of prestige planning. *Asia-Pacific Education Researcher* 16/2, 111-126.

Zhao, S. H, Liu, Y. B. and **Hong, H. Q.** 2007. Singapore Preschoolers' Oral Competence in Mandarin, *Language Policy* 6/1, 49-62.

POSTERS

APPLICATION OF CORPUS LINGUISTICS TO EFL TEACHER EDUCATION IN CHINA

Anping He[239]

Abstract

This paper reports the application of corpus linguistics to EFL teacher education in South China Normal University in the past 10 years. Focus is laid on solving problems including: where to get relevant corpus for learners, how to design corpus-aided activity for daily teaching goals, how to make corpus manageable in classroom teaching and how to enhance teaching practice by corpus research.

The practice is featured in four aspects: 1) Constructing and sharing EFL pedagogical corpora by both learners and teachers, making it as a component in teacher education courses. 2) Starting from textbook-corpus analysis and resulting in corpus-aided exercise design, taking it as one of the goals in teacher education. This includes investigating quantity and quality of textbook input and salient features in different types of exercise design. Linguistic forms and patterns retrieved from the corpus are further associated with pedagogical ideology embedded. 3) Implementing data-driven learning approach in classroom teaching by improving the teaching environment in terms of both soft ware and hard ware. 4) Reflecting teaching effect by corpus-based research.

Some examples of the above are presented and problems are discussed. All this indicates that extension from corpus linguistic research to teaching practice has to be initiated in language teacher education.

Keywords: EFL corpora, joint construction, course book analysis, exercise design, implementation

Introduction

"Teaching is a natural extend of research" (Leech 1997: 3).Yet the application of corpus linguistics to languge education, classroom teaching in particular, is not to an extent as is expected. One reason is that teachers themselves have not had the ideology and technique of corpora. In China, there is a population of 50 millions of learning English as a foreign language. We are now undergoing a new round of curriculum reform in English education, which is aiming at a 'big leap forward' development but based on a low level English proficiency as a whole throughout the country and poor teacher resources. South China Normal University is to educate pre-service and in-service EFL teachers for secondary and tertiary schools in Mainland China and we therefore should play an initial role to bring corpus linguistics into EFL teaching.

In the past 10 years of applying corpus linguistic to EFL teacher education, we have been trying to solve four problems: 1) where to get relevant corpus for teacher trainees, 2) how to design corpus-aided activity matching the ongoing daily teaching goals, 3) how to make huge corpus data manageable in classroom teaching, and 4) how to evaluate teaching effect. This paper is reporting our considerations and practice.

Construction of corpora for EFL teacher education

Together with importing huge amount of English corpora abroad as a reference corpus, we also build our own pedagogical corpora for routine course learning and teaching practice. This is a joint effort of teacher educator and teacher trainees. We make it as part of course practice by requiring in-service and pre-service teachers to bring in their small portion of data when they attend our courses such as "*Basic English Phonetics and Phonology', "EFL Syllabus Design and Teaching Material Development", "Corpus Linguistics & EFL Teaching*" *and "Discourse Analysis*'. The data includes scanned textbook, classroom teaching video transcription and students' written and

[239] He Anping is a professor of English in School of Foreign Studies, South China Normal University and Researcher (part-time) in the Centre for Linguistics and Applied Linguistics, Key Research Institute in Chinese Universities, Guangdong University of Foreign Studies, Guangzhou, China. She received her Ph.D. from Victoria University of Wellington, New Zealand. Her current research interests include corpus linguistics, English curriculum and methodology and discourse analysis. Her articles have appeared in the International Journal of Corpus Linguistics, British Journal of Social Psychology, Journal of Language and Social Psychology, RELC Journal and many CSSCI journals in Mainland China.

spoken performance in every year's examinations. All this is pooled into our *Corpora of EFL Education in China* (CEEC) and open to the trainees to do corpus analysis during the processing of the above courses. The corpora keep expanding every year and it now comes to a size as follows:

- Imported Corpora (300 million)
  - Native English speaker' spoken & written data
    - Adults / teenagers / children
    - English Literature: 3000 classic works …
- Self-built Corpora (9 million）
  - EFL teaching materials：(2.88 million)
    - 120 course books at tertiary, secondary & primary level at home & abroad
  - EFL classroom teaching (0.8 million)
    - 222 classes at tertiary, secondary & primary level at home & abroad
  - EFL learners' inter-language: (5.54 million)
    - Spoken & written data at tertiary, secondary & primary level at home & abroad

Text 1: Structure & Size of CEEC (He 2007)

The imported corpora offer us information on the centrality and typicality of the target language in use, while the self-built corpora connect corpus research findings with daily teaching objective and exercise design. Both contribute to fine resources to EFL education which we did not have ever before.

Design corpus-aided exercises

As information and communication technology (ICT) "can not longer be an added extra but rather an intrinsic part of teacher's methodological repertoire" (O'Keeffe & Farr 2003: 389), we teacher educators took a lead in cooperating corpus resource and techniques in the above courses and then require trainees to conduct corpus based research and exercise design with principles as follows:

1) Select specific teaching goals from current textbooks and further elaborate them as a problem to be solved;

2) Retrieve authentic information from the two types of corpora above either in form of concordances, collocation list, cluster list, wordlist or keyword list;

3) Edit the data and make it acceptable by students in terms of vocabulary and grammar structures;

4) Provide practical guiding instructions, i.e., questions to be answered step by step by different observation focus;

5) Write in teaching notes about possible answers and reasons.

Enlightened by concepts of corpus linguistic such as "frequency driven", "co-text meaning construction" and "lexicalgrammar" (Sinclair, 2004), trainees conduct corpus based research and design corpus-aided exercises with topics include but not limit to the following: "association between pronunciation and spelling", "word constructions", "lexical item with features of collocation, colligation and semantic preference", "micro-language skill training of guessing, catching gist, plotting and discourse structure", "cross-cultural language expressions comparison", "genre analysis", "literature style study". The following are a few examples of corpus-aided exercises based on a unit named *Nelson Mandela* from a popular EFL textbook for middle school in China, ranging from word study to skill training.

Case 1 Word construction

The word structure of "Adj. + *ness* → noun" is a learning target of the unit. By presenting a list of words ending with *ness* which is retrieved from the textbook corpus, students are to observe and answer questions including:

1) Which word is not an Adj. in its original? (→word class identification)
2) What change has to make to the Adj. in spelling after adding *ness* ? (→spelling regulation)
3) Is there any change in the word stress pattern after changing the part of speech? (→pronunciation regulation)
4) Is this type of Adj. sharing some meaning in common? (→ semantic preference)

Case 2 Word collocation

"Came to power" is a verb phrase highlighted in this unit, but the teaching can be extended to the delexicalization of *COME* , i.e., *COME to* + (none physical places). For more details, see those lexical items highlighted in Version 1 below. Guidance for observation includes:

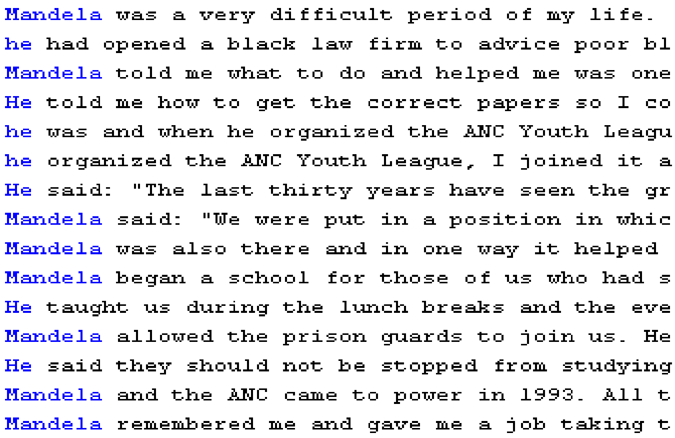
1) Observe nouns just after "come to" and think if they refer to some actual places or to certain situations? Name some of the later.
2) Find how many of these nouns are actually coming from verbs (e.g., "conclusion" is from the verb "conclude")?
3) Paraphrase the sentences by using the verb forms instead (e.g., change "came to a conclusion" into "concluded that …". Try at least 3 of them and think about the differences.
4) Translate some of the sentences into Chinese, such as "come to hand" "come to my mind", *"come to life"…*

Case 3 Retell the life story of Mandela

A keyword list of this unit in relates to the whole textbook reveals the 'aboutness' of the reading materials in the unit referring to Nelson Mandela. The concordances of *Mandela* and *he* demonstrate the major content points and ways of expressions which are useful in retelling this hero's life story. This can be obtained by attending to those action verbs and verbal verbs around the two node words, telling students about what Mandela has done and what he has said (as can be seen below).

No.	**Token**	**keyness**	**word**
1	**53**	**402.352**	**I**
2	**27**	**222.242**	**Mandela**
3	**22**	**181.086**	**Nelson**
4	**19**	**156.393**	**Elias**
5	**14**	**115.237**	**Africa**
6	**20**	**80.038**	**black**
8	**14**	**76.881**	**south**
9	**16**	**74.938**	**prison**
10	**8**	**65.850**	**ANC**
11	**9**	**48.123**	**guards**
13	**39**	**43.252**	**he**
14	**5**	**41.156**	**league**
16	**9**	**41.098**	**workers**
18	**10**	**37.259**	**government**
20	**5**	**35.782**	**youth**

Extract of Keyword list of *Story of Nelson Mandela*



```
Mandela was a very difficult period of my life.
he had opened a black law firm to advice poor bl
Mandela told me what to do and helped me was one
He told me how to get the correct papers so I co
he was and when he organized the ANC Youth Leagu
he organized the ANC Youth League, I joined it a
He said: "The last thirty years have seen the gr
Mandela said: "We were put in a position in whic
Mandela was also there and in one way it helped
Mandela began a school for those of us who had s
He taught us during the lunch breaks and the eve
Mandela allowed the prison guards to join us. He
He said they should not be stopped from studying
Mandela and the ANC came to power in 1993. All t
Mandela remembered me and gave me a job taking t
```

Extract of Concordances of *Mandela* and *he*

Case 4 Pattern and meaning

"--- *because* + (clause)" and "*because of* + (noun phrase)" are grammar patterns in the unit. After comparing the content of the clauses and noun phrases after the two patterns in the format of cluster lists retrieved from both textbook corpus and reference corpus, the designer of the exercise comments:

... we always assume that <u>because</u> + a clause and <u>because of</u> + a phrase can be used alternatively without any change of the flavor, the corpora has helped us to bring out the hidden knowledge of these two: excerpt for their similarities, <u>because of</u> has a negative semantic prosody, and it may be used to 1) express uncertainty of judgment, 2) to bring out undesirable consequences, or 3) to excuse oneself.

This is an in-depth learning of grammar pattern, which can be further extended to the study of "*owing to", "due to"* and *"thanks to"*.

Implementation in classroom teaching

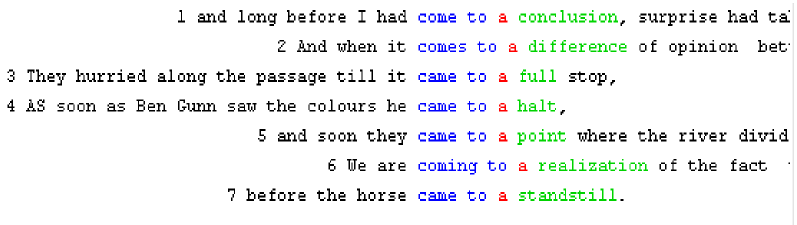
To make the corpus-informed or corpus-based language input manageable in classroom teaching, we have tried to improve the presentation format and technique by:

1) Use mini-file as input file in class, which is a practice modifying Sinclair's tasks design in his book *Reading Concordances* (2003)

2) Use *Antconc* (Anthony 2006) to justify font size, colour highlighting and techniques of cut-past print-screen on to doc. file or PPT file for class presentation.

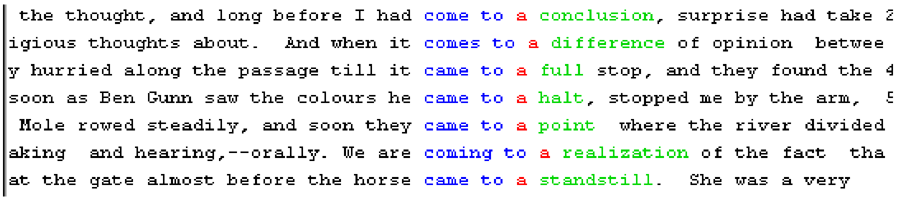
3) Use school inner-net to store all corpora above and offer free access to all users with only output files downloadable but no copy of the whole corpus for copy rights projection.

For example, the following version 1 is proved to be more acceptable than version 2 by middle school students.



```
                    1 and long before I had come to a conclusion, surprise had ta
                              2 And when it comes to a difference of opinion  bet
 3 They hurried along the passage till it came to a full stop,
 4 AS soon as Ben Gunn saw the colours he came to a halt,
                            5 and soon they came to a point where the river divid
                                   6 We are coming to a realization of the fact
                         7 before the horse came to a standstill.
```

Version 1: Extract of edited mini-file



```
 the thought, and long before I had come to a conclusion, surprise had take 2
igious thoughts about.  And when it comes to a difference of opinion  betwee
y hurried along the passage till it came to a full stop, and they found the 4
soon as Ben Gunn saw the colours he came to a halt, stopped me by the arm,  5
 Mole rowed steadily, and soon they came to a point  where the river divided
aking  and hearing,--orally. We are coming to a realization of the fact  tha
at the gate almost before the horse came to a standstill.  She was a very
```

Version 2: Extract of original mini-file retrieved by Antconc.

Reflection and research

After designing exercises or trying them in teaching practicum, trainees are to write reflection reports as course papers. This is a stage to enhance the course practice to a theoretical consideration of teaching materials development and classroom teaching practice. The following are some of the key phrase in the titles of their research papers or reflection reports, covering various aspects of EFL education.

On textbook analysis：

-- Ideology of humanism / gender equality / cross-culture awareness / globalization embedded in textbooks

-- Cognitive demanding in exercise design before and after curriculum reform

-- Basic vocabulary investigation in its frequency, central meaning and typical pattern

-- Features of orality in textbook dialogues

--“3-dimension grammar teaching” and grammar exercise design

-- “Lexical grammar” & lexical teaching design

-- Schema theory, corpus & reading skill training

On theme & stylistic features in literature reading:

-- Color words in Sons and Lovers

-- Beauty of color in Wilde’s fairy tales

-- Relationship between nature and humanity in Walden

-- Shakespeare's view point on the four seasons

On learners’ inter-language:

-- “Small words” in LINDSEI-Chinese corpus

-- Spelling errors / connective devices / attitudinal adverbials in learners English compositions

-- Text structures in abstract writing

On classroom discourse:

-- Negotiation sequence in classroom conversation

-- Questioning / feedback giving/ repairing / code switching / dis-fluency / in teacher talk

These are all corpus-based or corpus-aided papers and most of them have been published in books and journals in China.

Conclusion

The above description indicates a smooth circle from corpus building and research to practical teaching and again going back to corpus expanding and researching further. However, as a new way of teaching and learning language, corpus-using teachers have to “maintain control of a potentially large quantity of evidence while trying out generalizations and this requires intellectual skills that have not traditionally been taught (Sinclair 2003: vii). Such a demand challenges our corpus application. We still have many problems, including: 1) how to guide trainees to make corpus analysis, especially from simply identifying repeated forms to categorizing similar semantic or functional groups; 2) how to select and edit corpus examples to meet students' current proficiency; and 3) how to keep offering relevant corpora to the trainees when they graduate from our school. All this drives us to further study and practice.

References

Anthony, Laurence. 2006 *AntConc 3.2 Ow.bet a3(windows)*, a free software developed in School of Science and Engineering , Wesada University, Japan.

He, Anping. 2007 "Corpus-aided analysis of EFL course books." In *Curriculum. Teaching Material. Methodology*. Beijing: People's Education Press, 44-49.

Leech, G. 1998. "Preface." In *Learner English on Computer,* S. Granger (ed.). London: Longman.

O'Keeffe, MA. & F. Farr. 2003. "Using language corpora in initial teacher education: Pedagogic issues and practical applications." *TESOL Quarterly* 37/3:389-418.

Sinclair, John. 2003. *Reading Concordances*, London: Pearson Educational Limited.

Sinclair, John. 2004. *Trust the Text*, Routledge: London.

A CORPUS-BASED DESCRIPTION OF SUBTITLES FOR THE DEAF AND THE HARD OF HEARING (SDH) IN BRAZIL

Vera Lúcia Santiago Araújo[240]

Élida Gama Chaves[241]

Abstract

Two previous studies on SDH have suggested that the North-American closed caption model available on Brazilian open television is not as effective as it might be. These studies have also demonstrated that condensation and editing are key elements in enabling deaf viewers to enjoy a subtitled programme. A current reception research is trying to investigate which level of editing would increase accessibility for the Brazilian Deaf and Hard-of-Hearing. This research will try to find out a model of SDH that meets the needs of the country's deaf community.

Ordinary subtitles meant for the hearing audience normally have the maximum of 2 lines and their duration ranges from 1 to 6 seconds. In order to assure their readability, the number of characters per second (cps,) or the number of words per minute (wpm) should be about 13, 14 or 15 cps or 145, 160 or 180 wpm, depending on the required amount of editing. The subtitles exhibited in Brazil, henceforth closed captions, however, do not follow these patterns. They are a verbatim translation of the speech and are shown 2 seconds after speech and image. A group of Deaf from Fortaleza is testing the captions in order to create a preliminary model which will be evaluated by other Deaf from other parts of the country.

This article aims at presenting the corpus of the research built with the closed captions and the subtitles produced by the research team. The purpose is to compare the current model and the proposed ones to verify which changes are needed for Brazilian Deaf and the Hard of Hearing to watch TV comfortably. The work is still in progress, and the aspects analysed are lexical density (number of words per minute), segmentation (number of subtitles per speech) level of condensation (percentage of source speech translated), additions (explicitation) and deletions (deleted words and expressions).

Keywords: audiovisual translation, subtitling for the Deaf and Hard-of-Hearing, corpus-based description, closed caption model, reception research.

Introduction

The State University of Ceará, Brazil, has been investigating Subtitling for the Deaf and Hard-of-Hearing for eight years (SDH). Since then, two studies on SDH have been carried out (Franco and Araujo 2003; Araujo 2004, 2005 and 2007). The results suggested that condensation and editing are key elements in enabling deaf viewers to enjoy a subtitled film or programme. A current reception research is trying to investigate which level of editing would increase accessibility for the Brazilian Deaf and Hard-of-Hearing. This research will try to find out a model of SDH that meets the needs of the country's deaf community. A group of Deaf from Fortaleza is testing the subtitles in order to create a preliminary model which will be evaluated by other parts of the country.

[240] Vera Lúcia Santiago Araújo teaches English as a Foreign Language and Translation Studies at the State University of Ceará. Her main interest is on audiovisual translation. She carries out research on closed subtitling for the Deaf and Hard-of-Hearing in Fortaleza, Brazil. She has written many articles on the subject, either alone (Closed subtitling in Brazil. In Topics in audiovisual translation. John Benjamins Publishing Company; A legendagem para surdos no Brasil, EdUECE, 2005; Subtitling for the Deaf and Hard-Of-Hearing in Brazil Rodopi, 2007), or with Eliana Paes Cardoso Franco (Reading Television - Checking Deaf People's Reactions to Closed Subtitling in Fortaleza, Brazil. The Translator 2003). She also works with the use of audiovisual translation in language teaching and audiodescription for the blind.

[241] Élida Gama Chaves studies Translation and English at the State University of Ceará. She has been working with Professor Vera Lucia Santiago Araújo as an undergraduate junior researcher since March 2007. She has got a scholarship from a Brazilian Government Agency, FUNCAP (Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico,) to carry out research on closed subtitling for the Deaf and Hard-of-Hearing (SDH). Her undergraduate final paper will be on the use of corpora to describe SDH.

This article aims at presenting an eletronic corpus of this research built with the subtitles, from now on referred to as closed captions, produced by a Brazilian network (Globo Television) and the subtitles produced by the research team. The purpose is to compare the current model and the proposed ones to verify which changes are needed for the Brazilian Deaf and Hard-of-Hearing to watch audiovisual productions comfortably. The corpus may also be helpful in training future subtitlers by calling their attention to relevant aspects related to audiovisual acessibility.

SDH in Brazil

Subtitling for deaf and hearing audiences in Brazil differ a lot. For the Deaf, we use the North-American system of closed caption which inserts the title in line 21 of the vertical blanking interval – the black horizontal bar between individual television images – in the video signal. The titles – white characters in a dark background – are only visible by means of a decoder operated by the remote control of the TV set.

The model presents two types of captions. Roll-up captions scroll continuously from the bottom to the top of the screen, reaching up four lines at a time. Words come out from the left to the right side of the screen, and it is the type of caption used in programmes requiring real time translation, such as talk shows and news. Pop-on captions are similar to the open subtitles usually seen in Brazilian cinema and television. It is often used in fiction programmes: films, telenovelas (Brazilian famous soap operas), and sitcoms. Unlike roll-up captions, they come on and off the screen synchronised with speech and image.

Closed captions are produced by a professional called stenocaptioner, who operates a stenograph (a computerised piece of equipment with 24 keys that can be pressed simultaneously), a machine equipped with a stenotype keyboard (a keyboard similar to the one used in court), that allows fast typing. The stenocaptioner must be a skillful typist, because about 160 words a minute have to be typed. As s/he is trained to write verbatim transcriptions of the speech, the speech-image-subtitle synchronism is not an important feature. Roll-up captions are often displayed two seconds after the speech.

As one can see, having translational competence is not a required skill to work with SDH in the country, because this activity is not seen as translation. People ignore the fact that Translation Studies recognize three types of translation: interlingual (between two different languages), intralingual (within the same language) and intersemiotic (between different semiotic media, from verbal to visual and vice-versa). The result is that the model is ineffective and does not meet the needs of the Brazilian deaf community. (Franco and Araújo 2003; Araújo 2004, 2005 and 2007)

Ordinary subtitles meant for the hearing audience, on the other hand, have been succesfully used by Brazilian viwers for a long time. They normally have the maximum of 2 lines and its duration ranges from 1 to 4 seconds (in Europe they can reach 6 seconds). In order to assure their readability, the number of characters per second (cps) or the number of words per minute (wpm) should be about 13, 14 or 15 cps or 145, 160 or 180 wpm, depending on the required amount of editing. The closed captions, however, do not follow these patterns. They are a verbatim translation of what is being said and are shown 2 seconds after speech and image.

In order to show the differences between the two translations, table 1 presents the subtitles for a soap opera called Páginas da Vida (Life Pages):

Globo closed captions	Proposed subtitles
[LUCIANO] VOCÊ ME DISSE QUE NÃO TINHA AULA HOJE (1) (You told me you had no class today)	**[Luciano] Você não tinha aula?** (Didn't you have class?)
[GISELE] É, EU NÃO TENHO MESMO. MAS É QUE MEUS PAIS QUEREM QUE EU VÁ PRA ESCOLA... (2) (I really didn't. The thing is, my parents want me to go to school...)	**[Gisele] Não.** (No.)
MESMO QUANDO NÃO TENHO AULA. SABE, PRA FAZER ALGUNS ESPORTES, ALGUMAS ATIVIDADES SABE? (3) (....even when there's no class, you know. They say I can play some sports or do some other things there)	**Eu vou à escola mesmo sem aula.** (But I go to school anyway.) **Para fazer esporte.** (To play some sports.) **Outras atividades.** (Other activities)
SABE O QUE QUE É? EU NÃO SEI SE SEUS PAIS SÃO ASSIM MAS... (4) (You know what? I don't know if your parents are like this, but ..)	**Sabe o que é?** (You know what ?)
MAS OS MEUS ACHAM QUE TUDO DEVE SER FEITO EM FUNÇÃO DA ESCOLA E COM O PESSOAL DA ESCOLA. (5) (...mine think that everything I do must be related to school and school people.)	**Meus pais acham …** (My parents think...) **que tudo é em função da escola.** (that everything is related to school)

Table 1: Subtitling of Life Pages.

There are three main differences between ordinary subtitles and closed captions. First, Globo captions are a verbatim translation of the dialogues, as they lacked condensation. All of our translations were edited. Some were reduced (1, 2, and 4), and others were divided into two (5) or three (3), so that they harmonized with image and increased their readability. Second, Globo captions were written with capital letters. Normally, this type of letter is only used to translate written information within the film, like names of buildings, signs, headlines, titles of books, and so on. We used yellow subtitles with transparent background (the black background on table 1 is only to enhance their visibility), because the research Deaf participants regarded them as being more visible on the screen. Third, the maximum number of lines in the TV network is different, as the TV titles presented three lines in most of the examples showed in table 1 (2, 3, 5,).

The Corpus

The Making of

The corpora is composed of the transcription of the speech of some sequences of Globo TV programmes, and the subtitles produced to translate them - the ones by Globo and by the research team (see Table 1). Different genres of Globo programming were selected: Soap Opera - *Páginas da Vida* (Life Pages); Comedy - *A Grande Família* (The Big Familly); Movie - *O Auto da Compadecida* (The Dog's Will); Talk Show – *Programa do Jo* (Jo's Show); Documentary - *Globo Reporter*; and Variety Show - *Fantástico* (Fantastic).

The corpus is being built through the use of wordsmith tools. Its preparation follows three steps: transcription, tagging and aligning. The transcription of the speeches and the subtitles is being made in txt-format. The codification of the corpus will be done by tagging the following elements of SDH: 1) Identification of speaker; 2) Information on sound effects; 3) Translation of the speech by means of the different subtitlings. Every speaker must be identified for the Deaf to distinguish which one is speaking. In the proposed model it comes in brackets, not in the tagging, and every sound effect must be shown. As the Deaf is deprived of the auditory channel, everything that produces sound, and is relevant to follow the programme or film, must be translated. A transcription of the speech and its related translations will be provided.

The speakers will be tagged as <FAL1> **name of speaker 1** </FAL1> to speaker one and <FAL2> **name of speaker 2** </FAL2> to speaker two and so on:

<L6> <FAL2>GISELE</FAL2> MEUS PAIS ACHAM QUE. . <L6>
TUDO É EM FUNÇÃO DA ESCOLA.

The tagging of speaker identification.

The sound effects will receive an <ES> tag. To the first sound effect, <ES1>; to the second sound effect <ES2>:

<ES1> [Música Tema: "Wave" Bossa Nova] Sound Track Wave Bossa Nova
<ES2> [Toque de Piano] Piano Playing
<ES3> [Tiros] Gun Shots

<ES4> [Música Rock]

The tagging of sound effects.

Every speech and subtitle will receive an <L> tag: the first subtitle, <L1>, the second subtitle, <L2> and so on. For example, table 2 brings a dialogue between a boy and a girl in Life Pages:

Speech	Globo subtitles	Team subtitles
Por exemplo, aos sábados o pessoal lá sai junto, combinam de fazer alguma coisa. (For example they meet on Saturdays to go out together, and decide what to do .)	POR EXEMPLO, AOS SÁBADOS O PESSOAL LÁ SAI JUNTO, COMBINAM DE FAZER ALGUMA COISA.	O pessoal sai aos sábados, (People go out on Saturdays,)
Um vai pra casa do outro, comer uma pizza, sai pra dançar, mas eu sinceramente não gosto. (Go to a friend's house have some pizza or go out dancing. I honestly don't like it.)	UM VAI PRA CASA DO OUTRO, COMER UMA PIZZA, SAI PRA DANÇAR, MAS EU SINCERAMENTE NÃO GOSTO.	vai para casa do outro, (Go to each other's house,) come pizza, sai pra dançar. (have some pizza, go out dancing.) Eu não gosto. (I don't like it.)

Table 2: a dialogue from Life Pages.

This dialogue will be tagged, and the source text, the speech, and its translations, the captions and subtitles, will be aligned as follows:

<L7> POR EXEMPLO, AOS SÁBADOS O PESSOAL LÁ SAI JUNTO, COMBINAM DE FAZER ALGUMA COISA.
<L7> POR EXEMPLO AOS SÁBADOS O PESSOAL LÁ SAI JUNTO, COMBINAM DE FAZER ALGUMA COISA.
<L7> O PESSOAL SAI AOS SÁBADOS.
<L8> UM VAI PRA CASA DO OUTRO, COMER UMA PIZZA, SAI PRA DANÇAR. MAS EU SINCERAMENTE NÃO GOSTO.
<L8> UM VAI PRA CASA DO OUTRO, COMER UMA PIZZA, SAI PRA DANÇAR. MAS EU SINCERAMENTE NÃO GOSTO.

<L8> VAI PARA CASA DO OUTRO, <L8> COME PIZZA, SAI PRA DANÇAR. <L8> EU NÃO GOSTO.

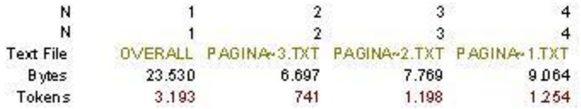
Tagging and aligning a dialogue from Life Pages.
The aligning will allow three types of comparison: source text and Globo; source text and team subtitles; captions and team subtitles.

The Analysis

The aspects to be analysed in the corpora are lexical density, level of condensation, segmentation, additions and deletions. Lexical density (LD) is related to readability. In order for the viwer to have time to read the subtitles and harmonize them with the images, they have to undergo a lot of editing. When exposed to a verbatim transcription of what is being said, the viewer may not watch the audiovisual production comfortably. The lexical density will be described by using the wordlist's statistics that will provide the number of words (tokens) which appear in each transcription. This figure will be divided by the duration in minutes of the programme or the film sequence transcribed.

$$\text{LD} = \frac{\text{number of words}}{\text{duration in minutes}}$$

For instance, the LD of Life Pages files are: 1) Speech = 180wpm (1198÷6.63); 2) *Globo* captions = 189wpm (1254÷6.63); 3) Team subtitles = 111.76wpm (741÷6.63). In spite of being a verbatim translation of the speech, Globo captions are denser because of the additional information (identification of speaker and sound effect).



N	1	2	3	4
N	1	2	3	4
Text File	OVERALL	PAGINA~3.TXT	PAGINA~2.TXT	PAGINA~1.TXT
Bytes	23.530	6.697	7.769	9.064
Tokens	3.193	741	1.198	1.254

Wordlist statistics for Life Pages.

The level of condensation (LC) deals with the percentage of source speech translated. When the corpus is finished, we will be able to know the amount of reduction necessary for a Brazilian Deaf to follow a subtitled film or programme. This amount will be obtained using the number of words given by the wordlist to the source and target texts. These figures will be compared so that we can have the percentage which is calculated by the following equation:

$$\text{LC} = \frac{\text{number of words of target text} \times 100}{\text{number of words of sourse text}}$$

The LC for the soap opera is: 1) *Globo* captions translated 104.67% of the speech (1254×100÷1198); 2) team subtitles translated 61.85% of the speech (741×100÷1198).

Segmentation refers to the number of subtitles that will be derived by each source text. The more subtitles per speech, the better it is. This procedure makes sure that the viewer will not have to process too much information and, as a consequence, will have more time to enjoy the show. The alignment of the dialogue from Life Pages shown above brings the difference between our segmentation and Globo's. One can see that in <L8> above, we produced three subtitles while *Globo* preferred to use just one.

The wordsmith's alignment resource will also provide information on what was added and deleted. Additions come in the form of the speaker's identification and the translation of sound effects. Deletions are necessary to achieve the desired level of condensation. Considering it is very relevant to know what kind information was taken out, tables 1 and 2 bring some examples of these aditions and deletions. In <L7> redundant ideas, such as the fact that people go out together and talk to each other to decide what to do before going out, were eliminated because they could be inferred by the context. As a result, our subtitles are more condensed and easier to read.

Final remarks

This article presented the use of electronic corpora to describe and analyse subtitles. As the SDH corpus is not finished, we do not have the complete picture of what kind of subtitles have produced so far. However, what we know for sure is that having quantitative data generated by the corpus will help us to train novice translators so that they can deal with accessibility in their future practice. The data may also be useful in convincing audiovisual producers that their products should be improved if they are to meet the needs of the Brazilian deaf community.

References

Araújo, V. L. S. Closed subtitling in Brazil *In Topics in audiovisual translation*. Amsterdam: John Benjamins Publishing Company, 2004, 199-212.

Araújo, V. L. S. A legendagem para surdos no Brasil. *In Questões de Lingüística Aplicada: Miscelânea.* Fortaleza:EdUECE, 2005, 163-188.

Araújo, V. L. S. Subtitling for the Deaf and Hard-of-Hearing in Brazil In *Media for All: Subtitling for the Deaf, Audio Description and Sign Language*, Kenilworth, Nova Jersey, EUA: Rodopi, 2007, 99-107.

Franco, E. P. C.; **Araújo, V. L. S.** Reading Television - Checking Deaf People's Reactions to Closed Subtitling in Fortaleza, Brazil. *The Translator*, Manchester, 2003, 9/2, 249 – 267.

COMPUTER-AIDED ERROR ANALYSIS AND STUDENTS' LEARNING DISORDERS: THE CASE OF SPELLING

María Belén Díez-Bedmar[242]

Abstract

Although error-tagging a computer learner corpus (CLC) is no doubt a difficult time- and effort-consuming task, the results obtained from Computer-aided Error Analysis (CEA) have demonstrated the value of such analyses in the improvement of the descriptions of the learners' language, their pedagogical materials, curricula, etc.

However, when analyzing CEA results, an aspect of vital importance is not normally taken into account, namely the students' learning disorders. This limitation has a direct consequence: the results of CEAs may be skewed, since errors may be attributed to the second language acquisition process rather than to a learning disorder. Hence, learner corpora can be claimed to be even more valuable resources, since they may help in the detection of students' learning disorders.

To exemplify so, this poster presents a computer learner corpus-based study of the possible dysorthographic spelling errors that a diagnosed student with dyslexia made during her four-year degree on English Studies (Filología Inglesa) at the University of Jaén (Spain). The data used for this poster is the subsection of the four-year longitudinal learner corpus compiled and error-tagged with the UCLEE at the University of Jaén (total amount of words, 283,623) which was handwritten by the diagnosed student along her degree (i.e. 25,096 words).

Keywords: computer-aided error analysis, interlanguage, learning disorders, spelling, dysorthography

Introduction

Computer learner corpora (CLC) have been used to describe the written or oral production by foreign or second language learners' so that experts in second language acquisition and foreign or second language teachers may obtain an in-depth description of the students' interlanguage or their process(es) of language acquisition. As a result, it is possible to improve their curriculum, the teaching materials implemented, the methodology used, etc.

The methodology used to undertake such research has frequently been, among others, Error Analysis (EA) or, more recently, Computer-aided Error Analysis (CEA), so that the learners' errors can be detected and classified either in a cross-sectional or longitudinal design. Although CEA involves the hard task of transcribing and error-tagging a (longitudinal) learner corpus (cf. Meunier 1998: 19-37; Granger 2002: 16-17; Prat Zagrebelski 2004: 93; etc.), the effort made is valuable because the annotator/researcher becomes familiar with the errors of the participants in the corpus,[243] even before running the statistical analyses. Once highlighted, most of those errors are normally explained as part of the student's second or foreign language learning process.

However, other factors, such as learning disorders, should be considered when analysing the student's learning process. For this reason, learners' profile forms would be more complete if this piece of information was required, and researchers were able to consider this variable when analysing the data. Nevertheless, learning disorders are normally unnoticed and underdiagnosed (both in the L1 and the L2 or FL), so students are not aware of their learning disorders and cannot report them. Consequently, the transcriber and/or human error-tagger of a (longitudinal) learner corpus may play a decisive role when detecting them, because the experience of transcribing and error-tagging a learner corpus is of help when determining which errors are frequent in the students' written production at a certain interlanguage stage and which ones are only found in the production by a particular student.

That was the case with the spelling errors by one student in the learner corpus used, who did not only show the common spelling problems that her classmates made, but also characteristic ones which followed repeated

[242] María Belén Díez-Bedmar is Assistant Professor at the Department of English Philology at the Unversity of Jaen (Spain). Her main research interest is the compilation, error-tagging, and pedagogical exploitation of (longitudinal) learner corpora. In fact, she compiled and error-tagged in full with the UCLEE her four-year longitudinal computer learner corpus (compiled at the University of Jaén from the academic year 2002-2003 to 2005-2006 and amounting to 283,623 words), which comprises the academic and general varieties of English. She is also involved in the International Corpus of Cross-Linguistic Interlanguage (ICCI), and participates in projects dealing with Computer-aided Error Analysis and Computer-Mediated Communication (CMC), as well as the implementation of the European Credit Transfer System. She is author and co-author of various chapters and articles on these topics and co-editor of the book 'Linking up Contrastive and Learner Corpus Research' (Rodopi, in press).

[243] The group of students in a learner corpus is supposed to be at the same interlanguage level, as it is normally determined by their institutional status (cf. Granger 2004: 130).

patterns. In fact, when qualitatively analysing those specific errors in an interdisciplinary team, involving two members of the Department of Psychology at the University of Jaén and myself, our conclusion was that they could be cases of dysorthographia. As a result, the student was informed of such possibility and eventually diagnosed with dyslexia by a psychologist.

Since dyslexia is related to both reading and writing disorders, and it was impossible to have access to the diagnosed student's reading skills, this paper focuses on the diagnosed student's spelling errors which are possibly due to dysorthographia,244 defined as

> […] a specific disorder of spelling which accompanies dyslexia; the cognitive dysfunction underlying the two disorders is probably common to both. In dysorthographia, the spelling of words is highly deficient, a direct consequence of the phonological disorder in dyslexic children. (Inserm Collective Expert Review 2007: 20).

Methodology

For the purposes of this study, the 4-year longitudinal computer learner corpus compiled at the University of Jaén from the academic year 2002-2003 to 2005-2006 was divided into two parts. The part used in this paper consists of the handwritten production (amounting to 25,096 words) by the student of the degree English Studies (Filología Inglesa) who has been psychologically diagnosed with dysorthographia based on DSM-IV-TR criteria (2000). The other part, which was handwritten by the other students taking the degree during those years, was not used for this poster.

The complete computer learner corpus, amounting to 283,623 words, was compiled, transcribed and error-tagged in full with the Error Editor developed at the Centre for English Corpus Linguistics (Hutchinson 1996) and its accompanying Error Tagging Manual (Dagneaux, Denness, Granger and Meunier 1996), with the help of a native speaker. Accordingly, spelling errors were annotated with the error tag "Form – Spelling" (FS) and the Concordance Tool in WordSmith Tools (Scott 2003) was used to retrieve all the instances of misspelling to undertake the qualitative studies.

Results

The first step to analyse the diagnosed student's possible instances of dysorthographic spelling was to retrieve the concordance lines with her spelling mistakes. Then, each case of misspelling underwent a qualitative analysis to check if her spelling mistakes showed the characteristics or "symptoms" of dysorthographia, as identified by the interdisciplinary team.

As a result, the thirty-seven types in the grey cells in Table 1 were identified.245 Although on most occasions each type only occurs once, there are twelve types which present various tokens, as indicated in the number between brackets. Hence, concieve, critize, fordward, posite and Romants present two tokens, ect three, believe and nowdays four, litle and thought five, and, finally, masculine presents six. The type which presented the highest number of tokens (22) was poety.

[244] The spelling errors of this student are referred to as "possible dysorthographic spelling errors" for two reasons. First, the student was diagnosed after finishing her degree, that is, after providing all her samples for the longitudinal CLC. Therefore, it ca nnot be scientifically claimed that those errors are caused by dysorthographia in case she developed it later. Second, the spelling errors described in this paper are found in the students' foreign language, whereas she was diagnosed in her L1.

[245] The target tokens are presented below for easier reference.

Critize (2)	Declarate	Demosnstate	Elicitist	Ect (3)	Exampe	Expansionsim
criticise	declarative	demonstrate	elitist	etc	example	expansionism
Exploteted	Feates	Flowe	Fordward (2)	Intest	Intonatio	Litle (5)
exploded	features	flower	forward	interest	intonation	Little
Masculin (6)	Mixting	Nowdays (4)	Peson	Phology	Pictres	Poety (22)
masculine	mixing	nowadays	person	phonology	pictures	poetry
Posite (2)	Recived	Romants (2)	Secon	Semantic	Speakin	Thougth (5)
positive	received	romantics	second	semantics	speaking	thought
Undesrtant	Worllds					
understand	words					

Table 1. Possible dysorthographic spelling errors

When analysing the words in Table 1, it can be noticed that they are frequent words in any student's repertoire. Apart from words clearly related to the field of Linguistics, which should be mastered by any student taking the degree in English Studies, i.e. "declarative", "phonology" or "semantics", etc., most words do not belong to any specialized field. Thus, a comparative study of the frequency of the thirty-seven types when correctly spelt and when showing the spelling above was conducted in the diagnosed student's section of the computer learner corpus to see if they could be cases of slips of the pen or they would rather stem from dysorthographia.

As seen in Figure 1, the instances of possible dysorthographic spelling of fourteen words outnumber the cases correctly spelt. In fact, if percentages are considered, the words which are always incorrectly spelt, possibly due to dysortography, are comparision, concieve, critize, exploteted, fordward, masculin, mixting, pictres and recieve, then followed by ect (60%), converstio, demosnstate, nowdays, and phology (50% each).

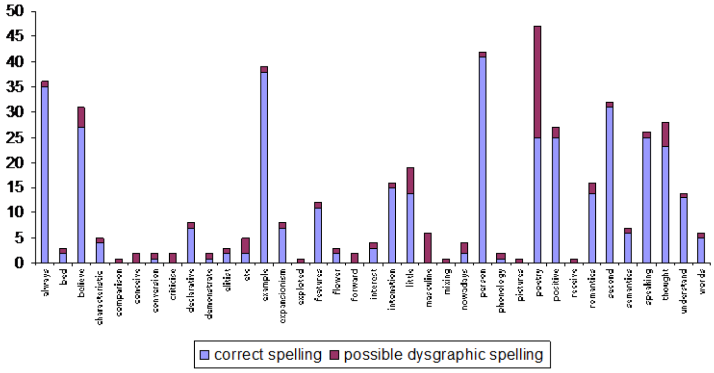


Figure 1. Number of cases presenting possible dysorthographic spellings per word

These results may point out to the relevance of those fourteen words, specially the ones which were always misspelt, i.e. comparision, concieve, critize, exploteted, fordward, masculin, mixting, pictres and recieve, when considering dysorthographic problems. However, the other twenty-three words which were recognized as possibly having spelling problems because of dysorthographia should also be taken into account, as they can follow a specific misspelling pattern.

Conclusion

The use of CEA to study the students' interlanguage cross-sectionally or longitudinally proves a helpful methodology to improve the description of the students' use of the second or foreign language, which may inform learner-corpus-based teaching materials. However, there are several aspects which need to be considered if good CEA results are to be obtained. Among them, a complete learners' profile form, i.e. one with as much information as possible, including the student's possible learning disorders is called for.

As seen in this paper, the omission of such information, which is frequently unknown, biases the interpretation of the CEAs conducted, since all the errors found in a learner corpus are misleadingly attributed to the students' second or foreign language acquisition process, rather than to their (possibly until then) underdetected and underdiagnosed learning disorders. Therefore, the task of the human error-tagger proves a possible first step to the detection and diagnosis of students' learning disorders, and, thus, a new approach to clinical linguistics.

No doubt, further research is needed. In fact, it would be interesting to classify the possible dysorthographic errors found in this student's production. Thus, it would be possible to highlight the most common type(s) of dysorthographic spelling errors so that human error-taggers may recognise them when doing their laborious task. Another line of research would be the comparison of the subsection of the corpus used here, i.e. that by the diagnosed student, with that produced by the other students in the corpus, i.e. the control corpus consisting of the handwritten production by the other 15 students in the four-year longitudinal corpus (258,527 words), to analyse if the possible dysorthographic errors are peculiar to the diagnosed student or they are also found (and which is their frequency) in the control corpus.

Acknowledgements

Thanks are due to Dr. Gómez Ariza and Dr. Gutiérrez Palma at the Department of Psychology in the University of Jaén for their time, help and encouragement.

References

American Psychiatric Association. 2000[4] (1952). *Diagnostic and Statistical Manual of Mental Disorders, Text Revision*. Washington, DC: APA.

Aston, G. 2001. "Learning with corpora: an overview." In *Learning with Corpora*, G. Aston (ed.). Houston: Athelstan, 7-45.

Dagneaux, E., Dennes, S. and Granger, S. 1996. *Error Tagging Mannual Version 1.1.* Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Granger, S. 2002. "A bird's-eye view of learner corpus research." In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung and S. Petch-Tyson (eds.). Amsterdam and Philadelphia: John Benjamins, 3-36.

Granger, S. 2004. "Computer Corpus Research: Current Status and Future Prospects." In *Applied Corpus Linguistics. A Multidimensional Perspective*, U. Connor and T. A. Upton (eds). Amsterdam and New York: Rodopi, 123-145.

Hutchinson, J. 1996. *UCL Error Editor*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Inserm Collective Expert Review. 2007. "*Dyslexia, Disorthography, Dyscalculia. Review of Scientific Data*". http://ist.inserm.fr/basisrapports/dyslexie/dyslexie-synthese-anglais.pdf [Access date 13/02/2008].

Meunier, F. 1998. "Computer tools for the analysis of learner corpora." In *Learner English on Computer*, S. Granger (ed.). London and New York: Addison Wesley Longman, 19-37.

Prat Zagrebelsky, M. T. 2004. *Computer Learner Corpora. Theoretical issues and empirical case studies of Italian advanced EFL learners' interlanguage*. Alessandria: Edizioni dell'Orso.

Scott, M. 2003. *Wordsmith Tools*. Oxford: Oxford University Press.

THE USAGE OF SUBCORPORA IN HIGH SCHOOL: INVESTIGATING SHORT-TERM LINGUISTIC CHANGES

Nina Dobrushina[246]

Abstract

The use of corpora in education may contribute to building a creative research atmosphere. One of the main objectives of a theoretical course of Russian for college students is to show language as a living system which is open to variation and change. A fresh look at Russian grammar and lexicon helps to encourage students' interest in the language and poses challenges inspiring them to start their own scholarly research. The paper discusses one specific way of using RNC: studies of short-term change in morphology (for students in linguistics) and quick tasks in semantic shifts. The procedure of studying the history of adjectives ***russkij*** *'Russian' (ethnic attribute),* ***rossijskij*** *'Russian' (derived from the name of country), and* ***otečestvennyj*** *'national', 'home-made' is described step by step. These three lexemes were widely used through 19th and 20th centuries, but their relative frequency and meaning was changing. The study of semantic changes in the Corpus helps to discover a new and powerful instrument for analyzing social changes. Skills in studying semantic change will be useful not only for linguists, but also for sociologists, journalists, political scientists and historians.*

Keywords: short-term linguistic change, variation, Russian National Corpus, L1 teaching, Russian language

Russian National Corpus in teaching

Russian National Corpus (RNC at www.ruscopora.ru) is a large collection of Russian texts of different time and genre provided with bibliographic, lexical, morphological and semantic markup. The Russian National Corpus comprises 52 392 texts containing 149 357 020 tokens and is a powerful tool of linguistic investigations. Detailed annotation allows making complex lexical, grammatical, semantic queries as well as working with various types of subcorpora.

RNC was launched three years ago as a free and open resource. Linguists have been vastly used the Corpus during the last three years, while most teachers of Russian are unaware of the Corpus and the opportunities it provides. In general, corpus-based approach to language teaching is new to the Russian educational system at all levels (see Dobrushina 2007).

Studies in variation and change

One of the main objectives of a theoretical course of Russian grammar and vocabulary for Russian college students is to show language as a living system open to variation and change. Such a fresh look at Russian grammar and vocabulary fosters students' interest in the language and poses challenges inspiring them to start their own scientific research.

Language corpus is probably the best source for this kind of inspiration, since variation and change are directly available here. The idea that the language may change and vary is new for most students, as the school grammar is based on the concept of language as a system of rules, norms and standards.

There are several ways to approach linguistic variation. One is to consider the studied element synchronically, using statistics to evaluate the relative frequency of variants and trying to understand the underlying linguistic and social mechanisms of change, attempting to model the trend of the change. Another is to study variation diachronically, consider changes within a certain period.

Investigating long-term changes. i.e. classical diachronical / comparative studies, requires a deep background knowledge of historical grammar. Short-term changes (one to two centuries) involve language states that are close to the students' own language. Another, and more practical, motivation for working with nineteenth-century

[246] Nina Dobrushina has been professor of Department of Language and Literature of State University Higher School of Economics (Moscow) since September 2003. She is lecturing a theoretical course of contemporary Russian language to journalists, sociolinguistics to sociologists, language of politics to students of political studies. She was also teaching Russian language to the students of School of Russian Studies from the USA and Germany (State University Higher School of Economics). She holds BA in philology from the Moscow State University (1991) and PhD degree from the Russian State University of Humanities. She has publications in the fields of Russian linguistics, teaching Russian Language, typology of verbal categories, and sociolinguis tics.

Russian is that RNC includes over 23 mln tokens from the period between 1800 and 1900, and over 123 mln tokens from between 1900 and 2008, which suffices for most scientific purposes, within the same interface. A linguist working on Russian is thus in a favorable position as compared to specialists in English studies. To work on both 19th and 20th century language, they are forced to use different corpora (Kytö & alt. 2006).

RNC has a rich system of bibliographic annotation of the texts. The user can easily specify a subcorpus that suits to his purposes basing on such attributes of text as the circumstances of creation, author, genre or function type, topic, etc. With RNC, students may create subcorpora of different periods and follow the change of the investigated phenomenon in a series of stages.

To sum up, the volume of the corpus, the efficient system of information retrieval make the study of short-term changes a possible component of the everyday teaching and learning practices for the Russian studies.

Topics of research

Depending on the students' interest and the topic of the course, the teacher can choose a specific linguistic phenomena for a study of change in grammar or semantics.

Grammatical changes

The students who are specialized in linguistics may be focused on grammatical change. Morphological processes are easy to study, since RNC allows to query many morphological features, including gender, animacy, transitivity, tense, case, degree of comparison and many other.

An example of studying short-term changes in grammar is provided by animacy shifts. Russian nouns have different inflectional pattern in the accusative depending on the animacy: animate nouns (names of humans and animals) normally have the accusative plural identical to genitive plural, while the accusative plural of inanimate nouns is identical to nominative plurals:

Nominative pl.	*starik-i* 'old man'	*venik-i* 'brooms'
Genitive pl.	*starik-ov*	*venik-ov*
Accusative pl.	*starik-ov*	*venik-i*

The intriguing point, however, is that some nouns form both animate and inanimate accusative plural. The distribution of the two forms is unclear. As an example, RNC provides contexts with animate and inanimate interpretation of the word *mikrob* 'microbe'. The word has the following set of case options:

Nominative pl.	*mikrob-y* 'microbe'
Genitive pl.	*mikrob-ov*
Accusative pl.	*mikrob-y* / *mikrob-ov*

It is interesting to find out whether this (or similar) variation is an innovation and which of the two variants of the accusative plural was more common in the 19th century, and to follow the evolution of this variation through the two centuries.

Semantic changes

The study of grammatical changes is primarily interesting for undergraduate linguists. Another type of tasks may attract students of other specializations, including journalism, history, political and social studies. The study of semantic changes in the Corpus helps students to discover a new and powerful instrument for analyzing social changes. Linguists have already started to use corpora in order to understand how the change in the society may be reflected in the language (cf. Bäcklund 2006:17-55). It is now time to teach specialists in other humanities to use corpus approach in their researches. Below I suggest a number of tasks of different level of difficulty which may help students to understand how to trace the changes in lexical meaning.

Simple tasks

On of the easiest and most fascinating ways to play with the Corpus is finding out when certain exactly specific words came into use. As an example, the Corpus provides 1188 of examples of the word *spiker* ‘speaker in the Parliament’. Automatically sorting the matches by creation date, the student will notice that, according to RNC, the word *spiker*, which is, obviously, an English loan, was only occasionally used before 1995 referring to the British parliament, while its Russian life begins after 1995. It takes not more than one minute to answer this question. The teacher should however prevent his or her students against trying to jump to conclusions about the exact date or text where the word is used for the first time. Indeed, the Corpus does not cover the whole diversity of the language (for instance, oral discourse is only partially covered).

Another task will take about half an hour: the student can investigate meaning shifts over time. For example, the well-known Russian word *perestrojka* was quite common before the late 80ies in the sense of ‘reconstruction’. To find the first occurrence of the new, political meaning, the student will have to look through some 3,300 matches. It does not take much time, though, because sorting the matches by date of creation allows to skip all irrelevant documents.

Short researches

At a more advanced level, the teacher suggests students to make a small research on the origins and development of certain meaning. Easy objects are designations of some notions of public social life, since changes in social life are accompanied by drastic changes in meaning. Some examples are *revolucija* ‘revolution’, *narodnyj* ‘folk, national’, *tovariš*^{j} ‘comrade’. As an instance of practical direction of the student’s activity, I will describe what the teacher does step by step.

Students are suggested to consider the change in the meanings of the adjectives *russkij* ‘Russian’ (ethnic attribute), *rossijskij* ‘Russian’ (derived from the name of country), and *otečestvennyj* ‘national’, ‘home-made’. These three lexemes were widely used through 19th and 20th centuries, but their relative frequency and meaning was changing (see an analysis of these words in Abrosimova, Kuzmina 2007: 140-141).

The teacher can start the study during the classes by suggesting his or her students to consider several examples which s/he found in the subcorpus of the early 19th century. They are then asked to explain in what ways this usage of the word is different from its current usage, to show that meanings of words are changing and to help students to develop some basic skills of understanding and tracing slight meaning shifts. For that, 5 to 10 contexts need to be chosen where some of the uses are similar to the contemporary usage, while some others are different.

(1) *No čto skazali by podpisčiki žurnala v nastojaš*^{j}*ee vrem*^{j}*a, jesli b uznali, čto on truditsa ne dl*^{j}*a nix, a dl*^{j}*a tex iz ix potomkov, kotorye vzdumajut kogda-nibud*^{j} *zanimat*^{j}*s*^{j}*a istorijej* ***otečestvennoj*** *literatury?* (V.N. Majkov)

‘But what would the subscribers to the magazine say now, if they knew that does not works for them but for those of their descendents who would one day make up their mind to study the history of the **Russian** literature’

(2) *Za r*^{j}*umkoju vina, v osobennosti* ***otečestvennogo****, nemec oživl*^{i}*aetsa, molodeet, rasskazyvaet, i, kak dit*^{i}*a tešitsa igruškoj, on tešitsa svojej starinoj*. (V.A. Sollogub)

‘Drinking a glass of wine, especially **homeland’s one**, a German becomes animated, young, starts telling stories and, as a child plays with a toy, enjoys his (his country’s) ancient history’

While the example (1) shows the usage of the word which is quite expectable in contemporary language (*otečestvennaja literatura),* the example (2) shows that in early 19th century the adjective *otečestvennyj* was used with reference to any country. Today its use is restricted to Russia, thus becoming a loose synonym of the adjective *russkij* (‘Russian’).

Next, students are asked to work with RNC themselves. If the teacher has access to a computer class, students may carry out a number of tasks under her/his supervision. As the first experience of working with RNC, the teacher asks them to find examples of one of these three words in the whole corpus, to sort the examples by the date of creation, and to note a number of different constructions (e.g. nominal heads) with this adjective which are most typical for specific periods. Assisted by the teacher, students compile the following table:

	1800-1830	1980 - 2008
otečestvennyj	*nravy* ‘customs, morals’, *istorija* ‘history’, *literature* ‘literature’, *prosveš^{j}enie* ‘education’, *pisateli* ‘writers’, *talanty* ‘talents’, *nebo* ‘sky’	*banki* ‘banks’, *predprijatija* ‘enterprises’, *metallurgija* ‘metallurgy’, *tovary* ‘wares’, *biznes* ‘business’, *kompanii* ‘companies’
rossijskij	*flot* ‘marine’, *or^{j}ol* ‘eagle’ (the symbol of the Russian Empire), *gosudarstvo* ‘state’, *goroda* ‘towns’, *poddannye* ‘nationals’, *zakony* ‘laws’, *posol* ‘ambassador’, *zeml^{j}a* ‘earth’	*zavody* ‘factories’, *ekonomika* ‘economy’, *biznes* ‘business’, *gosudarstvo* ‘state’, *neft^j* ‘oil’, *futbolisty* ‘football players’, *armija* ‘army’, *šaxmaty* ‘chess’
russkii	*baby* ‘country women’, *pesni* ‘songs’, *zeml^{j}a* ‘earth’, *jazyk* ‘language’, *ban^{j}a* ‘bath’, *odežda* ‘clothes’, *čelovek* ‘person’	*nauka* ‘science’, *krasavica* ‘beauty’, *tradicii* ‘traditions’, *turisty* ‘tourists’, *mentalitet* ‘mentality, intellectual attitudes and perception’, *klassika* ‘classical literature’, *skazki* ‘fairytales’, *jazyk* ‘language’

After a discussion of these preliminary results, the students are asked to build a subcorpus of 1900 - 1930 and to start looking for examples where the same adjectives are used in a way which is not common in contemporary Russian. For instance, example (3) shows how a Soviet writer uses the word *rossijskij* as synonym to *drevnij* ‘very old’:

(3) *Avtobus, peredelannyj iz gruzovika, s veselym ryčaniem obognal jego i zataraxtel po ulice mimo žalkix okrainnyx domišek. Čto-to drevnee,* ***rossiskoe****, bylo v etix lačugax, v zapaxe skotnogo dvora, čto isxodil ot nix.* (B. Guber)

‘The bus, converted from a lorry, left it behind with a cheerful roar and rattled on past wretched houses of the suburb. There was something ancient, **pre-revolutionary** in these cabins, in the smell of the farm-yards that came from them.’

In example (4), *rossijskij* is opposed to *sovetskij*: *rossijskij* means ‘one that existed before the revolution’ and is apparently used with negative connotation, while *sovetskij* means ‘contemporary’ and has a positive one.

(4) *Odnako* ***sovetskie*** *diplomaty okazalis^j sil^{j}nee i tverže mnogoopytnyx* ***rossijskix*** *poslov.* (B.Savinkov)

‘But the Soviet diplomats were stronger and firmer than the highly experienced ambassadors of the Russian Empire.’

Students will soon realize that, in the mid 20th century, *rossijskij* was used almost exclusively with reference to the pre-revolutionary Russia, since the country was re-named *Sovetskij Sojuz* (‘Soviet Union’). In the course of this little study, students learn how to use RNC to build subcorpora.

On later stages of research, the teacher helps students to obtain simple skills of statistical qualification of the data. For example, a simple count of relative frequencies of adjectives *otečestvennyj, russkij,* and *rossijskij* gives non-trivial results. The adjective *rossijskij* covers about 30% of the whole group of synonyms at the beginning of the 19th and in the end of the 20th century, while in the mid 20th century its usage drops down to 3%. This is a likely point to start a discussion of the connections between the changes in the semantics and social and historical processes in the country.

Conclusions

Skills in studying semantic change are useful not only for linguists, but also for sociologists, journalists, political scientists and historians. Analysis of short-term changes in the Corpus is an efficient way to develop creative approaches to language learning and to inspire students to start their own projects. The Corpus models a real scholarly research which may be carried out in a relatively short time because the students do not need to spend weeks to collect data. Working with the Corpus may help students gradually develop true research skills, for which the study of short-term changes in grammar or semantics is an appropriate training field.

References

Abrosimova E.A., Kuzmina N.A. 2007. Izučenie aktivnyx processov semantičeskoj derivacii s pomoš^{i}ju NKRJa. In: Dobrushina N.R., ed. Nacional^{i}nyj korpus russkogo jazyka i problemy gumanitarnogo obrazovanija. Sbornik statej. Moskva, Teis. Pp. 136-149

Bäcklund, Ingegerd. 2006. Modifiers describing woman and man in Nineteenth-century English. In: Kytö, Merja, Mats Rydén & Erik Smitterberg, eds. Nineteenth-Century English. Stability and Change. Cambridge University Press.

Dobrushina N.R., ed. 2007. Nacional^{i}nyj korpus russkogo jazyka i problemy gumanitarnogo obrazovanija. Sbornik statej. Moskva, Teis.

Kytö, Merja, Mats Rydén & Erik Smitterberg, eds. 2006. Nineteenth-Century English. Stability and Change. Cambridge University Press.

Leech, Geoffrey. 2004. Recent grammatical change in English: data, description, theory. Advances in Corpus Linguistics. Karin Aijmer and Bengt Altenberg (Eds.), Rodopi, Amsterdam.

FOCUSING ON LEARNING OUTCOMES: USING CORPORA AT A UNIVERSITY OF TECHNOLOGY

Andreas Eriksson[247]

Abstract

The present paper describes the objectives behind a work-in-progress project on the implementation of corpus material in language courses at Chalmers university of technology, Göteborg, Sweden. Chalmers offers a challenging environment with students who are not language students primarily. The courses given by the Centre for language and communication are both elective courses and obligatory courses on a wide range of engineering programmes. The project covers three courses with quite different objectives and involves both written and spoken language. A central argument in the paper is that the implementation of corpora for teaching purposes in ESP environments can be facilitated if it is anchored in identified learning outcomes. The learning outcomes of the three courses and how corpora might contribute to the fulfilment of these outcomes are discussed in the paper.

Keywords: corpora, learning, ESP, learning outcomes

Introduction

The value of corpora for language teaching purposes as well as the limitations of corpora for such purposes are issues that have been discussed and commented on by many scholars (e.g. Chambers 2007:6-7, Gabrielatos 2005:25, Johansson 2007:26, Lee and Swales 2006:57, O'Keefe, McCarthy and Carter 2007:246-247). One point that has been made explicitly by several authors is that corpora are no magic wands that inevitably generate language learning (see e.g. Conrad 2000:548, Gaskell and Cobb 2004:315, Mauranen 2004:103). There is thus generally great awareness of the problems and limits of corpora for teaching purposes among teachers and researchers. Still, since one of the major strengths of corpora is that they lend themselves to providing students with new and innovative types of language input, there is always a risk that emphasis is put on instructional input rather than learning outcomes. There is obviously no inherent contradiction between improved input and learning outcomes, but without basing the use of specific material in particular learning outcomes, the value of the material may not be as strong as it could have been.

The aim of the present work-in-progress report is to account for the early stages of a project where particular learning outcomes have been identified in three courses and to describe in what way corpora are believed to help students reach these outcomes. The idea is thus to base the use of corpora in learning outcomes and make the use of corpora outcome driven.

The study is carried out at the Centre for language and communication (CLC) at Chalmers university of technology and involves three different courses and EFL/ESP students from several engineering programmes. CLC has long experience of providing English proficiency courses, academic and technical writing courses as well as of integrating communication practice into engineering education (see e.g. Börjeson et al. 2007, Carlsson and Wranne 2008, Evertsson et al. 2007). However, the centre has comparatively little experience of using corpora or corpus-related material in their teaching, and the implementation and adaptation of such material into various courses thus partly means breaking new ground. However, it should be emphasised that the decision to use corpora stems from the identification of particular learning outcomes. The three courses described below are courses where certain learning outcomes were identified and where it is hypothesized that corpora could enhance learning.

Since students at Chalmers are students of engineering, language is not their major subject at university. It is likely that many of these students differ from language students in terms of motivation, objectives and familiarity with terminology. Consequently, the students that participate in the present investigation are different from the students in most other studies which have dealt with the use of corpora for teaching purposes. Students in such

[247] Andreas Eriksson obtained his PhD of linguistics at Göteborg university in 2004 with the thesis Tense and Aspect in Learner Writing. Advanced Swedish learners' use of tense and aspect in English argumentative text. He has been working as a senior lecturer at the Centre for language and communication at Chalmers university of technology since 2007. His research interests are second language acquisition (SLA), genre studies, corpus linguistics, tense, and aspect.

investigations have typically been university language students and the studies have been carried out as action research projects at departments of language at universities (Chambers 2007:7-8, Mauranen 2004:90-91).

Description of courses and desired learning outcomes

The first course to be investigated is an academic writing course open to all doctoral students at Chalmers. The course has been given for several years and covers the writing of texts common in academic writing, for example abstracts, research articles and conference papers. Overall the course has been very successful but one learning outcome that could be strengthened is students' knowledge about writing in their own disciplines, i.e. what is often referred to as genre knowledge. Genre knowledge is diverse in nature and involves several features. Hyland (2004:84), for instance, recognizes eight major aspects of genre knowledge. One of these is knowledge of grammar and phraseology. Corpora should be good sources for enhancing this particular aspect of genre knowledge, as they can give information about how and where particular words and phrases are used. The strengthening of this aspect would have to involve active participation from students, and therefore the collection of both individual and discipline specific corpora, in line with the study carried out by Lee and Swales (2006), are seen as potentially useful activities. As a result, the work will have to include the use of text analysis software, such as WordSmith Tools in order to facilitate comparisons of for instance collocations and clusters (Scott and Tribble 2006).

The second course is an elective proficiency course primarily aimed at students at basic or lower intermediate level. It involves both spoken and written proficiency but there is a focus on spoken proficiency as the course is supposed to help international Master's students during their first year at Chalmers. Master's courses at Chalmers are often project-based and thus involve a great deal of both formal and informal spoken English (all Master's courses are taught in English). The fairly general learning outcome of the spoken part of the course is to prepare students for future studies at Chalmers, particularly in terms of improving their ability to participate in discussions and give oral presentations. The material that corpora can provide, and which has not been used at Chalmers previously, consists of patterns common in spoken English. The effect of teaching such patterns, often referred to as formulaic sequences (Wray 2000, Meunier and Gouverneur 2004), is not fully clear but the arguments that are usually put forward are that the use of formulaic sequences can free processing capacity and that they can be beneficial for students in handling speech events (Ellis 1996, Wray 2000, Wray and Perkins 2000, Mauranen 2004). The corpus material in the present course will most likely be corpus-based and taken from sources like O'Keefe, McCarthy and Carter (2007), who list a number of formulaic sequences (referred to as 'chunks' by O'Keefe, McCarthy and Carter 2007:65-67) common in spoken language. The advantage of using this type of material is that it consists of common and naturally occurring sequences of spoken language. In other words, students are shown examples of language as it is actually spoken. The learners' use of chunks in both formal and informal university settings can be compared with material from a corpus like the Michigan Corpus of Academic Spoken English (MICASE) (Simpson et al. 2002).

The third course is called *Safety communication* and is given at the programme of nautical science. In the course, students learn to use a set of phrases and terminology published by the International Maritime Organization (IMO) as the IMO Standard Marine Communication Phrases (IMO 2002).[248] The phrases constitute one aspect of Maritime English and have been developed in order to make communication at sea as simple as possible so that also seamen with limited knowledge of English can communicate effectively in commonly occurring situations at sea and in harbour. The number of phrases is limited and all of them have been compiled and published in one single document. As a consequence, it might at first seem fairly clear what the students need to learn. Not surprisingly, however, it has turned out that knowing the phrases is seldom enough in real life situations. In these situations, the user often needs to apply language structures and vocabulary that extend far beyond the limits of the IMO phrases. One particular aspect of the need for a general knowledge of English concerns the cotext surrounding the fixed IMO phrases. Minor investigations have shown that the cotext of IMO phrases often contains formulaic sequences, but that both students and teachers are unaware of these sequences. The hypothesis is therefore that knowing more about the cotext of the IMO phrases would help students use these phrases more correctly and more effectively. It seems as if not only the phrases but also the words co-occurring with these phrases are fairly fixed, and being able to clarify the relationships with phrases and their cotext is believed to help students in their use of the IMO phrases. This part of the project is more extensive than the first two parts described, since it involves the identification of formulaic sequences as well as decisions about how the use of sequences can be learnt.

[248] Examples of IMO phrases are: "You are proceeding at dangerous speed" and "Stand by for assistance" (IMO 2002).

Critical factors

All three courses come with a number of critical factors that may hinder learning and which therefore have to be addressed. *Time* is one such factor. Both the doctoral student course and the safety communication course are intensive and run over seven weeks only. Several studies have emphasised the need for extensive training periods if students are to use corpora independently (e.g. Gaskell and Cobb 2004, Johansson 2007:25). Considering such findings, it is likely that the three courses in the present project will be corpus-based or corpus-informed rather than corpus-driven.

Another critical factor is students' level of English, particularly in the proficiency course. There is some doubt as to whether corpus material is useful for low-level learners or not, particularly when it comes to learner consultation of corpora (Chambers 2007; see also O'Keefe, McCarthy and Carter 2007:24). This is thus yet another factor which supports the use of selected corpus material rather than data driven learning in order to reach the learning outcome identified.

The third critical factor is *comparability* and primarily concerns the doctoral student course. If the students are to collect a mini-corpus of their own work and compare that with other material from their field, it is important that the material is comparable. It may not be possible to monitor this process fully and it is therefore a factor that needs to be considered when evaluating the learning outcome.

Evaluating the learning outcome is in itself perhaps the most difficult part of the whole process and a well-known problem in research on learning (Barr and Tagg 1995). The use of spoken language is particularly difficult to evaluate for a number of reasons. First of all, it involves recording, which in itself may affect participants. Secondly, a great deal of material is needed to evaluate the use of particular phraseology, and thirdly, transcribing recorded material is tedious and time-consuming work.

Another difficult aspect of the effects of corpus material and corpus-related teaching is to determine what generates learning. However, even if it is not always possible to show that a particular type of methodology results in a particular type of learning, it might be possible to show that a course where corpora have been used has resulted in particular learning outcomes. This might be a reasonable first step to indicate that corpora facilitate learning. The results can then be used to change certain parameters in order to be able to gain more knowledge about learning effects (cf. Gaskell and Cobb 2004:315-317).

Conclusion

Meunier and Gouverneur (2007:132) aptly show how general learning processes can be linked to exercises for the learning of formulaic sequences. This type of linking and the emphasis on learning outcomes made in the present paper may not be revolutionary approaches to corpora, but grounding the use of corpora in learning outcomes and general learning processes may strengthen teachers' and researchers' awareness of how corpora can be employed for teaching purposes.

Another insight from more general approaches to learning that seem to be worth remembering when dealing with corpora is Gardener's (1993:24) and Biggs'(2003:46) emphasis on the risks of trying to cover too much. According them, coverage is one of the greatest enemies of deep learning, as it prevents students from analysing the material carefully enough. This is certainly one of the risks of corpora as students can quickly be presented with a great deal of unanalysed material. Another aspect of this problem is addressed by Bowden and Marton (1998:24), who argue that "[f]or each phenomenon there is a limited number of critical aspects that can be discerned and focused on simultaneously. So differences in how this phenomenon is experienced reflect differences in what critical aspects are discerned and focused on simultaneously". Guiding inexperienced corpus users to focusing on critical aspects and to noticing use in various situations seems to be essential if corpus material is to foster learning in many ESP environments. Corpora in ESP contexts often require a great deal of contextualisation, but if such contextualisation can be based in explicitly stated learning outcomes, corpora can be used for a variety of purposes and in many different ESP contexts.

References

Barr, R. and **Tagg, J.** 1995. "From teaching to learning — a new paradigm for undergraduate education." *Change* 27/6: 12–25.

Biggs, J. 2003. *Teaching for Quality Learning at University*. London: The Society for Research into Higher Education & Open University Press.

Börjeson, F., Eriksson, A-M., Erlandsson, J., Hillman, K., Molander, S. and **Rex, E.** 2007. "Ger koppling av kunskapsinlärning och färdighetsövning ökad djupinlärning? – Utveckling av kursmoment i "Miljö- och resursanalys för hållbar utveckling V2"." In *ESA rapport 2007:12,* Göteborg: Chalmers University of Technology.

Bowden, J. and **Marton, F.** 1998. *The university of learning. Beyond quality and competence in higher education.* London: Kogan Page.

Carlsson, C-J. and Wranne, O. 2008. "Using writing as a cognitive tool in a car design project". In *Proceedings of the 4th International CDIO Conference, Hoogeschool Gent, Gent Belgium, June 16-19, 2008*. Gent, Belgium

Chambers, A. 2007. "Popularising corpus consultation by language learners and teachers." In Hidalgo, E, Quereda and Santana, J. *Corpora in the foreign language classroom. Selected papers from the sixth international conference on teaching and language corpora (TALC 6). University of Granada, Spain, 4-7 July, 2004.* Amsterdam: Rodopi, 3-16.

Conrad, S. 2000. "Will corpus linguistics revolutionize grammar teaching in the 21st century? " *TESOL Quarterly* 34/3: 548-560.

Ellis, R. 1996. "Sequencing in SLA. Phonological memory, chunking, and points of order." *Studies in Second Language Acquisition* 18:91-126.

Evertsson, M., Bankel, J., Enelund, M., Eriksson, A., Lindstedt, P. and **Räisänen, C.** 2007. "Design-Implement Experience from the 2nd Year Capstone Course "Integrated Design and Manufacturing"." In *Proceedings, Third International CDIO Conference, June 11-14, 2007*. MIT, USA.

Gabrielatos, Costas. 2005. "Corpora and language teaching: Just a fling or wedding bells?" *Teaching English as a Second or Foreign Language (TESL-EJ)* 8/4:A-1

Gardener, H.W. 1993. "Educating for understanding". *The American School Board Journal*, 180/7: 20-24.

Gaskell, D. and **Cobb, T.** 2004. "Can learners use concordance feedback for writing errors?" *System* 32: 301-319.

Hyland, Ken. 2004. *Genre and second language writing*. Ann Arbor: University of Michigan Press.

IMO. 2002. *IMO Standard Marine Communication Phrases*. London: International Maritime Organization.

Johansson, S. 2007. "Using corpora: from learning to research." In Hidalgo, E, Quereda and Santana, J. *Corpora in the foreign language classroom. Selected papers from the sixth international conference on teaching and language corpora (TALC 6). University of Granada, Spain, 4-7 July, 2004.* Amsterdam: Rodopi, 17-28.

Lee, D. and **Swales, J. M.** 2006. "A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora." *English for Specific Purposes*, 25: 56-75.

Mauranen, A. 2004. "Spoken corpus for an ordinary learner." In Sinclair, J.M. *How to use corpora in second language teaching.* Philadelphia, PA: John Benjamins, 89-105.

Meunier, F. and **Gouverneur, F.** 2007. "The treatment of phraseology in ELT textbooks." In Hidalgo, E, Quereda and Santana, J. *Corpora in the foreign language classroom. Selected papers from the sixth international conference on teaching and language corpora (TALC 6). University of Granada, Spain, 4-7 July, 2004.* Amsterdam: Rodopi, 3-16.

O'Keefe, A., McCarthy, M. and **Carter, R.** 2007. *From corpus to classroom*. Cambridge: Cambridge University Press.

Simpson, R.C., Briggs, S. L., Ovens, J. and **Swales, J. M.** 2002. *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.

Scott, M. and **Tribble, C.** 2006. *Textual patterns. Key words and corpus analysis in language education.* Amsterdam: John Benjamins.

Wray, A. 2000. "Formulaic sequences in second language teaching: principle and practice." *Applied Linguistics* 21/4: 463-489.

Wray, A. and **Perkins, M. R.** 2000. "The functions of formulaic language: an integrated model". *Language and communication* 20/1: 1-28.

PORTUGUESE AS A FOREIGN LANGUAGE: LANGUAGE TEACHING FOR SPECIFIC PURPOSES BASED ON CORPORA

Telma de Lurdes São Bento Ferreira[249]

Luciene Novais Mazza[250]

Abstract

This study focuses on the development of teaching materials for language specific purposes supported by the research areas of Corpus Linguistics and English for Specific Purposes. Our main objective is to elaborate a unit for foreigner students of Portuguese Language within business context in Brazil. For this analysis, we have made use of a study corpus of authentic texts composed of pharmaceutical registers and a reference corpus from Banco de Português do Brasil (CEPRIL). By the computational software WordSmith Tools (Scott, 1997) we could identify appropriated linguistics features to be applied in preparation of exercises. The study shows a different application of language learning-teaching based on Data-Drive Learning (Johns 1994) compared to the traditional methodology that has been found in teaching materials, especially, for those developed for business in Brazil.

Keywords: Portuguese as a Foreign Language, LSP, needs analysis, corpus-based teaching, materials development.

Introduction

The increasing use of *corpora* has been noticed in researches for teaching material elaboration, especially for English Language Teaching. However, these researches show that the use of *corpora* in teaching Portuguese as a Foreign Language (PFL) has not been effectively exploited. Although, a reasonable number of researches in Portuguese as a Foreign Language (Paes Almeida 2007: 49-50) can be found, the use of electronic *corpora* has been little used in elaboration of teaching materials. Amongst researches found, we highlight Cavalcante (2006) who use Corpus Linguistics basis to analyze verbal forms of a textbook to teach Portuguese for foreigners. We also highlight a research carried out by Dell'sola (Júdice et al. 2002), who, besides discussing how resources available on the Internet may be used as useful information source to learn Portuguese, also mentions a CD-ROM developed for teaching Portuguese containing interviews with native speakers. However, although the author uses a *corpus*, data collected for material creation does not seem to be previously prepared according to the basis of Corpus Linguistics.

Another concerning found in researches on evaluation and preparation of teaching materials is about language for specific purposes, mainly materials designed for learners involved in business areas. During our research, only one (01) material available in Brazil to teach Portuguese for business was found. In that material, both texts and exercises were not contextualized to the unit topics and, in addition, such texts did not seem authentic; maybe they had been adapted for publishing. Another important point noticed in this teaching material was a lack of exercises to exploit linguistics characteristics to understand the texts, like vocabulary and grammar. Probably, this gap is due to reading skill emphasis only. Therefore, we know that for commercial issues, editors must meet a more comprehensive public, making impossible a more specific work.

Thus, the general goal of this study is to create an activity to be used to teach Portuguese as a Foreign Language for intermediate and advanced levels, supporting future works to be developed. Consequently, this study should also help the needs of language teachers in Brazil, mainly, those who teach Portuguese for Specific Purposes.

[249] Telma de Lurdes São Bento Ferreira is a candidate for the degree of Master of Arts in Applied Linguistics at Catholic University of Sao Paulo and she is a specialist in Translation in English/Portuguese at University of Sao Paulo. She has been working on teaching and translation for ten years and is one of the owners of Lexikos Cursos e Traduções. Her research interest areas ar e Portuguese as a Foreign Language, Materials Development and Corpus Linguistics.

[250] Luciene Novais Mazza is a candidate for the degree of Master of Arts in Applied Linguistics at Catholic University of Sao Paulo. She has a lato sensu degree in Translation and Portuguese Language and works as a technical translator for multinational comp anies in Brazil. Her research interest areas are English for Specific Purposes and Materials Development.

Considering as a starting point: (i) lacks found in materials published in Brazil and (ii) use of learning method suitable to Portuguese as a Foreign Language for Specific Purposes, we decided to use a pharmaceutical specialized *corpus* and adopt the Data-Driven Learning (DDL) approach to analyze data. For this, we used theoretical basis of Language for Specific Purposes and Corpus Linguistics.

Fifteen (15) samples of registers about pharmaceutical segment were collected, with almost 6,500 words (tokens) processed by WordSmithTools 3.0 application (Scott, 1997). Also theoretical support by authors Hutchinson and Waters (1987), Robinson (1991), Johns (2004), Dudley-Evans and St.John (1998), Berber Sardinha (2004), Ramos (2008), and others was used.

Literature review

A Language for Specific Purposes (LSP) Approach

According to Dudley-Evans and St.John (1998: 2), *LSP has focused on the teaching of languages such as French and German for specific purposes, as well as English. In many situations the approaches used are very similar to those used in ESP; some, however, place a much greater emphasis on the learning of vocabulary*. But in this paper, we tried not only to analyze vocabulary, but also concordances and frequency patterns amongst different registers, corroborating basic hypothesis expressed by Allen and Widdowson (1974, *apud* Hutchinson and Waters, 1987: 10) when referring to English Language, that states: [...] *the difficulties which the students encounter arise not so much from a defective knowledge of the system of English, but from an unfamiliarity with English use, and that consequently their needs cannot be met by a course which simply provides further practice in the composition of sentences, but only by one which develops a knowledge of how sentences are used in the performance of different communicative acts.* Nowadays, we can understand those different communications (oral and written) as different registers amongst different contexts of situation and production.

In ESP approach, we have some authors like Robinson (1991) who says that *ESP is normally goal directed*, that is, students study the language not because they are interested in the language or its culture, but because they need the language for academics or professional purposes. In this way, Hutchinson and Waters (1987: 12) say that the purpose of a course for specific purposes is to provide the students *to function adequately in a target situation.* In Brazil, Ramos (2008) states that would be desirable for an ESP teaching a pedagogical proposal using strategies based in specific genres. Following such statements, we should also associate genres or registers use with the building of specialized *corpora* making use of methodologies on the basis of Corpus Linguistics.

Corpus Linguistics (CL)

According to Berber Sardinha (2004: 296), since the 70s corpus-based language description has been presented an ongoing growth in language teaching and learning, where *there are several applications from Corpus Linguistics for teaching.* According to the author, *corpora* use in teaching is focused in four areas: (1) native language description; (2) learner's language description; (3) methodology migration from research to the class; and (4) development of teaching materials, approaches and syllabus. The last item - interesting of this study, uses Corpus Linguistic concepts by three branches – the Lexical Syllabus, the Lexical Approach and the Data Drive Learning (DDL).

Data-Driven Learning (DDL)

According to Johns (1994: 296), data-drive learning (DDL) is an innovative approach in two aspects (technological and methodological). In the nineties, with technology approach and software and computing tools appearance, capable to store a huge number of data, it was possible to automate *corpora* entry via scanners and OCR (optical character recognition), making easy its development, as Johns (1994: 296) says that with the concordancing programs we are able *to recover rapidly from a corpus all the contexts within which a particular linguistic element – word, morpheme or phase – occurs.*

Johns (1994: 297) also states that main advantages which makes DDL different from the other approaches are: Direct access to data so that the learner can take part in building up his or her own profiles of meaning and uses; a form of linguistic research, and that the concordances offers a unique resource for the stimulation of inductive learning strategies of perceiving similarities and differences. The author says that these methodology implications should be considered when preparing teaching materials emphasizing real language and using concordance-based exercises, so that students can develop inductive strategies helping learning outside the class.

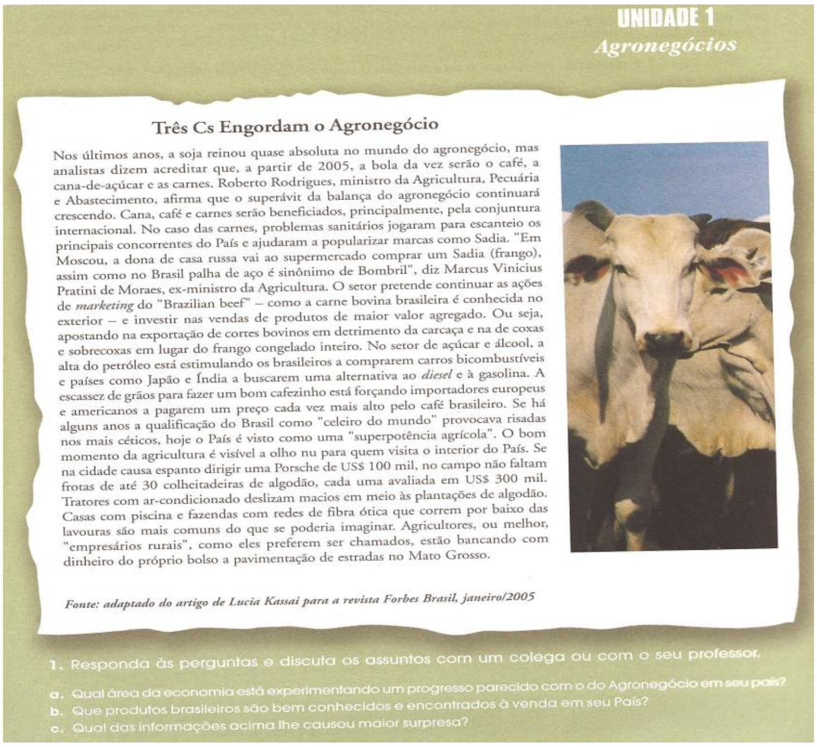
So, if materials for Portuguese as Foreign Language for Specific Purposes adopted methodologies like DDL when elaborating teaching exercises with specialized corpora, teachers should provide their students a better content, offering linguistic variations in formal and informal situations, attending then one of the lacks noticed in this research related to Portuguese learners in Brazil.

Background to the Portuguese Language for Foreigners in Brazil

From a brief survey applied to teachers of Portuguese as a Foreign Language in Brazil, two (02) relevant pieces of information about needs and wants of foreign students were chosen as analysis criteria. The first is about students´ wants in learning language in use, i.e., the Brazilian routine, they want to understand what Brazilian say; and the second one is students´ needs in reading more fluently in Portuguese, both in business and private situations. Then, we assumed that teaching materials for foreign language teaching are weak and teachers most of the time need to prepare activities without any support of the textbook, as there is no materials available on the market. So, we are supposed to say that, teachers knowing other methodologies can be more productive and materials that can helpful in the classroom.

On the other hand, students need materials based on authentic texts of language in use, their real needs mainly to know ‘why’ and ‘what for’ they need to know such language.

Unfortunately, no find teaching material concerned about such needs nor this material provide a support in using the language in several situations. In specific areas, such as the pharmaceutical, the lack is much deeper. As the teacher can not be a specialist in every area according to the student’s professional area, negotiation should be established with students in order to develop a suitable syllabus, finding out provides supports for teaching of language for specific purposes was found. Particularly, materials for Portuguese as a Foreign Language available for business area are rare and incomplete. As mentioned in introduction above, in Brazil one (01) teaching textbook for business was found, only for intermediate and advanced students. In this textbook, we noticed that the content does not properly approach the texts presented in the Units. For a better example of this reference, we present below a figure describing an exercise of the first Unit of the Brazilian textbook entitled “Panorama Brasil: Ensino do Português do Mundo dos Negócios”.



UNIDADE 1
Agronegócios

Três Cs Engordam o Agronegócio

Nos últimos anos, a soja reinou quase absoluta no mundo do agronegócio, mas analistas dizem acreditar que, a partir de 2005, a bola da vez serão o café, a cana-de-açúcar e as carnes. Roberto Rodrigues, ministro da Agricultura, Pecuária e Abastecimento, afirma que o superávit da balança do agronegócio continuará crescendo. Cana, café e carnes serão beneficiados, principalmente, pela conjuntura internacional. No caso das carnes, problemas sanitários jogaram para escanteio os principais concorrentes do País e ajudaram a popularizar marcas como Sadia. "Em Moscou, a dona de casa russa vai ao supermercado comprar um Sadia (frango), assim como no Brasil palha de aço é sinônimo de Bombril", diz Marcus Vinicius Pratini de Moraes, ex-ministro da Agricultura. O setor pretende continuar as ações de *marketing* do "Brazilian beef" – como a carne bovina brasileira é conhecida no exterior – e investir nas vendas de produtos de maior valor agregado. Ou seja, apostando na exportação de cortes bovinos em detrimento da carcaça e na de coxas e sobrecoxas em lugar do frango congelado inteiro. No setor de açúcar e álcool, a alta do petróleo está estimulando os brasileiros a comprarem carros bicombustíveis e países como Japão e Índia a buscarem uma alternativa ao *diesel* e à gasolina. A escassez de grãos para fazer um bom cafezinho está forçando importadores europeus e americanos a pagarem um preço cada vez mais alto pelo café brasileiro. Se há alguns anos a qualificação do Brasil como "celeiro do mundo" provocava risadas nos mais céticos, hoje o País é visto como uma "superpotência agrícola". O bom momento da agricultura é visível a olho nu para quem visita o interior do País. Se na cidade causa espanto dirigir uma Porsche de US$ 100 mil, no campo não faltam frotas de até 30 colheitadeiras de algodão, cada uma avaliada em US$ 300 mil. Tratores com ar-condicionado deslizam macios em meio às plantações de algodão. Casas com piscina e fazendas com redes de fibra ótica que correm por baixo das lavouras são mais comuns do que se poderia imaginar. Agricultores, ou melhor, "empresários rurais", como eles preferem ser chamados, estão bancando com dinheiro do próprio bolso a pavimentação de estradas no Mato Grosso.

Fonte: adaptado do artigo de Lucia Kassai para a revista Forbes Brasil, janeiro/2005

1. Responda às perguntas e discuta os assuntos com um colega ou com o seu professor.

a. Qual área da economia está experimentando um progresso parecido com o do Agronegócio em seu país?
b. Que produtos brasileiros são bem conhecidos e encontrados à venda em seu País?
c. Qual das informações acima lhe causou maior surpresa?

Exercise of Unit 1, *Panorama Brasil: Ensino do Português do Mundo dos Negócios*

As we can see above, the text in Unit 1 is not exploited based on linguistic elements, but using three questions that students must know according to their knowledge of world. Maybe for such answers students should need other approaches, not necessarily texts. In this situation, probably teacher adopts warm-up strategies for the text before or during activities with students.

From this point of view, we can assume that a corpus-based study considering students needs and wants and meeting teacher’s expectations on teaching materials can provide satisfactory results for language teaching.

Outline of a specialized corpus analysis processed by computer

The pedagogic *corpus* used in this paper was built from journals related to medicines, vaccines and input. They were selected from the Virtual Library of the Brazilian Ministry of Health. The analysis was processed in three steps. In the first step, we collected fifteen (15) registers, almost 6,500 words (tokens), changed into text files and analyzed by using the WordSmithTools 3.0 application (Scott, 1997). In the example below, we present a sample of the wordlist with lemma frequency and percentage.

N	**Word**	**Freq.**	**% Lemmas**
1	DE	2.858	4,73
2	E	1.497	2,48
3	A	1.394	2,31
4	EM	1.011	1,67
5	COM	568	0,94
6	QUE	538	0,89
7	DA	492	0,81
8	DO	422	0,70
9	NÃO	372	0,62
10	SE	325	0,54
11	OS	306	0,51
12	PARA	281	0,46
13	PACIENTES	279	0,46
14	POR	276	0,46
15	IN	275	0,45
16	MAIS	260	0,43
17	NA	256	0,42
18	RISCO	244	0,40
19	NO	237	0,39
20	É	232	0,38

Partial wordlist from Pharmaceuticals Journals *Corpus*

Then, we manually identified the most frequent words, of which we chosen the main words: *pacientes, risco, uso, tratamento, estudos, placebo, medicamentos, prevenção, saúde, eficácia.* Empirically, such words have been foreseen, because the main subject of the journals was the rational use of medicines. In the second step, we identified and processed one of the most frequent words in the corpus (*risco*) and generated a concordance list for it. In the chart below, we present a partial list of concordances of the word *risco.*

1	rado em indivíduos com alto	**risco**	(2 a 10%) de fratura
2	de progesterona determina	**risco**	2 a 3 vezes maior de c
3	se populacional encontrou	**risco**	23 vezes maior de agr
4	árias de TRH apresentaram	**risco**	50% maior de morte
5	lados determinaram maior	**risco**	(65%), seguidos das vi
6	TRH, correspondendo ao	**risco**	absoluto de sete event
7	IC: 0,59-0,93; redução de	**risco**	absoluto = 0,07;IC =
8	corresponde à redução de	**risco**	absoluto de 5 casos a
9	benefícios. O excesso de	**risco**	absoluto foi de 8 caso
10	5%: 0,67-0,89; redução de	**risco**	absoluto: 0,03; IC: 0,01-
11	ez associou-se a discreto	**risco**	adicional de malformaç
12	a, mesmo em crianças sem	**risco**	adicional para convuls
13	se pela estratificação de	**risco,**	além dos níveis de LD
14	primária de pacientes com	**risco**	anual de novos eventos
15	a vançado somente quando o	**risco**	anual do evento cardiov
16	dependência dos fatores de	**risco**	apresentados (ver quad
17	a despeito da evidência de	**risco**	associado, pode ser c
18	e transdérmica (24%). O	**risco**	aumenta com a duraçã
19	coagulopatia, apresentam	**risco**	aumentado de hemorr
20	indivíduos. Indivíduos com	**risco**	aumentado de eventos

Chart of a partial list of corcordances with *risco*

When analyzing the concordance list of the word *risk,* a pattern called our attention because it was different from the pattern *benefício/risco.* We present below the occurrence found.

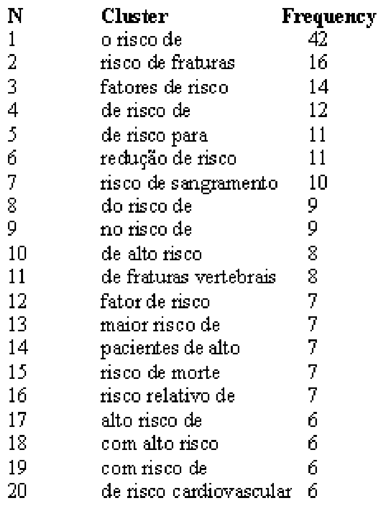
gastrointestinal, nos quais a relação **benefício/risco** pudesse ser maximizada. A prescrição

In order to check this occurrence, we accessed the concordancing program of the Portuguese Database (PUC-SP/LAEL/CEPRIL/DIRECT), that contains 233 million words of Portuguese Language spoken and written in Brazil, and we found two (02) occurrences of the binomial *risco/benefício* and zero (0) of *benefício/risco,* providing the assumption that this occurrence, typically, is not common. In the example below, we present two occurrences found in the CEPRIL Portuguese Database.

"Há muitos produtos melhores e com melhor relação **risco/ benefício"**, afirma.

cercado por não possuírem uma relação favorável de **risco-benefício.** Na Suécia, o registro da

However, at the moment, we will not consider these findings because for this analysis we are interested to identify the most relevant lexical patterns found in the study *corpus.* In the third and last step of the analysis, it was processed clusters of three words, also using the word *risco.* The table below shows a sample of this analysis.



N	Cluster	Frequency
1	o risco de	42
2	risco de fraturas	16
3	fatores de risco	14
4	de risco de	12
5	de risco para	11
6	redução de risco	11
7	risco de sangramento	10
8	do risco de	9
9	no risco de	9
10	de alto risco	8
11	de fraturas vertebrais	8
12	fator de risco	7
13	maior risco de	7
14	pacientes de alto	7
15	risco de morte	7
16	risco relativo de	7
17	alto risco de	6
18	com alto risco	6
19	com risco de	6
20	de risco cardiovascular	6

Chart of list of clusters with *risco*

Rationale

By processing clusters and analysis concordance lines, four relevant patterns were found: (a) adjective + noun + preposition + *risco*; (b) noun + preposition + *risco*; (c) *risco* + preposition + noun; (d) *risco* + adjective, and (e) verb + *risco*, as showed in the sequence of tables below:

(a) adj + noun + preposition + *risco*

Adjective	Noun	Preposition	Research word
alto	*potencial*	*de*	*risco*
baixo			
maior			

Table of clusters of adjective + noun + preposition + *risco*

(b) noun + preposition + *risco*

Noun	Preposition	Research word
redução	*de*	*risco*
Aumento	*do*	
fator	*de*	
pacientes	*de*	

Table of clusters of noun + preposition + *risco*

(c) *risco* + preposition + noun

Research word	Preposition	Noun
risco	*de*	*efeitos*
		eventos
		morte
		recorrência
	para	*fraturas*
		hemorragias

Table of clusters of *risco* + preposition + noun

(d) *risco* + adjective

Research word	Adjective
risco	*relativo*
	potencial
	absoluto
	anual
	adicional

Table of clusters of *risco* + adjective

(e) verb + *risco*

Verb	Research word
haver	*risco*
equilibrar	
reduzir	
acarretar	
apresentar	
encontrar	
ter	
dobrar	

Table of clusters of verb + *risco*

Then in order to find lexical and grammar elements to prepare the exercises, it was processed other concordances list of the word *risco*. Such list had their lines selected for a possible exercise from them. We present below an example of concordances list processed by WordSmith Tools 3 application.

.

1	aixas doses, associa-se a	**risco**	de complicações gastri
2	es ao dia, também acarret	**risco**	cardiovascular, e a F
3	de 2400 pacientes com alto	**risco**	de doença de Alzheime
4	lar em pacientes com alto	**risco**	para o desenvolvimento
5	a para pacientes com alto	**risco**	gastrintestinal.
6	psia em mulheres com alto	**risco**	de hipertensão gestac
7	rado em indivíduos com alto	**risco**	(2 a 10%) de fratura
8	nas em pacientes com alto	**risco**	para fraturas e refratar
9	anejo de pacientes de alto	**risco**	. No entanto, há dado
10	servada a pacientes de alto	**risco**	gastrintestinal, nos qua
11	smo em pacientes de alto	**risco**	, não se justifica. Ev
12	á indicado em casos de alto	**risco**	de fraturas e refratarie
13	úde para pacientes de alto	**risco**	de desenvolver eventos
14	lidade em pacientes de alto	**risco**	. Referências Bibliogra
15	mias em pacientes de alto	**risco**	de desenvolver eventos
16	em 287 pacientes de alto	**risco**	para hemorragia, comp
17	da 32 33 DMO e alto	**risco**	para fraturas . O ris
18	pacientes com artrite e alto	**risco**	de sangramento. Apó
19	nido de osteoporose e alto	**risco**	para fraturas. Em paíse
20	e TRH,correspondendo ao	**risco**	absoluto de sete event

Chart of list of concordances with risco

Exercises details

Using concord and cluster tools from WSTools version 3 (Scott, 1997), we chose and selected some lexical elements to be applied when elaborating the teaching unit proposed in this paper. We also decided to work intermediate and advanced levels as we believe that intermediates learners have a fluency to perform exercises under minimum interference of teacher and they are able to practice their inferential capacity - one of the DDL approach criteria.

In the sequence, we present two examples (02) of exercises elaborated for the unit proposed in this study. It is important to emphasize that all statements have been written in the target language – Portuguese, because we believe it may also help and motivate students to face foreign language learning as a challenging, like that corroborating the student as a researcher.

<u>Exercício 1</u>

(a) Analise os termos abaixo e tente descobrir o assunto do texto a seguir.

1 ANTICONCEPCIONAIS

2 ORAIS

3 MEDICAMENTOS

4 USO

5 EVIDÊNCIAS

6 ADVERSOS

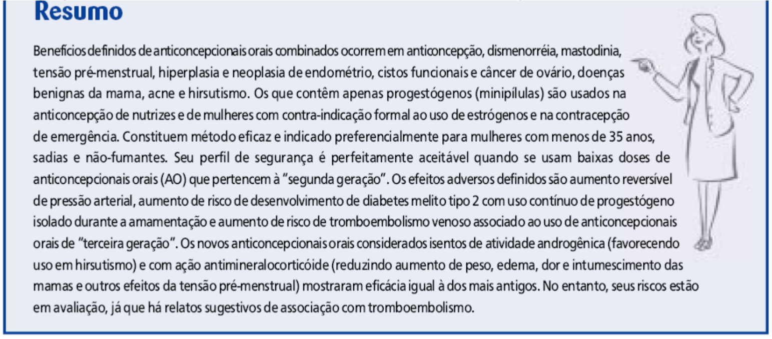
7	BENEFICIOS
8	EFEITOS
9	NOVOS
10	RISCOS

(b) Escreva nas linhas abaixo ao menos três assuntos referente aos termos analisados.

____

____

(c) Agora, leia o texto abaixo e confira se as suas deduções estavam corretas.



Resumo

Benefícios definidos de anticoncepcionais orais combinados ocorrem em anticoncepção, dismenorréia, mastodinia, tensão pré-menstrual, hiperplasia e neoplasia de endométrio, cistos funcionais e câncer de ovário, doenças benignas da mama, acne e hirsutismo. Os que contêm apenas progestógenos (minipílulas) são usados na anticoncepção de nutrizes e de mulheres com contra-indicação formal ao uso de estrógenos e na contracepção de emergência. Constituem método eficaz e indicado preferencialmente para mulheres com menos de 35 anos, sadias e não-fumantes. Seu perfil de segurança é perfeitamente aceitável quando se usam baixas doses de anticoncepcionais orais (AO) que pertencem à "segunda geração". Os efeitos adversos definidos são aumento reversível de pressão arterial, aumento de risco de desenvolvimento de diabetes melito tipo 2 com uso contínuo de progestógeno isolado durante a amamentação e aumento de risco de tromboembolismo venoso associado ao uso de anticoncepcionais orais de "terceira geração". Os novos anticoncepcionais orais considerados isentos de atividade androgênica (favorecendo uso em hirsutismo) e com ação antimineralocorticóide (reduzindo aumento de peso, edema, dor e intumescimento das mamas e outros efeitos da tensão pré-menstrual) mostraram eficácia igual à dos mais antigos. No entanto, seus riscos estão em avaliação, já que há relatos sugestivos de associação com tromboembolismo.

Fonte: Biblioteca Virtual do Ministério da Saúde http://bvsms2.saude.gov.br/php/index.php

(d) De acordo com o texto, quais são os riscos dos novos anticoncepcionais?

____

____

____

Exercício 2

(a) Com base nas linhas de concordância a seguir, identifique as linhas que a palavra risco ocorreu da mesma forma que no texto apresentado no exercício anterior.

1	aixas doses, associa-se a	**risco**	de complicações gastri
2	es ao dia, também acarret	**risco**	cardiovascular, e a F

3	de 2400 pacientes com alto	**risco**	de doença de Alzheime
4	lar em pacientes com alto	**risco**	para o desenvolvimento
5	a para pacientes com alto	**risco**	gastrintestinal.
6	psia em mulheres com alto	**risco**	de hipertensão gestac
7	rado em indivíduos com alto	**risco**	(2 a 10%) de fratura
8	nas em pacientes com alto	**risco**	para fraturas e refratar
9	anejo de pacientes de alto	**risco**	. No entanto, há dado
10	servada a pacientes de alto	**risco**	gastrintestinal, nos qua
11	smo em pacientes de alto	**risco**	, não se justifica. Ev
12	á indicado em casos de alto	**risco**	de fraturas e refratarie
13	úde para pacientes de alto	**risco**	de desenvolver eventos
14	lidade em pacientes de alto	**risco**	. Referências Bibliogra
15	mias em pacientes de alto	**risco**	de desenvolver eventos
16	em 287 pacientes de alto	**risco**	para hemorragia, comp
17	da 32 33 DMO e alto	**risco**	para fraturas . O ris
18	pacientes com artrite e alto	**risco**	de sangramento. Apó
19	nido de osteoporose e alto	**risco**	para fraturas. Em paíse

Examples of exercises elaborated for this study

Considerations

This study has investigated others possibilities to teach a foreign language using an approach based on authentic texts and then discussed some relevant needs and wants pointed by learners and practitioners on the basis of teaching materials available in Brazil. Based on learners involved in the Brazilian business context and their needs for future language use, a preparation of tasks or activities which considers grammatical and lexicon functions by means of concordances of a specialized corpus allow learners to deduce part of the speech used for specific work areas. Besides, learners should also be motivated to research the meaning of that function into different context. Finally, this study forward a proposal for an elaboration of an activity addressed to foreign learners of Portuguese Language acting as a professional in Brazil. Therefore, it is hoped that this study may offer new strategies in materials development and contribute to help practitioner in the classroom.

References

Centro de Pesquisa, Recursos e Informação em Linguagem, CEPRIL. Departamento de Lingüística Aplicada e Estudos da Linguagem, LAEL, PUCSP.

http://www2.lael.pucsp.br/corpora/bp/index.htm [Access date 25/05/2008].

Berber Sardinha, A.P. 2004. *Lingüística de Corpus*. São Paulo: Manole.

Biblioteca Virtual do Ministério da Saúde. Medicamentos, Vacinas e Insumos. http://bvsms2.saude.gov.br/php/index.php [Access date 21/05/2008].

Dudley-Evans, T. and **St John, M. J**. 1998. *Developments in English for Specific Purposes: A multi-disciplinary approach*. Cambridge: Cambridge University Press.

Hutchinson, T. and **Waters, A.** 1987. *English for Specific Purposes: A learning-centred approach*. Cambridge: Cambridge University Press.

Johns, T. 1994. “From printout to handout: Grammar and vocabulary teaching in the context of Data-driven learning”. In **Odlin, T.** (ed.), *Perspectives on Pedagogical Grammar*. Cambridge: Cambridge University Press, 293-313.

Ponce, H. 2006. *Panorama Brasil: Ensino do Português do Mundo dos Negócios*. São Paulo: Editora Galvão.

Ramos, R.C.G. (in preparation). “ESP in Brazil: history new trends and challenges”. In: *Current State of Play vs Actual Needs and Wants*. British Council: IATEFL.

Robinson, P. 1991. *ESP Today: A Practitioner’s Guide.* New York: Prentice Hall.

Scott, M. 1997. *Wordsmith Tools.* Version 3. Oxford: Oxford University Press.

A SPECIAL-PURPOSE CORPUS APPLICATION

Cecília Monteiro Fróis[251]

Abstract

Corpora have been used as a resource for language teaching since the 1990s and several corpora applications to language learning and teaching have been developed. However the use of corpora as sources of terminology is not yet widely accepted as complementary to the use of glossaries.

This paper begins by referring to some published research on applications of corpora to the learning and teaching of languages and translation. It then discusses the usefulness of applying similar methodologies to the teaching of terminology in genetics and to concept acquisition by Portuguese students of science. To achieve the goal of providing such a pedagogical corpus application, a bilingual contemporary and comparable corpus was built using online articles from several sources, and the Corpógrafo (an online suite of tools) was used to store all data and to analyze them. A learner-centred approach and an active methodology were the perspectives on teaching adopted for the realization of this motivation activity. Search words were used for KWIC concordances and the terminological units found were compared with those present in a glossary.

The results show that the compiled corpus is a richer and more up-to-date source of terminology. The explanations retrieved using the Corpógrafo may also be used, in the future, by students to establish semantic relations between the several terminological units in order to actively construct a conceptual basis of basic contemporary genetics.

Keywords: corpus pedagogical application, contemporary special-purpose corpus, genetics, Corpógrafo, motivation activity

Introduction

The main purpose of this paper is to present a classroom activity, based on corpus data, addressed to Portuguese students of science studying texts in English.

The interest in using corpora as a resource for language teaching began in the 1990s when language and linguistic studies became more empirical. (Chambers 2005). Since then, as Chambers relates, several authors have published studies on the applications of corpora in language teaching and learning: Johns (1986), Tribble and Jones (1990), Burnard & McEnery (2000), Kettemann & Marko (2002). These specialized corpora may also be used to teach translation students as Maia (2003: 148, 149) points out.

However, little attention has been paid to the use of corpora applied to the teaching of terminology and concepts to science students. This paper will focus on one possible application of corpus data and corpus analysis to this area. Students in scientific areas have to acquire both the terminology of the area and the conceptual structure referred to by the terminological units. In areas where scientific knowledge evolves quickly, using a corpus could provide access to up-to-date terminology.

This paper begins with a brief description of the corpus built, then it gives a brief overview of the most widely used approaches to corpora and language teaching and finally it describes how the corpus built could be used for students of genetics.

The objective isto show that a contemporary corpus of Genetics may provide more up-to-date terminology and explanations of the most recent terminological units than an online glossary. It also argues that corpus consultation and corpus analysis will enable students of genetics to grasp the conceptual structure of the area more efficiently.

[251] Cecília Monteiro Fróis is a PhD student of Terminology and Translation at the Faculdade de Letras da Universidade do Porto (FLUP). She is working for her thesis on the description of the terminology used in the domain of genetics in English and in Portuguese, under the supervision of Professor Belinda Maia. The Fundação para a Ciência e a Tecnologia (FCT) awarded her a doctoral grant in 2006. She is the co-author of the article "A Case of Meaning Extension" in Jacek Walinski, Krzysztof Kredens and Stanislaw Gozdz-Roszkowski (eds.) *PALC 2005:Corpora and ICT in Language Studies* Frankfurt am Main, Peter Lang Publishing Group, pp. 251-260.

Description of a corpus built for pedagogical applications

The corpus that was constructed is a small bilingual comparable corpus of Genetics, in European Portuguese and in English. The texts included are texts addressed to the general public because, as Vargas Sierra (2006) points out, these texts are richer in definitions and explanations. The span of time between the conclusion of the first draft of the human genome (2000) and the present day was considered contemporary for this corpus. In order to assure the quality of the terminology extracted from the corpus as well as the possible definitions and explanations, great care was taken in the choice of the texts included in the corpus. Texts included are mainly newspaper articles on genetics, science news published in online newspapers, science blogs and material acquired from online academic publishers. The majority of the texts mention their authors and the authors' affiliations and the metadata were registered in our database. However, some texts from reliable sources do not mention the authors.

The authors of the selected texts are mainly science journalists, but some authors are specialists writing for the general public, students or high school teachers. Every effort was made to construct a balanced corpus and, to achieve this goal, texts of a certain size were selected. Texts from several sources and from several authors were chosen in order to have a representative and balanced sample. To obtain a conceptually balanced corpus the texts selected are from the sub-domains of genetics which have become more important in recent years, such as molecular genetics and its areas of application - genetic engineering and forensic genetics - and population genetics.

Pedagogical perspectives

By the end of the 20th century a constructivist perspective on knowledge acquisition had become progressively popular among educators. According to this perspective knowledge is constructed individually as a result of individual experience. There is however, a social dimension which influences both the construction of meaning and its integration in previous acquired knowledge, as well as its influence upon new learning experiences. (Arends, 1995) The Cognitive-Gestalt approach is another learning theory that influenced greatly the learning methodologies. It considers learning as an active process where experience and problem-solving activities are crucial to promote learning. (Hutchinson & Waters 2005, Dunn 2002) These perspectives on learning are a result of psychology research and have implications for teaching methodology.

Hutchinson & Waters (2005) consider that since learning has several facets, using different aspects of different theories is probably the best approach. As a result different methods apply different aspects of the theories. However, the active methods have gained preference since evidence demonstrated that learning becomes more effective when practical activities are introduced in the learning process. (Silva, 1992) However, 'active', in this domain, implies using cognitive capacities, as Hutchinson & Waters (2005) say:

> "In practical terms this means that 'activity' should not be judged in terms of how much learners say or write, but in terms of how much the learners have to think – to use their cognitive capacities and knowledge of the world to make sense of the flow of new information."(Hutchinson & Waters, 2005: 128)

A learner-centred approach to language teaching acknowledges the needs, learning styles and interests of the students and recommends the adoption of a flexible teaching methodology which suits best the interests of the students. (Richards & Rodgers, 2005) The integration of corpus consultation as a language learning activity fits in this learner-centred approach to language teaching and learning about a special domain in English.

A corpus-based LSP teaching application

The LSP bilingual comparable corpus was explored for data using the Corpógrafo. The Corpógrafo, "an integrated web-based tool for corpus linguistics and knowledge engineering", (Sarmento et al. 2004) was used to build and store the corpus and metadata. The Corpógrafo allows the use of raw texts, enables terminological extraction, the creation of terminological databases and the finding of possible definitions occurring in the corpus and also the finding of semantic relations between terms. (Maia & Sarmento 2005).

The corpus built was used to confirm the hypothesis that applying semantic relations analysis of terms would benefits students of science and raise their awareness of the conceptual field of Genetics and, at the same time, show them that some recent terminological units are found in the corpus and not in recent glossaries such as The Talking Glossary of Genetic Terms from the National Human Genome Research Institute.

The intention was to illustrate, for the students, the interest in establishing semantic relations between the concepts referred to by the terms. The explanations found in the corpus would be used to establish the semantic relations between the terminological units. The corpus analysis presented here is a very limited one although we intend to broaden the scope of the present analysis. At this stage this application could be used as a motivation activity that could interest students in the consultation and analysis of corpus-data in the future.

To check the validity of the hypothesis a KWIC concordance search was done using the Corpógrafo. The keyword chosen for the search was the well known term known by its abbreviation “DNA”.

For this paper, the focus was narrowed to the English corpus for the intention was to show the interest of this application to facilitate the student’s task of acquiring terminology in a second language which is English. As most terminological units are compound nouns, a search for collocations of nouns associated with “DNA” seemed a good method to explore new terminological units. After a search of KWIC concordances at sentence level, some interesting collocations were found:

> Researchers have begun to suspect that some of this noncoding **DNA** consists of regulatory sequences that can affect, in different ways, when different genes are "expressed" and actually able to produce specific proteins.
>
> Scientists and venture capitalists are comparing RNA interference, or RNAi, to the recombinant **DNA** revolution that launched the entire biotechnology industry in 1976 when Genentech began its operations.
>
> It turns out that **DNA** generates far more RNA than the standard dogma predicts it should - even some "junk" DNA gets transcribed.
>
> **DNA** vaccines are a simpler version, made of just the virus' genes placed into a circular DNA structure called a plasmid, then grown in bacteria.
>
> In fact, that answer depends on the **DNA** chip that 23andMe uses to scan customer
>
> genomes.
>
> Researchers can use **DNA** sequencing to search for genetic variations and/or mutations that may play a role in the development or progression of a disease.

A search through The Talking Glossary of Genetic Terms from the National Human Genome Research Institute shows, at letter D. the terminological units “deoxyribonucleic acid” (DNA), “DNA replication” and “DNA sequencing”. At letter R we find “recombinant DNA” and at letter N “non-coding DNA”. However, the terminological units “junk DNA”, “DNA vaccines” and “DNA chip” cannot be found in this glossary.

Using as keyword “RNA” the results at sentence level concordances were:

> The accelerating number of discoveries in miRNA can be linked to another recently discovered RNA phenomenon, called RNA interference, or RNAi, which also blocks protein function.
>
> RNA is a chemical relative of DNA found in the nucleus and cytoplasm of cells.
>
> In addition to the messenger- and micro- varieties, it has several other iterations including transfer RNA and ribosomal RNA.
>
> By 2000, researchers figured out how it works: Cells chop RNA into pieces, which are called small interfering RNA.
>
> In 2001, Rockefeller University researcher Thomas Tuschl published a paper explaining how to design small interfering RNA that could inhibit any chosen gene.
>
> Specifically, an enzyme copies the information in a gene's DNA into a molecule called messenger ribonucleic acid **RNA** (mRNA).

The Talking Glossary of Genetic Terms from the National Human Genome Research Institute presents the terms “ribonucleic acid (RNA)” and “messenger RNA (mRNA)” which are also present in the corpus. On the other hand the corpus has “small interfering RNA”, “transfer RNA”, “ribosomal RNA”, “RNA interference” or “RNAi”, and “miRNA”. The variety and relevance of terminological units is greater in the corpus.

The corpus presented here cannot be accessed directly due to copyright, but students would have access to it through a password. Another solution could be the distribution of print-outs, of the data retrieved from the corpus, to the students who could then search for the semantic relations between the terminological units and compare them with those found on The Talking Glossary of Genetics accessible at the following site: http://www.genome.gov/10002096

Concluding remarks

This paper argues that corpora are useful resources for teaching Languages for special purposes to students of genetics. Returning to the hypothesis posed earlier, it is possible to state that the corpus built provides recent terminology which is not found in the NHGRI glossary and it also provides explanations for some terminological units which will be useful for the conceptual organization of the domain.

However, as the corpus is small, and the data presented are few, it would be necessary to enlarge the corpus and the extraction of data in the future, in order to evaluate better the benefits of its use. Nevertheless the results presented here may be used as a motivation activity which could persuade students of the benefits of corpus consultation.

References

Arends, R. (1995). *Aprender a Ensinar*. Lisboa: McGraw-Hill de Portugal Lda.

Chambers, A. (2005). "Integrating corpus consultation in language studies". In *Language Learning & Technology*, Volume 9, Number 2, May 2005, pp.111-125 at:

http://llt.msu.edu/vol9num2/chambers/

Dunn, L. (2002). "Theories of learning". In OCSLD at: http://www.brookes.ac.uk/services/ocsd/2_learntch/briefing_papers/learning_theories.pdf

Hutchinson, T. & Waters, A. (2005). *English for Specific Purposes: a learning-centred approach.* Cambridge: Cambridge University Press. 21st printing.

Maia, B. (2003). "Using Corpora for Terminology Extraction: Pedagogical and Computational Approaches". In Barbara Lewandowska-Tomaszczyk (ed): *PALC 2001:Practical Applications in Language Corpora*, Frankfurt am Main, Peter Lang Publishing Group, pp.147-162

Maia, B. & Sarmento, L. (2005). "The Corpógrafo - an Experiment in Designing a Research and Study Environment for Comparable Corpora Compilation and Terminology Extraction". In *Proceedings of eCoLoRe / MeLLANGE Workshop, Resources and Tools for e-Learning in Translation and Localisation* (Centre for Translation Studies, University of Leeds, UK, 21-23 de Março 2005), pp. 45-48. pdf

Richards, J. C. & Rodgers, T. S. (2005). *Approaches and Methods in Language Teaching.* Cambridge: Cambridge University Press. 2nd edition, 10th printing.

Sarmento L. & Maia, B. & Santos, D. (2004). "The Corpógrafo - a Web-based environment for corpora research". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation (LREC'2004)* (Lisboa, Portugal, 26-28 de Maio de 2004), pp. 449-452.

Silva, M. G. (1992) *Manual de Métodos e Técnicas Pedagógicas*. Lisboa: CNS.

Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (Eds.). (1997). *Teaching and Language Corpora.* London: Longman.

MOTIONAL AND ASPECTUAL MEANINGS OF COME + TO_INFINITIVE IN NATIVE AND NON-NATIVE VARIETIES OF ENGLISH

Sara Gesuato[252]

Abstract.

The catenative construction COME + to_infinitive encodes the notions of 'goal-oriented motion' (literal meaning) or 'culmination of an event' (aspectual meaning). This paper explores the motional and aspectual usage of the construction in six components of the International Corpus of English representing one native and five non-native varieties (Great-Britain; East-Africa, Hong-Kong, India, Philippines, Singapore).

Of the 351 instances of the construction retrieved in the corpus), about 65% instantiate the aspectual meaning ("we came to understand that people's love cannot be quantified") and 30% the motional meaning ("One day, he came to visit"), while the remaining 5% are compatible with either interpretation (i.e. 'decide to' vs 'happen to, end up V-ing'; "people who want to come to this House will come to uphold its powers and its responsibilities"). The specific meaning activated correlates with the semantic nature of the event encoded: deliberate acts performed by agents are associated with literal COME ("if a customer comes to change the goods"), while involuntary processes experienced by sentient participants or undergone by patients are associated with metaphorical COME ("Later on I came to know what literature was"; "it came to be used in schools"). However, the correlation may be overruled by the larger co-text such as a temporal subordinating expression or embedding structure ("By the time you come to exercise the power of sale"; "why it was that they came to nominate the company").

In conclusion, like other motion verbs, COME can be used metaphorically to represent events as paths. In all varieties considered, its non-literal usage is prominent, suggesting an incipient grammaticalization. It presents the completed development of a process as a projected outcome, thus acting as a forward-oriented marker of perfective-resultative aspect.

Keywords: motion verb, catenative construction, aspectualizer, English varieties, corpus linguistics

Introduction

Temporal concepts are often understood in spatial terms. In particular, events are represented as paths through motion verbs (McIntyre 2001: 150). These metaphorically encode the notion of reaching a non-physical destination or achieving a goal, whether followed by adverbials or *to*_infinitives (e.g. "A pay rise will lead to job cuts"; "The dangerous political situation led me to leave the country"; "Let's get to the point"; "We got them to resign").

One motion verb subject to such metaphorical usage is *COME*. Followed by an adverbial, it can encode the notions of 'reaching a state' (e.g. "We came to a final decision") or 'beginning to operate' (e.g. "When did it come into power?"). Followed by a *to*_infinitive, it expresses the achievement of a result or culmination of an event (e.g. "I gradually came to see him as a friend"). In this case, *COME* is part of a catenative construction encoding a single event qualified with an aspectual, culminative nuance. The construction means 'end-up V-ing' or 'get to the point when X occurs'.

However, COME can still be used literally in the same syntactic environments. More specifically, when followed by a *to*_infinitive, *COME* can express 'goal-oriented motion' (e.g. "She had come to visit her in-laws"). In this case, the sequence *COME* + *to*_infinitive encodes two events: motion and a subsequent goal. The sequence means 'move closer the deictic centre of the speech event so as to X'.

[252] Sara Gesuato is associate professor of English language at the University of Padua, Italy, where she teaches in the school of education. Her research interests include pragmatics, genre analysis, lexical semantics, corpus linguistics and academic discourse. Recently, she has co-edited a volume on the use of technologies for research and teaching purposes (with Taylor Torsello and Busà: "Lingua inglese e mediazione linguistica", Cleup, 2004), published an anthology of speech acts ("L'inglese nelle contrattazioni private", Aracne, 2005), written a monograph on dissertation acknowledgements ("Giving credit where credit is due", UMI, 2005). She has also investigated evaluation in book blurbs and paper review guidelines, the content and structure of titles of academic publications and calls for abstracts, and the co-text of English near-synonyms of Latin and Germanic origin.

Aspectual and motional COME combine with different semantic-syntactic phraseologies. The former can co-occur with expressions denoting the passing of time (e.g. “Slowly I came to understand the situation”), and is compatible with *how*-questions whose scope is the whole construction (e.g. “How did she come to understand the situation?”). Literal *COME* is compatible with expressions denoting the single point in time at which movement takes place (e.g. “I came to visit you at 6”), is compatible with the insertion of a place adverbial before the infinitive (e.g. “I came [here] to visit you at 6”), and can be queried with *how*-questions relevant to the matrix clause (e.g. “How (by what means) did she come [here] to visit you?”).

The aspectual usage of COME + to_infinitive is exemplified in Huddleston, Pullum and Bauer (2002: 1207); its meaning is glossed as ‘come about’ or ‘happen’ in Quirk and Biber (1999: 708); finally, it is classified as an ingressive aspectualizer in Brinton (1988: 4). However, to my knowledge, no comparative study has been carried out on the literal and non-literal meanings of COME + to_infinitive.

This paper illustrates the frequency of occurrence, distribution and partial co-textual environment of the literal and aspectual usage of *COME* + *to*_infinitive with corpus-based data illustrating one native and five non-native varieties of present-day English. The goal is to describe the phraseological patterns of the two meanings across English varieties so as to trace their syntactic-semantic profiles.

Data and findings

I collected concordances of *come*, *comes*, *came* and *coming* + *to*_infinitives in the Great-Britain (GB), East-Africa (EA), Hong-Kong (HK), India (IND), Philippines (PHI) and Singapore (SIN) 1,000,000-word components of the International Corpus of English (ICE).[1] I filtered out from the query output ambiguous concordances with the word *work*, which could be classified as either a noun or a verb. Example:

(1) “the able-bodied people coming to work with disabled people” (ICE-GB\s1a-001).

I retrieved 351 instances, i.e. about 88 per component, on average. The four forms of the lexeme *COME* are represented in the corpus components. Examples:

(2) “Those who came to demand the licence back” (ICE-EA\w2e018k)

(3) “The farmer had come to report that his calf was sick” (ICE-EA\pptech-k)

(4) “a society that is coming to believe that the external examination result is the only gauge of intelligence” (ICE-EA\rep-feature-k)

(5) “the referee Alain Chaconne comes to pull the front rows up” (ICE-GB\s2a-002).

However, the two most frequently instantiated are *come* and *came*, together accounting for 86% of the data. The frequency values are similar in the GB, EA, PHI and SIN components, ranging from 37 to 47 instances. However, higher values are found in IND and, especially, HK (see Table 1).

	come	*comes*	*came*	*coming*	*Total*
GB	8	1	17	11	37
EA	19	4	16	2	41
HK	40	2	17	7	66
IND	59	6	47	6	118
PHI	29	1	11	1	42
SIN	22	3	17	5	47
Total	177	17	125	32	351

Table 1: Distribution of *COME* forms

Various tense-aspect combinations are attested (e.g. present and past continuous, present conditional, *will*-future, bare and *to*-infinitives, past perfect, present perfect continuous, gerund), which include the occasional use of modal auxiliaries. Examples:

(6) “curious people would come to stare where they had been” (ICE-GB\w2f-015)
(7) “the ambassadors of Saudi Arabia which originally asked for the allied coalition to come to save his country (ICE-GB\s2a-019)
(8) “it’s coming to be a serious issue” (ICE-HK\s1a-004)

(9) “the community has come to expect a lot more from senior officials” (ICE-HK\s2b-032)

(10) “in cases where blood or blood products do not suit the patients, PFCs can come to help” (ICE-IND\w2b-037)
(11) “they should come to see the world properly” (ICE-IND\s1a-062).

However, only three tense-aspect forms account for over 72% of the data, namely the simple present, the simple past and the present perfect. Two others, the *will*-future and the past perfect, are instantiated in only about 9% of the concordances. Table 2 shows that the frequencies of occurrence of these structures differ across components. GB and EA reveal a similar, strong preference for the simple past, which determines an under-representation of other forms. On the other hand, PHI and SIN stand out for a fairly balanced representation of the simple present, simple past and present perfect, although they both favour the simple past over other tenses/aspects. Finally, HK and IND display distinctive frequency patterns. The former shows a balanced preference for the simple present and simple past; the latter is strongly focused on the simple present.

	Simple present	*Simple past*	*Present perfect*	*Will-future*	***Past perfect***	***Pretotal***
GB	3	17	2	1	1	24
EA	6	17	6	1	4	34
HK	21	18	6	5	1	51
IND	63	4	17	9	5	98
PHI	11	14	11	1	2	39
SIN	12	18	9	2	0	41
Total	116 (33%)	88 (25%)	51 (14.5%)	19 (5.4%)	13 (3.7%)	287 (81.7%)

Table 2: Most frequent syntactic realizations of *COME* +*to*_infinitive

The sequence *COME* + *to*_infinitive typically co-occurs with animate, especially human subjects. Only 60 concordances (i.e. 17%) contain inanimate subjects. Examples:

(12) “When the history of literature of our own country comes to be written there is sure to be a page in it” (ICE-IND\s2b-048)
(13) “the allowance has come to be seen as a reward for senior citizens' past contribution to the community” (ICE-HK\w2e-002)

(14) “the role of ownership came to rest in the unwilling hands of the Civil Servants” (ICE-GB\w2b-016).

I manually tagged the concordances for their literal and aspectual meanings. I chose the former interpretation when a suitable paraphrase for a given *COME* + *to*_infinitive sequence was “purposefully move closer to the addresser/addressee in order to X”. I chose the latter when the sequence could be understood as “happen to X” or “end up X-ing”. When either interpretation was applicable, I glossed the concordance as ambiguous. Examples:

(15) “she left after a week and left with Sharad and Medha who came to escort her back home” (literal; ICE-IND\w1b-001)
(16) “and the junior comes to know who you are” (aspectual; ICE-IND\s1a-090)

(17) “when I come to read this thing” (ambiguous; ICE-IND\s2a-069).

The specific aspectual nuance of *COME* depends on the temporal nature of the processes it qualifies. With a punctual verb (e.g. *realize*), *COME* hints at an intervening process leading up to the realization of an instantaneous event. With a stative verb (e.g. *believe*), it signals the achievement of a state at the end of a timespan characterized by a different state. With a dynamic durative verb (e.g. *develop*), it expresses the gradual unfolding of a process, which results in its continuation or completion. Examples:

(18) “It is during this period that the South came to realize that the "aid" package from the North had done little to alleviate their poverty” (dynamic punctual; ICE-EA\ldsocscience-k)
(19) “Kenyans have come to believe that it is their MPs who are supposed to, for example, tarmack their roads” (stative; ICE-EA\w2e011k)
(20) “a form of learning in which voluntary responses come to be controlled by their consequences” (dynamic durative; ICE-HK\w1a-004).

Table 3 shows that, overall, the aspectual meaning is much more frequently instantiated (57%) than the literal one (37%), with a few ambiguous cases. However, the distribution of these meanings is not homogeneous across the corpus components. For instance, IND and PHI show a strong preference for the aspectual meaning, relevant to about 3/4 of their data. EA and SIN, instead, instantiate the aspectual meaning just over 50% of the time. Finally, HK and GB more frequently encode the literal meaning, which accounts for about 50% and 60% of their data, respectively.

	Literal	*Aspectual*	*Ambiguous*	*Total*
GB	22 (59.5%)	9 (24.3%)	6 (16.2%)	37 (100%)
EA	18 (43.9%)	22 (53.7%)	1 (2.4%)	41 (100%)
HK	33 (50.0%)	28 (42.4%)	5 (7.6%)	66 (100%)
IND	29 (24.5%)	83 (70.0%)	6 (5.5%)	118 (100%)
PHI	9 (21.4%)	32 (76.2%)	1 (2.4%)	42 (100%)
SIN	19 (40.4%)	26 (55.3%)	2 (4.3%)	47 (100%)
Total	130 (37.0%)	200 (57.0%)	21 (6.0%)	351 (100%)

Table 3: Distribution of literal vs aspectual meanings

Of the 200 concordances instantiating the aspectual meaning of *COME*, 46 (i.e. 23%) are accompanied by linkers or adverbials expressing temporal or causal relationships. These expressions signal the (chrono)logical link between the process and result components of the unitary events being represented. The formulas may focus on the unfolding of a process over time (e.g. *as, gradually, increasingly, more and more*), its connection to a preceding cause (e.g. *after than, later on, so that*) or its conclusion and resultant effect (e.g. *finally, thus, then*). The two most frequent ones are *when* (14 instances), and *so* (5 instances).

The corpus instantiates 133 lexemes, that is, on average, one every 2.6 concordances. Eighty-three (i.e. over 62%) are exemplified only once. The five most frequent ones are *know* (74 instances), *see*, both in the sense of ‘visually perceive’ and in that of ‘visit’ (26 instances), *think*, only in the sense of ‘reflect’ (19 instances), *be* (13 instances), and *visit* (10 instances). Together, they account for over 40% of the data.

The lexemes exemplified in the concordances encode deliberate acts performed by agents, or involuntary processes undergone by experiencers or patients. Examples:

(21) “my parents are coming to collect my sister” (deliberate act; ICE-GB\s1b-012)
(22) “We came to show our respect for him and offer our condolence to his family” (deliberate act; ICE-HK\s2b-024)
(23) “a number of factors, each of which has come to be tainted by certain developments” (involuntary process; ICE-IND\w2d-005)
(24) “the Philippine government came to recognize the extent and impact of the disease” (involuntary process; ICE-PHI\w2a-027).

	Act	*Experience*	*Unclear*
GB	26 (70.3%)	9 (24.3%)	2 (5.4%)
EA	22 (53.7%)	19 (46.3%)	0 (0.0%)
HK	41 (62.1%)	24 (36.4%)	1 (1.5%)
IND	50 (42.4%)	65 (55.1%)	3 (2.5%)
PHI	9 (21.4%)	33 (78.6%)	0 (0.0%)
SIN	24 (51.1%)	22 (46.8%)	1 (2.1%)
Total	172 (49.0%)	172 (49.0%)	7 (2.0%)

Table 4: Distribution of deliberate acts vs experiences

Their frequency patterns closely match those of the literal and aspectual meanings of *COME* (compare Tables 3 and 4). And indeed, there appears to be a correlation between the meaning activated in any specific case and the semantic nature of the event being represented. Events involving agents tend to combine with motional *COME*, while events involving experiencers or patients tend to co-occur with aspectual *COME*. Examples:

(25) "people of diverse cultures, religions and languages came to co-exist within the same political boundaries" (aspect + experience; ICE-SIN\w2a-011)
(26) "from Julius Caesar's time to ours principles and tenets have developed and come to be engraved in every enlightened society" (aspect + experience; ICE-PHI\s2b-025)
(27) "other doctors came to help" (motion + act; ICE-SIN\w2b-009).

A comparison between Table 3 and Table 5 shows the strength of the correlation between the motional and aspectual meanings of *COME*, on the one hand, and the voluntary or involuntary nature of the events represented, on the other. That is, very similar intra-corpus differences emerge with regard to the distribution of meanings.

	Act + motional	*Experience + aspectual*
GB	22 (59.5%)	8 (21.6%)
EA	15 (36.6%)	20 (48.8%)
HK	33 (50.0%)	25 (37.9%)
IND	23 (19.5%)	78 (66.1%)
PHI	7 (16.7%)	32 (76.2%)
SIN	18 (38.3%)	22 (46.8%)
Total	118 (33.6%)	185 (52.7%)

Table 5: Correlation between meanings of *COME* and event types

There appears to lack a perfect fit between the values in Table 3 and those in Table 5. This is mainly because the compatibility of a *COME* + *to*_infinitive sequence with an aspectual interpretation may correlate with co-textual features other than the semantic nature of the events represented in the *to*_infinitive. For examples, the presence of a temporal subordinating expression or a *how-*/*why*-headed embedding structure may affect the interpretation of a concordance. If such a phrase co-occurs with of verb denoting a deliberate act (in the active voice), it determines an aspectual (re-)interpretation of the text. If it occurs with a verb denoting an involuntary process (either an experience or a deliberate act in the passive voice), it simply confirms the aspectual interpretation of the text. Such introductory phrases mark about 6% of the data. Examples:

(28) "did not explain uhm the nature of the transaction or why it was that they came to nominate the company" (voluntary process + aspectual reinterpretation; ICE-HK\s2b-045)
(29) "but how does one explain how it came to be there in the first place" (involuntary process + aspectual interpretation confirmed; ICE-GB\s2a-036)
(30) "And that's why through this particular scene we come to see her rebellious character" (involuntary process + aspectual interpretation confirmed; ICE-HK\s1b-010)

(31) “Uh how is it for example that students come to develop reading skills in Filipino or English” (ambiguous process + aspectual reinterpretation; ICE-PHI\s1b-001).

Additional co-textual features possibly affecting the interpretation of a concordance include the choice of subjects with inanimate referents as the unintentional cause of otherwise voluntary processes, and also the representation of potentially deliberate acts as the outcome or endpoint of gradually evolving situations. Examples:

(32) “They were killed by armed people the likes of whom Kenyans have come to refer to with dread as bandits” (ICE-EA\ten-editorial-k)

(33) “another important inevitable event – death, which came to control the future course of life itself was ushered in” (ICE-IND\w2b-024)

(34) “But in fact when you come to think of it uh closely it’s redesigning the inside of our body rig” (ICE-HK\s2a-058).

Discussion and conclusion

The *COME* + *to*_infinitive sequence is fairly frequently attested in ICE components representing one native and five non-native varieties of present-day English. In all components, it encodes both a literal, motional meaning and an aspectual, culminative meaning, which can be identified by means of relevant paraphrases. The former expresses goal-oriented motion (i.e. “move closer so as to X”). The latter signals the gradual completion of a process or the final achievement of a goal resulting from the full development of an introductory phase (i.e. “end up X-ing” or “happen to X”), and its specific semantic nuances depend on the types of verbs it combines with: attainment of a result with stative verbs, inception of a process with dynamic durative verbs, and realization of a process with dynamic punctual verbs.

The activation of one or the other meaning strongly correlates with the nature of the events encoded in the *to*_infinitive: the representation of a deliberate act vs an involuntary experience favours, respectively, a motional vs an aspectual interpretation. At the same time, other co-textual features affect the interpretation of given instantiations; for example, the presence of a time adverbial, a *how*-/*why*-headed embedding formula or a passive favours an aspectual interpretation. Finally, a few occurrences remain ambiguous, i.e. interpretable both literally and non-literally.

In the occurrences examined, the aspectual meaning is instantiated more frequently than the motional one, together accounting for about 57% and 37%, respectively, of the data. In line with this, most of the matrix clauses and all their complements are realized in non-progressive, forms. In addition, three perfective verb phrases, the simple present, past and present perfect, account for almost 3/4 of the data. These syntactic co-choices are particularly suitable for encoding the notion of completion-culmination of a process.

However, not all corpus components display the same phraseological patterns as the corpus as a whole. For instance, unlike the other components, GB and HK encode more often the literal meaning of *COME*; *came* is the most common verb-form of *COME* in GB, while *come* is in all the others; the most frequent tense is the simple present in HK and IND, but the simple past in GB, EA, PHI and SIN; and the only lexeme represented in all components is *KNOW*. Moreover, these distinct preferences do not trace a consistent profile of the components, in the sense that different co-textual features determine different groupings of the components. For example, GB, EA, HK and SIN exemplify more deliberate acts than involuntary processes; yet, of these, only GB and HK more often instantiate the motional meaning of *COME*. In particular, distinct native vs non-native co-textual patterns have not emerged.

The data collected from six varieties of present-day English shows that *COME* + *to*_infinitive can encode both a motional and, more frequently, an aspectual meaning. The latter expresses the gradual transition from an evolving process to its endpoint or target outcome, i.e. a change of state implying the realization of an event. Because of its focus on both the development of an event and its later, prospective completion-accomplishment, aspectual *COME* + *to*_infinitive counts as a forward-oriented culminative aspectualizer. More generally, *COME* + *to*_infinitive illustrates the partial grammaticalization of a lexico-syntactic spatial expression into a prospective marker of resultativity-perfectivity. It contributes to the encoding of aspect, which is not consistently realized through morpho-syntactic means in English.

Credit is given to the ICE-EA corpus and the Technische Universität Chemnitz; the ICE-HK Corpus, the Department of Linguistics, The University of Hong Kong, and the Department of English, The Chinese University of Hong Kong; the ICE-IND Corpus, Shivaji University, Kolhapur, and the Freie Universität Berlin; the ICE-PHI Corpus and the College of Liberal Arts, De La Salle University, Manila, The Philippines; the ICE-SIN Corpus and the Department of English Language & Literature, The National University of Singapore; and the ICE-GB corpus and The Survey of English Usage, Dept. of English Language and Literature, University College London.

References

Brinton L. J. 1988. *The Development of English Aspectual Systems*. Cambridge: CUP.

Huddleston, R., Pullum G. K. and **Bauer L.** (eds.). 2002. *The Cambridge Grammar of the English Language*. Cambridge: CUP.

McIntyre, A. 2001. "Argument Blockages Induced by Verb Particles in English and German: Event Modification and Secondary Predication." In *Structural Aspects of Semantically Complex Verbs*, N. Dehé and A. Wannen (eds.). Berlin/Frankfurt/New York: Peter Lang, 131-164.

Quirk R. and **Biber D.** (eds.). 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

FEEDBACK FROM INTEGRATING CORPUS CONSULTATION IN TEACHING GREEK AS MOTHER TONGUE AT SCHOOL

Maria Giagkou[253]

Ioanna Antoniou-Kritikou[254]

Abstract

The current paper describes an attempt to introduce data-driven learning of Greek as mother tongue in the context of primary and secondary education. We will present a preliminary small-scale evaluation of corpus use and will record pupils' feedback from their first hands-on corpora experience. Our investigation involved a total of 85 pupils and 3 teachers. The evaluation tools included questionnaires, aimed at depicting the pupils' overall appreciation of the corpus-based activities they were engaged in, and the teachers' reports.

The results allow for an optimistic view of the potentials of corpus integration in L1 teaching contexts. Some of the arguments of the advocates of the use of corpora in language teaching can be attested through the results of this case study: the corpus did seem to promote collaborative learning; the pupils did inductively discover language norms by examining corpus evidence. A very promising result concerns the intention of future corpus use not only in the language classroom, but also in their homework. However, the limitations of using the corpus in primary and secondary education settings were also revealed: learning how to use the corpus and interpreting the results is a demanding task for non-expert users; it needs time to achieve a certain level of familiarization with the principles and methodological aspects of corpus linguistics.

Keywords: corpora; Greek; mother tongue; primary and secondary education; user feedback

Introduction

In the past two decades the use of corpora in language teaching and learning is gaining increasing interest. Since Tim Johns (1986; 1991) introduced data-driven learning and suggested the exploitation of corpus data in language teaching, the value of the corpus as an educational tool has been established by various studies. To summarize some of the main arguments of corpus advocates, corpora are a resource for authentic (in contrast to invented) examples in language teaching (Conrad 1999), thus being in line to Sinclair's famous saying that "Language cannot be invented, it can only be captured" (1997: 31). Furthermore, "the corpus encourages the student to act as the producer of research, rather than its passive receptacle" (McEnery and Wilson 1997: 6). McEnery and Wilson (ibid.) also argue that corpora facilitate both divergent and collaborative learning. Divergent in terms of allowing for self-paced and self-directed learning; collaborative because students share their corpus discoveries in classroom settings. This aspect of discovery learning is also emphasized by Hunston (2002: 170) who argues that discovering answers in corpus data maximizes student motivation. As for the teachers, not only does corpus evidence help in deciding what should be taught (Gavioli & Aston 2001) but also the use of corpus redefines the role of teacher as mediator and co-worker in the learning process. The teacher is regarded as an expert in the process (teach how to learn), rather that an expert in a specific subject (Bernardini 2004).

[253] Maria Giagkou is a linguist. She obtained a Master's degree in Language Technology in 2000. She is currently a PhD student at the Linguistics Department of the University of Athens. Her PhD research involves corpus-based L1 pedagogy. For the last 10 years she has been working as a research fellow at the Institute for Language and Speech Processing / "Athena" Research Centre (ILSP), being involved in a number of research projects at the ILSP Educational Technology Department, mainly in the fields of teaching of Greek as L1 and L2 and the teaching of Ancient Greek. Her research interests include technology-enhanced language teaching, corpus linguistics, text comprehension and readability.

[254] Ioanna Antoniou-Kritikou received her M.A. degree and her PhD in the field of Theoretical Linguistics from the University of Paris 7, Department of Linguistics. She worked at the Commission of the European Union (Division of translation) and as a trainer of secondary education teachers (PEK) in the field of language teaching methodology. Since 1994, she has been working as a researcher at the Educational Technology Department of the Institute for Language and Speech Processing (ILSP), where she has been developing educational multimedia software in the context of many research programmes. Ioanna Antoniou-Kritikou is currently Deputy Head of the Educational Technology Department of ILSP and has published books and research papers in the area of language teaching methodology and Educational Technology.

While the field is well supported by various important studies it is nonetheless true that most of them concern foreign language learning and tertiary education. So far as the authors are aware, there are only few contributions to the integration of corpora in pre-tertiary education, for example the works of Sabine Braun (Braun 2007) and Jan Rohrbach (Rohrbach 2003). Studies on corpus-based mother tongue (L1) education in primary or secondary school are also rare, these being mainly the relevant publications of Paul Thompson and Alison Sealey (Thompson et al. 2004; Sealey and Thompson 2004a; 2004b; 2006; 2007).

The current study concerns an attempt to introduce data-driven learning of Greek as L1 in the context of primary and secondary education[1]. We will present a preliminary small-scale evaluation of corpus use and will record pupils' feedback from their first hands-on corpora experience. The main research question that guided our investigation was: How do young learners respond to using a corpus for their linguistic inquiries in the classroom? Additionally, we tried to gain feedback on the functionality and user-friendliness of the designed interface. As will be apparent, this is an exploratory investigation of the young learners' reactions to the introduction of corpus-based L1 teaching. A generalized evaluation of corpus use was by no means in the scope of our research, neither was our goal to compare the corpus-based approach to other teaching methods.

The ETHEK corpus

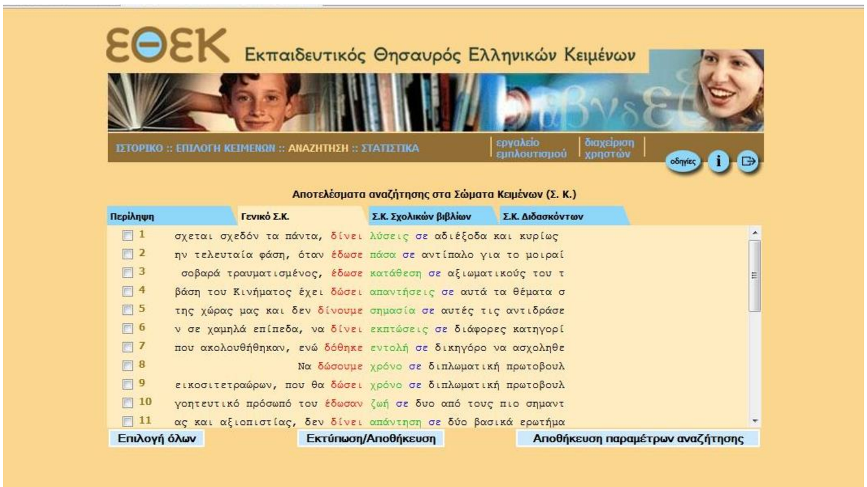
For the case study described herein we exploited the Educational Thesaurus of Greek Texts corpus (ETHEK), which consists of two sub-corpora:

I. A part of the Hellenic National Corpus (HNC: www.hnc.ilsp.gr, Institute for Language and Speech Processing) comprising 35,169,629 words. HNC is a general corpus of written Greek from various sources (books, newspapers and the internet). The texts cover a wide range of genres and topics (Hatzigeorgiou et al. 2000), thus making HNC quite representative of the Modern Greek language.

II. A corpus of instructional texts comprising 2,268,134 words. These texts come from the official textbooks used in the Greek educational system, particularly in the last two years of the primary school and in all six years of the lower and upper secondary school. They cover various curriculum subjects, science, geography, music, chemistry, etc.

The corpus is morphologically annotated. Furthermore the texts are annotated as to means of publication, writer, publisher, date of publication, genre and detailed genre, and topic and detailed topic. Especially as far as texts from school textbooks are concerned, information on the grade level the text is aimed at is included.

The ETHEK corpus can be accessed through a web-based user interface (www.xanthi.ilsp.gr/ethek) that was designed especially for children and adolescents. The ETHEK environment offers simple and advanced word/lemma/PoS searches, retrieval of word/lemma frequencies, a concordancer and a sub-corpus selection tool. A teacher module is also available that enables the teacher upload and tokenize new texts to compile his/her own corpus. The ETHEK corpus is going to be made available to all public schools in Greece by the Greek Ministry of Education.



Αποτελέσματα αναζήτησης στα Σώματα Κειμένων (Σ. Κ.)

```
□ 1       σχεται σχεδόν τα πάντα, δίνει λύσεις σε αδιέξοδα και κυρίως
□ 2       ην τελευταία φάση, όταν έδωσε πάσα σε αντίπαλο για το μοιραί
□ 3        σοβαρά τραυματισμένος, έδωσε κατάθεση σε αξιωματικούς του τ
□ 4       βάση του Κινήματος έχει δώσει απαντήσεις σε αυτά τα θέματα σ
□ 5       της χώρας μας και δεν δίνουμε σημασία σε αυτές τις αντιδράσε
□ 6       ν σε χαμηλά επίπεδα, να δίνει εκπτώσεις σε διάφορες κατηγορί
□ 7       που ακολουθήθηκαν, ενώ δόθηκε εντολή σε δικηγόρο να ασχοληθε
□ 8                             Να δώσουμε χρόνο σε διπλωματική πρωτοβουλ
□ 9       εικοσιτετραώρων, που θα δώσει χρόνο σε διπλωματική πρωτοβουλ
□ 10      γοητευτικό πρόσωπό του έδωσαν ζωή σε δυο από τους πιο σημαντ
□ 11      ας και αξιοπιστίας, δεν δίνει απάντηση σε δύο βασικά ερωτήμα
```

Screenshot from the ETHEK corpus user interface:
concordance output for the query: [lemma: *δίνω (to give)* + noun + word: *σε (to)*]

Corpus integration in L1 teaching settings: the case study

Our case study involved a total of 85 pupils distributed in 3 classes: 29 pupils from the 6th grade of primary school (11-12 years-olds), 33 pupils from the 3rd grade of lower secondary school (14-15 years-olds) and 23 pupils from the first grade of upper secondary school (15-16 years-olds).

Three two-hour sessions were held in the three classes respectively. They were conducted by the authors of the current paper, while the Greek language teacher of each class was monitoring the process. The pupils worked in groups of two or three per workstation. The sessions were structured in two parts. The first part was an introduction to corpora and their application in linguistic research, followed by a presentation of the ETHEK interface and search functionalities, since, as suggested by Fligelstone (1993), the application of corpora in teaching presupposes both *teaching about* and *teaching to exploit* corpora. The second part focused on specific language phenomena relevant to the respected grade syllabus. However, it was organised based on educational scenarios that were created in the framework of the ETHEK project and employed corpus-based language activities, i.e. study of concordance lines and word frequency data.

More specifically, the primary school students worked on the spelling of certain Greek words that, due to historic orthography, appear in orthographic variants, e.g. the word [*avγo’*] (egg) is encountered either as *αυγό* or *αβγό*, and the word [*tre’no*] (train) is spelled either as *τρένο* or *τραίνο*. The young learners searched for such words and were asked to compare their frequencies in different sub-corpora representing earlier or later texts and draw conclusions on which variant is more common. The lower secondary school students worked on the notion of polysemy. They searched for certain words and were asked to complete a dictionary entry where the multiple definitions of the word were present but the examples of use were missing. The pupils selected from the search results examples of use indicative of the different meanings of each word. Finally, the upper secondary school students were engaged in activities designed to help them discriminate between general vocabulary and terminology. They worked on a sub-corpus of texts from the physics, mathematics and chemistry textbooks and were asked to search for nouns, study their authentic examples of use and then categorize them accordingly. Towards the end of the session the pupils were given time to perform their own free corpus searches.

Feedback on the above procedure came from: (i) the pupils who were asked to fill-in a questionnaire and (ii) three teachers who attended the lessons and were asked to compile a report on their observations. The pupils’ questionnaire comprised open and close questions structured in two sections: (a) evaluation of the corpus user interface and functionalities and (b) overall appreciation of the corpus as a resource in the L1 classroom settings. The teachers’ reports included comments on the observed attitudes of their pupils during the hands-on corpus activities and their personal view on the potentials of corpus exploitation in L1 teaching, according to the experience gained during the sessions.

Results and discussion

Since the design of the ETHEK interface does not concern the current presentation, we will limit the discussion only to the results depicting the pupils’ overall appreciation of the corpus-based L1 lesson. The pupils’ questionnaire included six relevant close questions and an open one. These were linked to some of the arguments of the advocates of corpora presented in the introduction of this paper. One of the questions was linked to the argument that corpora promote collaborative learning: *Do you agree with the following statement? “Today I collaborated with my classmates more than in a typical language classroom”* (possible answers: I agree – I don’t agree – I am not sure). 67% of the pupils responded that they did collaborate with their classmates while consulting the corpus and analyzing the results.

Furthermore the argument that ‘hands-on’ corpora enhances the notion of student as researcher was attested. The pupils did not respond enthusiastically here. Less than half of them (41%) answered that they did feel like language researchers, while 34% of the pupils could not answer the relevant question. This did not come as a surprise, since it takes time to cultivate research behaviour to pupils, and the two-hour session was certainly not enough, especially if one considers the fact that this was the first time the pupils heard about what a corpus is and how it can be used in linguistic research. It is worth mentioning that when comparing the responses of the three different groups of pupils (primary, lower and upper secondary school), the latter group presented the most positive answers, whereas mainly the primary school students could not express an opinion on the question (Table 1). This is certainly not irrelevant to the maturity of the older group in terms of L1 acquisition level and familiarization to research methods.

At the same time, more than half of the pupils (51.8%) stated that they did inductively discover language norms by examining corpus evidence (*Do you agree with the following statement? “Today I discovered a language norm or attribute from the search results”*). This finding is in line with the argument of discovery learning as a significant merit of using corpora in the language classroom. When comparing the three different groups of students, those from the primary school presented more positive responses (Table 2). We would expect that the older pupils due to their ability of more abstract and inductive reasoning would find it easier to reach to generalized conclusions based on corpus evidence. This should be relevant to the kind of activities the pupils were engaged in. It seems that the activities on orthographic variants the primary school students worked on achieved to address an interesting and at the same time confusing for the young learners topic that the corpus evidence illuminated.

	I agree	**I don't agree**	**no opinion**
primary	41.4%	10.3%	48.3%
lower secondary	27.3%	39.4%	33.3%
upper secondary	60.0%	21.7%	17.4%
Total	**41.2%**	**24.7%**	**34.1%**

Table 4. Students' answers to the question "Do you agree with the following statement: Today I felt like a language researcher"

	I agree	**I don't agree**	**no opinion**
primary	65.5%	17.2%	17.2%
lower secondary	45.5%	27.3%	27.3%
upper secondary	43.5%	26.1%	30.4%
Total	**51.8%**	**23.5%**	**24.7%**

Table 5. Students' answers to the question "Do you agree with the following statement? Today I discovered a language norm or attribute from the search results"

A very promising result concerns the intention of future corpus use both in the language classroom and in their homework. The majority of the pupils answered that they would either definitely or probably like to use the corpus more often in both settings (Tables 3 and 4). Relevant to these findings was the fact that many pupils, after the end of the sessions, asked if they could use the usernames and passwords they were given during the classroom session, in order to access the ETHEK website from their homes. The statistically significant correlation between the answers to the two respective questions (Spearman's rho=0.489, significant at the 0.01 level), indicates that when the pupils appreciate the corpus as an educational tool in the classroom are more likely willing to use it in self-learning settings.

The answers to the question evaluating the pupils' overall impression of the lesson (*"Did you like today's Greek language lesson which used a corpus?"*) were generally positive, as almost 90% of the pupils at least *adequately* liked it (Table 5). Some interesting findings came from answers to the open question, which required a justification of the above answers. Some of the pupils' comments like, "*We spent too much time learning the program instead of using it"* and *"It takes time to learn how to use it"*, confirm what has already been emphasized by Braun (2007) and Kaltenböck & Mehlmauer-Larcher (2005): non-expert users need time, often not available in the strict school curriculum, to achieve a certain level of familiarization with the principles and methodological aspects of corpus linguistics.

	Frequency	**Percent**	**Cumulative percent**
Definitely yes	25	29.4%	29.4%
Maybe yes	34	40.0%	69.4%
Maybe no	18	21.2%	90.6%
Definitely no	2	2.4%	92.9%
no opinion	6	7.1%	100.0%
Total	85	100.0%	

Table 6. Students' answers to the question: "Would you like to use ETHEK more often in the Greek language lesson?"

	Frequency	**Percent**	**Cumulative percent**
Definitely yes	23	27.1%	27.1%
Maybe yes	30	35.3%	62.4%
Maybe no	11	12.9%	75.3%
Definitely no	0	0.00%	75.3%
no opinion	21	24.7%	100.0%
Total	85	100.0%	

Table 7. Students' answers to the question: "Would you use ETHEK to find answers to your questions about the Greek language in your own study at home?"

	Frequency	Percent	Cumulative Percent
very much	17	20.0%	20.0%
a lot	35	41.2%	61.2%
adequately	24	28.2%	89.4%
a little	9	10.6%	100%
not at all	0	0%	100%
Total	85	100,0	

Table 8. Students' answers to the question:
"Did you like today's Greek language lesson which used a corpus?"

Finally, the teachers reported their impressions of the integration of corpus-based activities in the language classroom. All three teachers realised the potentials of corpus use and argued that the ETHEK corpus could be employed to achieve many of the pedagogical and instructional goals outlined in the curriculum. In particular, they found the option to input their own texts very useful as a means of designing novel and personalized activities. In terms of the observed students' participation, they underlined that most of their pupils were more active in the lesson than usual. For instance, one of the teachers wrote: "*[The lesson] motivated even those pupils that due to a natural diffidence don't usually participate in the traditional language lesson…*"

Although the results of the investigation presented in the current paper are encouraging, we are not in the position to generalize them in terms of the receptiveness of corpora by primary and secondary education students. A large-scale evaluation that implements a larger number of corpus-based educational activities and a more balanced and representative students sample are necessary to provide fuller support for the arguments made herein.

Furthermore, additional effort must be devoted in recognising the contribution of corpus use to the achievement of certain educational and instructional goals outlined in L1 curriculum and in validating the argument that the use of corpora can enhance the learning process both in guided teaching and in self-directed learning as a research tool complementary or alternative to more traditional reference resources such as the lexicon, the grammar and the textbook.

Note

1. The study is part of the 'Educational Thesaurus of Greek Texts' research project, funded by the European Union and the Greek Ministry of Education (3rd CSF Operational Programme "Information Society").

References

Bernardini, S. 2004. "Corpora in the classroom: An overview and some reflections on future developments." In *How to Use Corpora in Language Teaching,* John McH Sinclair (ed.). Philadelphia: John Benjamins, 15-38.

Braun, S. 2007. "Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora." *ReCALL* 19/3: 307-328.

Conrad, S.M. 1999. "The importance of corpus-based research for language teachers." *System* 27/1: 1-18.

Fligelstone, S. 1993. "Some reflections on the question of teaching, from a corpus linguistic perspective." *ICAME Journal* 17: 97-109.

Gavioli, A. and **Aston, G.** 2001. "Enriching reality: language corpora in language pedagogy." *ELT Journal* 55/3: 238-246.

Hatzigeorgiou, N., Spiliotopoulou, A., Vacalopoulou, A., Papakostopoulou, A., Piperidis, S., Gavriilidou, M. and **Karayanis, G.** 2000. "Hellenic National Corpus (HNC): a Modern Greek corpus on the internet." In *Proceedings of the 21st Annual Meeting of the Linguistics Department,* School of Philology, Faculty of Philosophy, Aristotle University of Thessaloniki, May 2000, Thessaloniki, 812-821. (in Greek)

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Johns, T. 1986. "Microconcord: a language-learner’s research tool." *System* 14/2: 151–162.

Johns, T. 1991. "‘Should you be persuaded’: two examples of data-driven learning materials." In *Classroom Concordancing*, T. Johns & P. King (eds.). ELR Journal 4. University of Birmingham, 1-16.

Kaltenböck, G and **Mehlmauer-Larcher, B.** 2005. "Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching." *ReCALL* 17/1: 65-84.

McEnery, T. and **Wilson A.** 1997. "Teaching and Language Corpora." *ReCALL* 9/1: 7-14.

Rohrbach, J. 2003. "‘Don’t miss out on Göttingen’s nightlife’: Genreproduktion im Englischunterricht." *Praxis des Neusprachlichen Unterrichts* 50: 381–389.

Sealey, A. and **Thompson, P.** 2004a. *An investigation into corpus-based learning about language in the primary school (R000223900).* Full Research Report. http://www.esrcsocietytoday.ac.uk [Access date 14/05/2008].

Sealey, A. and **Thompson, P**. 2004b. "‘What do you call the dull words?’ Primary school children using corpus-based approaches to learn about language." *English in Education* 38/1: 80–91.

Sealey, A. and **Thompson, P.** 2006. **"**‘Nice things get said’: corpus evidence and the National Literacy Strategy." *Literacy* 40/1: 22–28.

Sealey, A. and **Thompson, P.** 2007. "Corpus, concordance, classification: Young learners in the L1 classroom." *Language Awareness* 16/3: 208-223.

Sinclair, J.M. 1997. "Corpus evidence in language description." In *Teaching and Language Corpora*, A. Wichmann, S. Fligelstone, T. McEnery & G. Knowels (eds.), London: Longman, 27-39.

Thompson, P., Sealey, A. and **Scott, M.** 2004. "Kids, concordance, collocation." Paper presented at the *6th Teaching and Language Corpora Conference*, Granada, Spain, 6-9 July 2004.

CONSTRUCTING KNOWLEDGE VIA METAPHOR IN SINGAPOREAN STUDENT WRITING: A CORPUS-BASED STUDY

Libo Guo[255]
Huaqing Hong
Shanshan Wang
Siti Azlinda

Abstract

This paper reports on work in progress of a large-scale study which seeks to examine and compare knowledge construction and the development of grammatical metaphor in Secondary 3 (Year 9) student writing in English and Social Studies. Through a combination of qualitative (systemic-functional) and quantitative (via computer-supported tool MMAX2) analyses of a sample of 42 student writings, it is shown that arguing in subject English and arguing in Social Studies employ different grammatical resources and point to different directions. Compared with subject English, which employs rankshifted embedding, Social Studies (and its parent disciplines such as History and Sociology) depends to a greater extent on grammatical metaphors to argue. This kind of work can have important implications for developing students' advanced literacy in that it can deepen our understandings of the textual features of different subject areas and their different underlying value systems.

Keywords: grammatical metaphor, knowledge, learner corpora, writing across the curriculum.

Introduction

A number of researchers (e.g., Christie 2002; Derewianka 2003; Foley 1998; Halliday 1993a, 1993b, 1994, 2004; Halliday and Matthiessen 1999; Painter, Derewianka and Torr 2007) have pointed out that mastery of grammatical metaphor, i.e. reconstrual of experience into more abstract, general level represents a landmark in the development of children's writing ability and affords them access to educational and school knowledge. Focusing on literacy demands of secondary school subjects, Martin (1993; 2007) further notes that grammatical metaphor serves different purposes in different subject areas. The discourses of history for example show a strong tendency to express cause-effect relationship within a clause rather than between clauses through grammatical metaphor. There is, however, scant comparative, empirical research on how students learn to develop grammatical metaphor across the curriculum. This paper reports on work in progress of a large-scale study which seeks to examine and compare knowledge construction and the development of grammatical metaphor in Secondary 3 (Year 9) student writing in English and Social Studies. Specific questions addressed include:

1. How do students employ grammatical metaphor in their school writing tasks?
2. How does metaphorization differ from English to Social Studies?

Grammatical metaphor

Related to but distinct from the approach taken by Deignan (2005) who focuses on lexical metaphor, we adopt the general definition that Halliday (1994: 342) gives for grammatical metaphor: 'for any given semantic configuration there will be some realization in the lexicogrammar – some wording – that can be considered CONGRUENT; there may also be various others that are in some respect "transferred", or METAPHORICAL'. In other words, once a construal of experience and an enacting of social relations are completed in the form of lexicogrammatical wording, such semantic relations can be RE-construed and RE-enacted in the form of a range of other lexicogrammatical alternatives; grammatical metaphor expands the language's resources to make meaning. It follows that grammatical metaphor falls into two broad types: ideational and interpersonal. By ideational meaning is meant what a text or part of it is about, its content, or subject matter. And interpersonal meaning of a text refers to the manner in which it addresses the intended reader or listener and the subject matter.

[255] After obtaining his first degree in English language and literature, Libo GUO taught English to university students of scienc e and technology for a number of years before pursuing his Master's degree and then PhD. His research has been mostly on language, language education, and multimodality in science texts. At the Centre for Research in Pedagogy and Practice (CRPP) at the Nat ional Institute of Education, Singapore, he is working as a Research Fellow involved in the analysis of student writing across the curriculum and classroom discourse.

An example of an ideational metaphor may be seen in the phrase ‘engine failure’, where the noun ‘failure’ serves to represent a blend of process (i.e., ‘failing’) and thing (i.e., an act of ‘failing’), as distinct from the congruent version of ‘an engine fails’, where the verb ‘fails’ serves to represent a process.

In tracing the language development of children from early childhood to adolescence, Halliday (1993b: 111) has proposed a three-step model of human semiotic development: (1) grammatical generalization as ‘the key for entering into language, and to systematic commonsense knowledge’; (2) grammatical abstractness as ‘the key for entering into literacy, and to primary educational knowledge’; and (3) grammatical metaphor as ‘the key for entering into the next level, that of secondary education, and of knowledge that is discipline-based and technical’. Further work (e.g., Derewianka 2003; Painter, Derewianka and Torr 2007) has found that before children grasp the metaphorical mode of meaning, they may have to grapple with some protometaphorical forms, which include rankshifted embeddings and faded metaphors. Rankshifted embeddings refer to ‘a mechanism whereby a unit may come to serve to realize an element of a unit of the same rank or of a lower rank’ (Derewianka 2003: 190). For example, in ‘I likede the letter that you gave me’, ‘that you gave me’ would be a clause on its own but serves now only as part of a clause, i.e., at a lower rank than before (Derewianka 2003:191). And faded metaphors are those instances ‘which were in origin metaphorical but which have since become established as the norm’ (Derewianka 2003:192), e.g., ‘do a dance’ (versus a process verb ‘to dance’), ‘make a mistake’ (versus a process verb ‘to err’), ‘take a bath’ (versus a process verb ‘to bathe’). These protometaphors are believed to model ‘the nature of grammatical metaphor for the child’ (Derewianka 2003: 192) and hence developmentally significant, although they are in themselves not yet motivated use of grammatical metaphor.

Method

Selection of linguistic features

Halliday (1998: 208-211) and Halliday and Matthiessen (1999: 244-249) categorize grammatical metaphor into thirteen types of elemental transference. Among them, Type 1 is the transference from quality (for instance, ‘unstable’) to thing (‘instability’) and Type 2 is that from process (for instance, ‘absorb’) to thing (‘absorption’). Ravelli (1988: 139) incorporates process types into the categorization of grammatical metaphor to give 19 types. In analyzing nominalization in scientific writing, Banks (2003) follows Ravelli (1988) in distinguishing different process types, and so does Derewianka (2003) in analyzing the development of grammatical metaphor from early childhood to adolescence.

In the present study, Halliday and Matthiessen’s (1999) categorization was followed as it was our purpose to identify the broad subject area variation in students’ writing. Specifically, drawing on Derewianka (2003), Halliday and Matthiessen (1999: 246-248) and Halliday (1994), an annotation scheme for ideational metaphor was devised, available from the first author upon request.

Selection of students’ essays

As part of a large-scale study of pedagogic practices in Singapore schools of a variety of geographical and socioeconomic backgrounds (Luke, Freebody, Lau and Gopinathan 2005), from 2004 to 2005, researchers at the National Institute of Education, Singapore, observed and audio-recorded more than 1200 authentic lessons of Primary 5 and Secondary 3 classroom interactions in 56 schools, and collected over 6500 pieces of students' writings (homework, class work, tests, major assignments and projects) from these lessons. This provides us with a huge database of evidence of contemporary classroom practices and students' performances in Singapore schools. For the purpose of this paper, 24 Secondary 3 student essays in English and 18 Secondary 3 student essays in Social Studies were selected. The type of writing, the genre, selected in both subjects was argumentation.

Subject	No. of Students	No.of Essays	Total runningwords
English	24	24	8830
Social Studies	18	18	8627

Table 1: Students’ essays used in the study

Analytical procedures

First, the classroom interaction was examined in order to obtain an overview of the lessons and how the writing tasks were set. Second, the associated student writing was analyzed for the occurrence of metaphorical mode of meaning. Finally, similarities and differences were established between student writings in English and Social Studies.

Three annotators were involved in annotating the 42 essays. Before the actual annotation of the student work, extensive training in grammatical metaphor and annotation tools was provided and pilot annotation carried out to ensure a high rate of agreement among the annotators. The selected linguistic features were annotated with MMAX2 tool (Müller & Strube 2006). And finally, the annotated output was uploaded to the SCoRE online query package (Hong 2005) to extract the results, which were further tabulated for statistical analysis in the next section.

Results

The findings of the study are presented in two sub-sections. First, we present a selective analysis of one Social Studies essay in terms of the use of grammatical metaphor and protometaphor. This serves to illustrate the annotation scheme, the annotation process and the interpretation of the analysis. Second, we present the pattern emerging from the corpus-based analysis of the 42 sample essays.

A sample analysis

The following figure presents a sample analysis of the first two paragraphs of a Social Studies essay. Some explanations are provided for the annotations.

Line no.	The original text	Annotated linguistic features
1 2 3 4 5 6 7 8 9 10 11 12 13	I think [1] territorial [2] dispute was the most important [3] cause [4] of [5]international [6] conflicts as compared to the other [7] causes like [8] conflicts over ideology, scarce natural resources, [9]historical [10] animosity and environmental issues.	[1]. ‘territorial’: adjective from preposition. [2]. ‘dispute’: noun from main verb. [3]. ‘cause’: noun from conjunction. [4]. ‘of international conflicts’: qualifier. [5]. ‘international’: adjective from preposition. [6]. ‘conflicts’: noun from main verb. [7]. ‘causes’: noun from conjunction. [8]. ‘conflicts’: noun from main verb. [9]. ‘historical’: adjective from preposition. [10]. ‘animosity’: noun from main verb. [11]. ‘territorial’: adjective from preposition.
14 15 16	[11]Territorial [12]dispute often [13] results from	[12]. ‘dispute’: noun from main verb. [13]. ‘results from’: verb from conjunction.

17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37	other factors like scarce natural resources and [14]historical [15] animosity. An example is the [16] dispute [17] between Malaysia and Indonesia [18] over the gas-rich area in the Ambalat region of the Sulawesi, [19] which is [20] a result of [21] territorial [22] dispute [23] over scarce natural resources. In 1962, India and China went to war [24] as a result of [25] disputes [26] over national boundaries.	[14]. 'historical': adjective from preposition. [15]. 'animosity': noun from main verb. [16]. 'dispute': noun from main verb. [17]. 'between Malaysia and Indonesia': qualifier [18]. 'over the gas-rich area in the Ambalat region of the Sulawesi': qualifier [19]. 'which is a result of territorial dispute over scarce natural resources': embedding [20]. 'as a result of': preposition from conjunction. [21]. 'territorial': adjective from preposition. [22]. 'dispute': noun from main verb. [23]. 'over scarce natural resources': qualifier [24]. 'as a result': preposition from conjunction. [25]. 'disputes': noun from main verb. [26]. 'over national boundaries': qualifier.

Table 2

Sample annotation. (Notes: The student's essay is reproduced verbatim, and the errors (if any) in the essay are retained. For ease of reference, line numbers are inserted on the left and serial numbers in square brackets (e.g., [1]) are inserted in front of those sentences whose linguistic features are commented upon in the 'Annotated linguistic features' column.)

In Lines 1 and 2, 'territorial dispute' contains two instances of grammatical metaphor. 'dispute' is here used as a noun, denoting at once both a process and a thing. So it is metaphorical, of the type 'noun from main verb'. 'territorial' is an adjective but denotes a prepositional phrase ('about the territory') and so it is metaphorical. The clause that spans Lines 14-20 is highly metaphorical. Of the several instances of metaphor, 'results from' is a verbal group but denotes a logical relationship of cause-effect congruently realized through conjunction such as 'because'. As 'results from' is at once both a process and a conjunction, it is metaphorical, of the type 'verb from conjunction'. Altogether, these two paragraphs contain 112 running words and 26 instances of grammatical metaphor and protometaphor, on average one instance per 4.30 running words.

Corpus-based analysis

In order to determine the extent of variation of student writing from subject English to Social Studies, we took a corpus-based quantitative approach to analyze the 42 essays by dividing them into two groups (English and Social Studies) and calculating the normalized frequency and text coverage of protometaphor and metaphor across the two subject areas. Raw frequency, i.e., the actual occurrences of a certain type of metaphor and protometaphor in the texts, can be informative. But, given that not all texts are of the same length, following Biber, Conrad and Reppen (1998) and McEnery, Xiao and Tono (2006: 52-53), a norm of 400 words was decided upon as the typical text length. That is, we sought to compare the normalized frequencies of metaphors and protometaphors in the two groups of student essays. The raw and normalized frequencies of metaphors and protometaphors are presented below.

Categories		English	Social Studies	Total
Metaphor	Actual instances	424	817	1241
	Normalized frequency (Ave per 400w)	19.21	37.88	28.44
Proto-metaphor	Actual instances	165	103	268
	Normalized frequency (Ave per 400w)	7.47	4.78	6.14
Total	**Actual instances**	**589**	**920**	**1509**
	Normalized frequency (Ave per 400w)	**26.68**	**42.66**	**34.58**

Table 3: Frequencies of metaphor and protometaphor in the students' essays

As shown in the table, for every 400 words of argumentative text, Social Studies essays employ 37.88/19.21=1.97 times as much grammatical metaphor as English essays. Social Studies texts are nearly twice as metaphorical as English ones. As for protometaphor, the proportion is nearly reversed. That is, English essays employ 7.47/4.78 =1.56 times as much protometaphor as do Social Studies essays. While Social Studies strives for compactness realized in grammatical metaphor, subject English strives for diffuseness realized through protometaphor such as embedding. By reference to the Social Studies text analyzed above, for example, the phrase 'territorial dispute' condenses a considerable amount of information.

The following table lists and compares the frequencies per 400 words of various types of metaphors across English and Social Studies essays.

Type of Metaphor		English	Social Studies	Total
Noun from various forms	Actual instances	266	418	684
	Normalized frequency	12.05	19.38	15.67
Preposition from conjunction	Actual instances	15	28	43
	Normalized frequency	0.68	1.30	0.99
Verb from various forms	Actual instances	6	2	8
	Normalized frequency	0.27	0.09	0.18
Adjective from various forms	Actual instances	11	76	87
	Normalized frequency	0.50	3.52	1.99
*Plus Verb	Actual instances	21	7	28
	Normalized frequency	0.95	0.32	0.64
**Plus Noun	Actual instances	0	1	1

	Normalized frequency	0	0.05	0.02
Interpersonal	Actual instances	8	0	8
	Normalized frequency	0.36	0	0.18
Other	Actual instances	97	285	382
	Normalized frequency	4.39	13.21	8.75
Total	**Actual instances**	**424**	**817**	**1241**
	Normalized frequency	**19.21**	**37.88**	**28.44**

Table 4

Frequencies of different types of metaphor in the students' essays (Notes: *The sub-category 'Plus verb' refers to the phenomenon whereby the 'content' of an action (or state) is expressed as a noun and a verb is inserted ('added') to express the idea that this action (or state) exists or happens. Examples of this sub-category include the 'took place' in 'A serious accident took place' and the 'take' in 'take a bath'. ** The sub-category 'Plus noun' refers to the phenomenon whereby a noun is added to express the idea that some event is a fact, phenomenon, statement, etc. For example, 'the fact' in 'The fact that he passed his exams…')

It can be observed from the table that metaphors involving the shifts to nouns account for more than 50% of all metaphors in both subject area essays, making them the single most frequent metaphor type in the corpus.

At the same time, neither raw frequency nor normalized frequency gives an indication of what proportion of a text one instance of metaphor or protometaphor affects, i.e., its scope at the level of discourse, or how 'powerful' or extensive each instance is. The extent to which metaphors and protometaphors affect or spread across the texts can be captured through the notion of text coverage, which can be measured by the number of words affected by metaphors and protometaphors (i.e., tokens) divided by the running words of the texts and can be expressed in percentages. For instance, in a constructed clause 'This is not what John said at the meeting', 'what John said at the meeting' is an instance of protometaphor (i.e. embedding). The extent to which this clause is affected by this embedding can be obtained by the number of words of the embedding (6 words) divided by the total number of words (9 words), to give 6/9 ≈ 67%. The following table presents the results regarding text coverage of metaphors and protometaphors in the two subject areas.

Categories		**English**	**Social Studies**	**Total**
Metaphor	Tokens	424	817	1241
	Text Cov. (%)	4.80%	9.47%	7.11%
Protometaphor	Tokens	165	103	268
	Text Cov. (%)	1.87%	1.19%	1.54%
Total	**Tokens**	**589**	**920**	**1509**
	Text Cov. (%)	**6.67%**	**10.66%**	**8.64%**

Table 5: Text coverage of metaphor and protometaphor in the students' essays

As can be seen in the table, metaphors in Social Studies essays spread across or infiltrate the texts 9.47% /4.80% = 1.97 times as much as do the metaphors in English essays. But protometaphors in English essays cover the texts 1.87% / 1.19% = 1.57 times as much as do the protometaphors in Social Studies essays. In other words, comparatively speaking, Social Studies essays are metaphorical while English essays are protometaphorical, which is in accord with a point above.

Discussion and conclusion

This paper has responded to Nesselhauf's (2004: 136) call "to investigate certain areas of grammar, lexis or discourse and go beyond single words" and "to start from functions, not from forms" and to Granger's (2002: 28) call for more interdisciplinary collaboration in the compilation and exploitation of learner corpora. Through a combination of qualitative and quantitative analyses of a sample of student writings, it is shown that arguing in English and arguing in Social Studies employ different grammatical resources and point to different directions. Compared with subject English, which employs rankshifted embedding, Social Studies (and its parent disciplines such as History and Sociology) depends to a greater extent on grammatical metaphors to argue. This provides empirical support for Martin's (1993; 2007) observations based on the analysis of a small number of texts. This kind of work has important implications for teaching advanced literacy as it deepens our understandings of the textual features of different subject areas and their different underlying value systems.

Acknowledgments

This paper makes use of data from the research project "Towards a corpus-based study of intellectual quality in Singapore classrooms" (CRP 7/07 AK), funded by the Centre for Research in Pedagogy and Practice, National Institute of Education, Singapore (http://www.crpp.nie.edu.sg). The views expressed in this paper are the authors' and do not necessarily represent the views of the Centre or the Institute.

The students' written work analyzed here was collected in the Core Research Project, Centre for Research in Pedagogy and Practice, Singapore (Luke, Freebody, Lau and Gopinathan, 2005). We are particularly grateful to Koh Kim Hong for granting the permission for secondary use of students' work. Thanks also go to Ms Zheng Jianzhen for her work in the query system.

References

Biber, D., Conrad, S. and **Reppen, R.** 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Christie, F. 2002. "The development of abstraction in adolescence in subject English." In *Developing Advanced Literacy in First and Second Languages: Meaning with Power*, M. J. Schleppegrell and M. C. Colombi (eds). Mahwah, NJ: Lawrence Erlbaum Associates, 45-66.

Deignan, A. 2005. *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Derewianka, B. 2003. "Grammatical metaphor in the transition to adolescence." In *Grammatical Metaphor: Views from Systemic Functional Linguistics*, A. Simon-Vandenbergen, M. Taverniers and L. Ravelli (eds). Amsterdam: John Benjamins, 185-219.

Foley, J. 1998. "Moving from 'common-sense knowledge' to 'educational knowledge.'" In *Language, Society and Education in Singapore: Issues and Trends*, S. Gopinathan, A. Pakir, W. K. Ho and V. Saravana (eds). Singapore: Eastern Universities Press, 245-268.

Granger, S. 2002. "A bird's-eye view of learner corpus research." In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung and S. Petch-Tyson (eds). Amsterdam: John Benjamins Publishing Company, 3-33.

Halliday, M. A. K. 1993a. "On the language of physical science." In *Writing Science: Literacy and Discursive Power*, M. A. K. Halliday and J. R. Martin (eds). London: the Falmer Press, 54-68.

Halliday, M. A. K. 1993b. "Towards a language-based theory of learning." *Linguistics and Education* 5/2: 93-116.

Halliday, M. A. K. 1994. *An Introduction to Functional Grammar* (2nd edn). London: Edward Arnold.

Halliday, M. A. K. 2004. "Language and the reshaping of human experience." In *The Language of Science: Collected Works of M. A. K. Halliday* (Volume 5), J. Webster (ed). London: Continuum, 7-23.

Halliday, M. A. K. and **Matthiessen, C. M. I. M.** 1999. *Construing Experience Through Meaning: A Language-based Approach to Cognition*. London: Cassell.

Hong, Huaqing. 2005. "SCoRE: A multimodal corpus database of education discourse in Singapore schools", In Proceedings of the Corpus Linguistics Conference Series Vol. 1, No. 1 (ISSN 1747-9398). Birmingham, UK, July 14-17, 2005.

Luke, A., Freebody, P., Lau, S., and **Gopinathan, S.** 2005. "Towards research-based innovation and reform: Singapore schooling in transition", *Asia-Pacific Journal of Education* 25/1: 7-29.

McEnery, T., Xiao, R. and **Tono, Y. 2006**. *Corpus-based Language Studies: An Advanced Resource Book.* London; Routledge.

Martin, J. R. 1993. “Life as a noun: Arresting the universe in Science and Humanities.” In *Writing Science: Literacy and Discursive Power*, M. A. K. Halliday and J. R. Martin (eds). London: the Falmer Press, 221-267.

Martin, J. R. 2007. “Construing knowledge: A functional linguistic perspective.” In *Language, Knowledge and Pedagogy: Functional Linguistic and Sociological Perspectives*, F. Christie and J. R. Martin (eds). London: Continuum, 34-64.

Müller, C. & **Strube, M.** 2006. “Multi-level annotation of linguistic data with MMAX2.” In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, S. Braun, K. Kohn, and J. Mukherjee (Eds.). Frankfurt: Peter Lang, 197-214.

Nesselhauf, N. 2004. “Learner corpora and their potential for language teaching.” In *How to Use Corpora in Language Teaching*, J. M. Sinclair (ed.). Amsterdam: John Benjamins Publishing Company, 125-152.

Painter, C., Derewianka, B., and **Torr, J.** 2007. “From microfunction to metaphor.” In *Continuing Discourse on Language: A Functional Perspective*, R. Hasan, C. Matthiessen, and J. Webster (eds). London: Equinox, 563-588.

CORPORA IN THE TEACHING OF ENGLISH IN FLEMISH SECONDARY SCHOOLS: CURRENT SITUATION AND FUTURE PERSPECTIVES

Liesbet Heyvaert[256]
An Laffut[257]

Abstract

In this paper we map out the position of corpora and corpus analysis in the teaching of English at secondary school level in Flanders (Belgium). We present the results of a questionnaire carried out among the 'key' people involved in English language teaching in Flanders, i.e. textbook writers, teachers, teacher trainees, and teacher trainers. We also discuss the official Flemish curricula for English, and report on an in-service training on corpus use that we organized for secondary school teachers.

Keywords: English, secondary level, questionnaire, in-service training

Introduction

In this paper we map out the position of corpora and corpus analysis in the teaching of English at secondary school level in Flanders (Belgium). We do this by presenting the results of a questionnaire carried out among the 'key' people involved in English language teaching in Flanders, i.e. textbook writers, teachers, teacher trainees, and teacher trainers. We also discuss the official Flemish curricula for English, and report on an in-service training on corpus use that we organized for secondary school teachers.

It is argued, firstly, that the average Flemish author of textbooks and Flemish teacher trainer/ teacher trainee are familiar with the linguistic notions that underlie corpus analysis and recognize the importance of using corpus data in language teaching, but fail to see the immediate relevance to their own practice and/or are afraid to embark on corpus analysis themselves. It is also revealed that a strong correlation exists, not so much between age and corpus-mindedness, but between familiarity with the new perspective on language underlying corpus linguistics (i.e. focus on frequency, variation, description and phraseology), on the one hand, and recognition of the need to work with corpora, on the other. Corpus skills, in other words, also depend on one's general view of the system of language. Finally, it is also shown that there appear to be *degrees of difficulty* of corpus use, depending on the teacher's/teacher trainer's/trainee's/author's familiarity with the linguistic notions underlying it: for a large majority of our subjects, searching for *collocational* or *lexical* patterning, for instance, turned out to be much easier than searching for *grammatical* or *colligational* patterning.

These findings bear out Mukherjee's (2004) conclusion that we need to systematically popularize corpus analysis among language teachers. However, the results of our survey also clearly indicate that what we need in Flanders is more than that: we also need corpus workshops that address the specific needs and problems of other target groups, i.e. textbook writers, teacher trainers and teacher trainees; we need to continue with more general introductory workshops elaborating on basic notions such as frequency, variation, phraseology; and throughout this, we have to be aware of the fact that for the average corpus user, some aspects of corpus analysis, such as looking for grammatical patterning or dealing with the statistics behind corpus data, are more 'scary' than others.

Corpora in secondary school teaching – theory vs. practice

[256] Liesbet Heyvaert obtained her MA in Germanic Languages at the University of Leuven in 1994. She also holds the degree of Master of Applied Linguistics (Macquarie University, Sydney, 1996). She completed her PhD on English nominalizations at the University of Leuven in 2002 and has since then worked as a postdoctoral researcher of the Fund for Scientific Research-Flanders (FWO). She is responsible for the English Teacher Training unit of the University of Leuven since 2004.

[257] An Laffut obtained an MA in Language and Literature (English / Dutch) from the University of Leuven in 1996, and in 2000 completed a PhD in English Linguistics with a cognitive-functional description of the locative, image impression and material/product constructions. She is currently ESP lecturer at the Leuven Language Institute and teacher trainer within the English Teacher Training unit of the K.U.Leuven.

In Mukherjee (2004), the author reports on a survey which he carried out among German secondary school teachers of English. He concludes that “the practice of English language teaching in Germany is still largely unaffected by descriptive corpus-linguistic research into authentic language use and applied corpus-linguistic suggestions of using corpus resources and corpus-based methods for teaching purposes” (Mukherjee 2004: 239). Mukherjee suggests that, for secondary school teaching of English to catch up with applied corpus linguistics, corpus linguistics needs to be ‘popularized’: it has to be ‘translated’ into the everyday needs and problems which secondary school teachers of English find themselves confronted with (from theory to practice); and it has to be taught to these teachers in a *systematic* way, i.e. through a kind of large-scale institutionalized type of workshop, which is open to any qualified teacher and organized by local teaching boards. The results of Mukherjee’s survey among German teachers of English also suggest that while current research in applied linguistics tends to focus on learner-oriented uses of corpora ('data-driven learning'), teachers are primarily concerned with mastering the ‘art’ of corpus analysis themselves. And, finally, Mukherjee's finding that the large majority of German teachers of English claims not to have come across corpus linguistics before while all making intensive use of corpus-based reference works, especially dictionaries, confirms earlier research (Tribble 2000).

The gap between corpus theory and teacher practice which Mukherjee's survey reveals has also been pointed out by others. O'Keeffe, McCarthy and Carter (2007:246), for instance, suggest that "a more critical response to the findings of corpus linguistics needs to come from teachers". In their view, corpus linguists cannot suffice by merely passing on their expertise, because "Just because a corpus linguist tells us that a certain structure is the most frequent in a corpus does not necessarily justify giving it prominence in a beginners' level course. (...) Teachers know that learners will need to learn all seven days of the week, and they know this from practice, not from theory. Their tacit knowledge needs to be brought to bear more explicitly in relation to corpus findings and their practical applications. Language teachers must continually assert their role as mediators between corpus findings and practice" (O'Keeffe, McCarthy and Carter 2007:246). What we teach about corpus linguistics and corpus analysis, in other words, crucially depends on what teachers need. In the discussion of our corpus workshop that we include below, this need for developing courses and workshops not *for* but *together with* the targeted audience will be foregrounded as one of the main conclusions to be drawn from the workshop.

To find out whether Mukherjee's findings also hold for the Flemish secondary school teacher of English, we decided to carry out a survey ourselves. In what follows, we first elaborate on the survey and the results that we obtained. We then briefly consider the Flemish curricula and discuss a corpus workshop that we organized in April 2008 in which we tried to implement some of the concerns and principles discussed in Tribble 2000, Mukherjee 2004, O'Keeffe, McCarthy and Carter (2007) and Frankenberg-Garcia (forthcoming). In a final section, we will formulate some conclusions and suggestions for future initiatives

Corpora in the teaching of English at Flemish secondary school level

It is important to make it clear from the start that the subjects of our questionnaire were not randomly chosen among the Flemish secondary school teacher population. First, unlike Mukherjee (2004), we did not so much focus on ‘the average teacher of English’, but we presented our survey to the whole set of professionals surrounding a teacher of English, i.e. major Flemish textbook writers, teacher trainers and teacher trainees. It was thereby hypothesized that this approach would enable us to find out more about the corpus attitude and expertise of people whose impact on the ‘ordinary’ teacher of English can reasonably be expected to be more direct than that of university lecturers specialized in corpus linguistics—people who, in other words, might be viewed as our natural allies in trying to 'spread the word' and introduce the teacher-in-the-field to corpus analysis. In the survey that we did carry out among teachers of English, moreover, we did not select a random group (as did Mukherjee 2004). Rather, we questioned teachers of English who attended a workshop on corpus use that was organized by the English Teacher Training Unit of the K.U.Leuven. These teachers are not *average* teachers. Even if not familiar with corpus analysis, they were assumed to have at least some idea of what a corpus is and can be in the context of language teaching.

In an attempt to obtain as much information and feedback as possible, we offered each of the groups of people that we questioned a slightly different set of questions, tailored to their expertise, needs and general background. We enclose the four different sets that we thus came up with in the appendix. In general, the questionnaire was designed to find out about the subjects'

1. familiarity with corpus analysis and with the notions underlying it;

2. knowledge of notions such as collocation, colligation, frequency, variation;

3. perspective on the potential of corpus analysis for language teaching and textbook writing;

4. future plans and ambitions with respect to corpus analysis;

5. opinion on the use of corpus data in the classroom (data-driven learning).

An overview of the main results of the survey

In general, the results of the survey are sobering and clearly show that we are nowhere near having a core of teaching professionals which themselves feel at ease with corpora and can pass on their expertise to the 'ordinary' language teacher. More particularly,

1. in spite of the fact that a large majority of our subjects admits to using corpus-based reference materials (dictionaries and grammars), they are not familiar with corpus analysis and feel insecure about tackling corpus data for teaching purposes themselves. Remarkably, this also goes for our teacher trainees, many of whom have been confronted with corpus analysis in their undergraduate studies. *The underlying problem here seems to be that even those people that frequently resort to the www or other corpora for individual (language) purposes are either unable to 'translate' teaching-related issues and questions into a corpus search; or they are simply unaware of the potential of corpus analysis in the context of their language teaching.*

2. Among the linguistic concepts that underlie corpus analysis, a distinction needs to be made between, on the one hand, the notions of collocation, frequency and variation, and, on the other hand, colligation or grammatical patterning. Except for the teacher trainees, none of our subjects was familiar with the notion of colligation. *Those people that use corpora consequently almost exclusively look for lexical patterning, data on frequeny and variation patterns (typically regional varieties). None of our subjects admitted to having ever used a corpus to check the grammatical functioning of a language item. Grammars and dictionaries remain the preferred reference materials here.*

3. A large majority of the people that filled out our survey are convinced that they will be consulting a corpus regularly in the future or they are determined to do so. *This is an encouraging result, which confirms Mukherjee's (2004) findings and further highlights the need for action and, in particular, for teacher-oriented, popularizing workshops on corpus analysis.*

4. While teachers find it most important to be able to work with corpus data to find balanced answers to questions that pupils ask, textbook writers focus on the potential of corpora, and in particular, of concordances for the development of authentic exercises and the generation of frequency-based wordlists. *Teachers thus seem particularly aware of the potential of corpora for dealing with issues to do with language change and variation, i.e. issues which they often find themselves confronted with in class and which are not (yet) dealt with in textbooks and reference works.*

5. Unlike the German teachers of English which Mukherjee (2004) questioned, the Flemish teaching professionals and trainees which we consulted were mostly in favour of paying attention to corpus analysis and search strategies in class. The ones that objected to it were typically involved in more technically-oriented classes and were afraid that corpus analysis would be too complicated for their pupils.

The Flemish curricula

In a way, it is not surprising to find that so few teachers, textbook writers and teacher trainers are familiar with corpus analysis since references to corpora in official curricula are scarce. The bibliography to the curricula for the 2nd and 3rd grade (i.e. years 3 to 6 in secondary schools), which lists books and reference material recommended for a teachers’ library, contains the only explicit mention of corpora: the *Cambridge International Dictionary of English, Collins COBUILD English Dictionary, Longman Essential Activator* and *New Longman Dictionary of Contemporary English* are all recommended as being corpus-based.

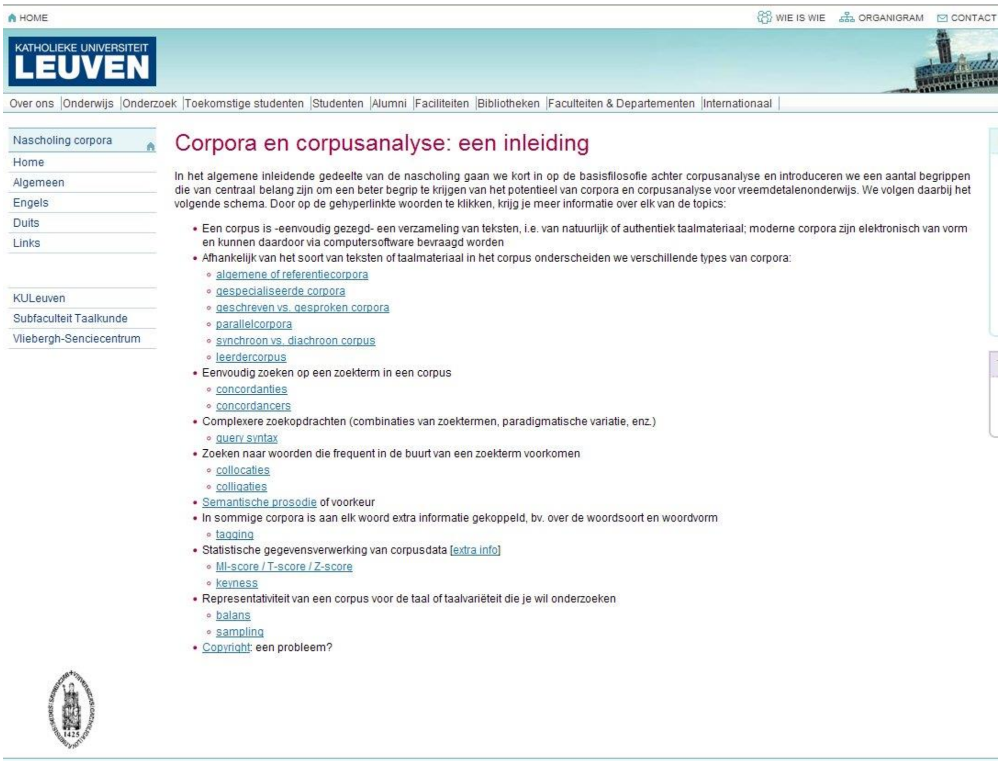
An introductory corpus workshop. Some reflections and findings.

To find out what an introductory workshop on corpus analysis ideally looks like, we recently developed an in-service training module for teachers of English and German (in cooperation with Geert Brône, Geert Stuyckens and Christine Vyncke of the University of Leuven, who were responsible for the German part of the workshop). In what follows we briefly describe its overall structure and exercise design.

Overall structure: from theory to practice

1. The workshop started off with a general introduction, in which the concept of corpora and the philosophy underlying corpus analysis was explained, and some elementary notions were introduced. In this way, participants were able to fully grasp the potential that corpus analysis holds for the teaching of foreign languages. To avoid an overload of plain, "dry" theory in the introduction, participants were encouraged to actively follow, either by

simply clicking on hyperlinks taking them to examples of the item under discussion, or by having them immediately try out certain actions (e.g. simple queries).



HOME
WIE IS WIE
ORGANIGRAM
CONTACT
KATHOLIEKE UNIVERSITEIT LEUVEN
Over ons | Onderwijs | Onderzoek | Toekomstige studenten | Studenten | Alumni | Faciliteiten | Bibliotheken | Faculteiten & Departementen | Internationaal
Nascholing corpora
Home
Algemeen
Engels
Duits
Links
KULeuven
Subfaculteit Taalkunde
Vliebergh-Senciecentrum

Corpora en corpusanalyse: een inleiding

In het algemene inleidende gedeelte van de nascholing gaan we kort in op de basisfilosofie achter corpusanalyse en introduceren we een aantal begrippen die van centraal belang zijn om een beter begrip te krijgen van het potentieel van corpora en corpusanalyse voor vreemdetalenonderwijs. We volgen daarbij het volgende schema. Door op de gehyperlinkte woorden te klikken, krijg je meer informatie over elk van de topics:

- Een corpus is -eenvoudig gezegd- een verzameling van teksten, i.e. van natuurlijk of authentiek taalmateriaal; moderne corpora zijn elektronisch van vorm en kunnen daardoor via computersoftware bevraagd worden
- Afhankelijk van het soort van teksten of taalmateriaal in het corpus onderscheiden we verschillende types van corpora:
  - algemene of referentiecorpora
  - gespecialiseerde corpora
  - geschreven vs. gesproken corpora
  - parallelcorpora
  - synchroon vs. diachroon corpus
  - leerdercorpus
- Eenvoudig zoeken op een zoekterm in een corpus
  - concordanties
  - concordancers
- Complexere zoekopdrachten (combinaties van zoektermen, paradigmatische variatie, enz.)
  - query syntax
- Zoeken naar woorden die frequent in de buurt van een zoekterm voorkomen
  - collocaties
  - colligaties
- Semantische prosodie of voorkeur
- In sommige corpora is aan elk woord extra informatie gekoppeld, bv. over de woordsoort en woordvorm
  - tagging
- Statistische gegevensverwerking van corpusdata [extra info]
  - MI-score / T-score / Z-score
  - keyness
- Representativiteit van een corpus voor de taal of taalvariëteit die je wil onderzoeken
  - balans
  - sampling
- Copyright: een probleem?

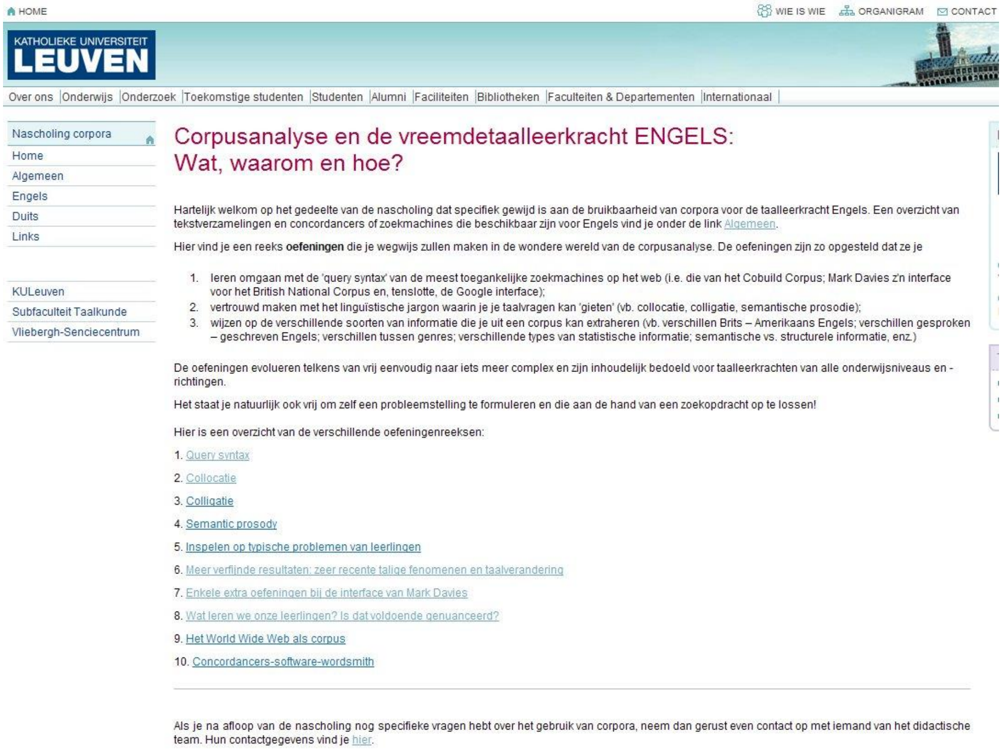
Figure 1: Corpus workshop: introductory page

2. After a brief discussion of the corpus-based nature of most contemporary dictionaries and certain grammar books, participants were shown which different types of corpora exist, which allowed them to get a taste of various types of corpus-based research (translation, diachronic change, etc.). This was followed by a hands-on illustration of basic notions such concordances, query syntax, tagging, statistical scores etc., and a demonstration of how all these features can be usefully exploited for didactic purposes (e.g. collocation, semantic prosody, colligation, frequency-based word lists). In the second part of the workshop, participants were given the opportunity to explore in more detail the concepts that were introduced in the first part.

Exercise design

The exercises were designed and ordered in a way that allowed participants to

1. work with the query syntax of some of the most accessible corpus interfaces that are freely available online (demo version of the Cobuild Corpus, the interface that Mark Davies created for the British National Corpus and the Corpus of American English, and – simply – Google); *this avoids a situation where participants cannot actually 'do' anything with what they have learned in the workshop at home because the corpora they have learned to work with are not available to them there.*
2. become familiar with the linguistic "*jargon*" in which questions about language have to be couched.
3. find out the *different kinds of information* that can be obtained from a corpus (e.g. differences between British and American English, differences between written and spoken languages, differences across genres; different types of statistical information, etc.)
4. learn to recognize the potential of corpus analysis for everyday language questions and problems which they find themselves confronted with as teachers, authors, teacher trainers—*we formulated the questions and exercises in such a way that the starting point was always a concrete situation, such as the correction of student essays, a lesson plan, tendencies observed among students in terms of variation, etc.*



Corpusanalyse en de vreemdetaalleerkracht ENGELS:
Wat, waarom en hoe?

Hartelijk welkom op het gedeelte van de nascholing dat specifiek gewijd is aan de bruikbaarheid van corpora voor de taalleerkracht Engels. Een overzicht van tekstverzamelingen en concordancers of zoekmachines die beschikbaar zijn voor Engels vind je onder de link Algemeen.

Hier vind je een reeks **oefeningen** die je wegwijs zullen maken in de wondere wereld van de corpusanalyse. De oefeningen zijn zo opgesteld dat ze je

1. leren omgaan met de 'query syntax' van de meest toegankelijke zoekmachines op het web (i.e. die van het Cobuild Corpus; Mark Davies z'n interface voor het British National Corpus en, tenslotte, de Google interface);
2. vertrouwd maken met het linguïstische jargon waarin je je taalvragen kan 'gieten' (vb. collocatie, colligatie, semantische prosodie);
3. wijzen op de verschillende soorten van informatie die je uit een corpus kan extraheren (vb. verschillen Brits – Amerikaans Engels; verschillen gesproken – geschreven Engels; verschillen tussen genres; verschillende types van statistische informatie; semantische vs. structurele informatie, enz.)

De oefeningen evolueren telkens van vrij eenvoudig naar iets meer complex en zijn inhoudelijk bedoeld voor taalleerkrachten van alle onderwijsniveaus en -richtingen.

Het staat je natuurlijk ook vrij om zelf een probleemstelling te formuleren en die aan de hand van een zoekopdracht op te lossen!

Hier is een overzicht van de verschillende oefeningenreeksen:

1. Query syntax
2. Collocatie
3. Colligatie
4. Semantic prosody
5. Inspelen op typische problemen van leerlingen
6. Meer verfijnde resultaten: zeer recente talige fenomenen en taalverandering
7. Enkele extra oefeningen bij de interface van Mark Davies
8. Wat leren we onze leerlingen? Is dat voldoende genuanceerd?
9. Het World Wide Web als corpus
10. Concordancers-software-wordsmith

Als je na afloop van de nascholing nog specifieke vragen hebt over het gebruik van corpora, neem dan gerust even contact op met iemand van het didactische team. Hun contactgegevens vind je hier.

Figure 2: Corpus workshop: exercises

All the exercises ranged from very simple, "spoon-feeding" tasks (e.g. tasks to help them find out whether a specific concordancer is case sensitive, or whether hyphens or apostrophes need to be added or omitted) to more complex yet very realistic situations, often based on our own classroom practice. Examples of the latter :

- Which adjectives and verbs are commonly used with nouns like *fun*, *question*, *problem*? Notice that these are high-frequency nouns often used by both beginning and advanced learners of English. But do your pupils actually *use* the collocating adjectives and verbs that you found?
- One of your pupils systematically uses *first off* instead of *first of all* in their essay. Find out whether this phrase is commonly used in written English. Does its frequency in written (academic) English warrant its use in an essay? Additionally, can you find out where this use originated / where it is more frequent (British or American English)?
- Research (Nesselhauf 1996) shows that non-native speakers often use "*suggest* + to-infinitive" (e.g *She suggested to leave the door locked*). Use a concordancer to check whether this is actually a frequently used pattern. Looking at the data, can you find another, much more frequent lexico-grammatical pattern?
- Use a concordancer to look at typical patterns for the noun *contact* (often followed by *with*) and the verb *to contact* followed by a noun. How do the semantics differ in these two grammatical contexts, i.e. *who* or *what* do you *contact* (verb), and when is the noun *contact* more appropriate? (From Hunston 2002:22)

Concluding remarks

An introductory workshop on corpus analysis, we found out, can only be successful if we look for inspiration among the teachers themselves and firmly ground the types of exercises and situations we offer in actual, every-day classroom situations. This is also what most participants identified as being most valuable about the workshop as we designed it. On the other hand, the participants also pointed out that the wealth of material that corpora offer is often overwhelming to them, and that they need more hands-on sessions dealing with the correct or possible interpretation(s) of corpus material.

In short, the organization of the introductory workshop on corpora as well as the survey that we carried out among teachers of English, textbook writers, teacher trainers and teacher trainees, show that much introductory corpus work remains to be done for English in Flanders' secondary schools. More importantly, however, they also suggest that if we are to bridge the gap from theory to classroom practice, what we as corpus linguists need to do most is listen to and work with the teaching professionals out there in the field.

APPENDIX

Questionnaire 1 – Teachers

1. Were you familiar with corpora or corpus linguistics before?
    - ○ Yes, I was already familiar with corpora / corpusllinguistics.
    - ○ No, I wasn't exactly familiar with this, but I had already heard about corpora.
    - ○ No, I didn't know anything about corpora.
2. If you already knew about corpora / corpus linguistics, where did you first encounter these notions?
    - ○ while studying for my Master's in languages
    - ○ through my teacher training seminars
    - ○ other: ....
3. Before participating in this workshop, were you familiar with didactic and linguistic concepts underlying corpus analysis, i.e. concepts such as
    - ○ collocation
    - ○ colligation
    - ○ frequency
    - ○ the importance of linguistic (e.g. regional, register, diachronic) variation

    Which of the concepts you were already familiar with did you acquire during your (under)graduate studies? Which ones have you encountered only recently,?
    - ○ (under)graduate studies: ....
    - ○ recently acquired: ....
4. Will you be consulting a corpus regularly as of now?
    - ○ yes / no
5. In which way / to which end would you consider using corpora?
    - ○ to develop exercises on the basis of concordances (collocations, typical patterns, …)
    - ○ to correct tests and papers (is something acceptable, idiomatic, etc.)
    - ○ to generate word lists
    - ○ to prepare a class or to find balanced answers to questions that pupils ask.
    - ○ other: ....
6. Do you feel that *learners* too can benefit from (analysing) corpus data, and that to this end textbooks should pay attention to (search) strategies, e.g. via google?
    - ○ yes / no

Questionnaire 2 – Teacher Trainees

1. Were you familiar with corpora or corpus linguistics before?
   - Yes, I was already familiar with corpora / corpusllinguistics.
   - No, I wasn't exactly familiar with this, but I had already heard about corpora.
   - No, I didn't know anything about corpora.
2. If you already knew about corpora / corpus linguistics, where did you first encounter these notions?
   - while studying for my Master's in languages
   - through my teacher training seminars
   - other: ....
3. Before participating in this workshop, were you familiar with didactic and linguistic concepts underlying corpus analysis, i.e. concepts such as
   - collocation
   - colligation
   - frequency
   - the importance of linguistic (e.g. regional, register, diachronic) variation

   Which of the concepts you were already familiar with did you acquire during your (under)graduate studies? Which ones have you encountered only recently,?
   - (under)graduate studies: ....
   - recently acquired: ....
4. Will you be consulting a corpus regularly as of now?
   - yes / no

5 In which way / to which end would you consider using corpora?
   - to develop exercises on the basis of concordances (collocations, typical patterns, …)
   - to correctt tests and papers (is something acceptable, idiomatic, etc.)
   - to generate word lists
   - to prepare a class or to find balanced answers to questions that pupils ask.
   - other: ....

6. Do you feel that *learners* too can benefit from (analysing) corpus data, and that to this end textbooks should pay attention to (search) strategies, e.g. via google?
   - yes / no

Questionnaire 3 – Textbook writers

1. Have you ever worked with (existing or self-compiled) corpora? If so, which ones?

2. If you were already familiar with corpora, where did you first hear about corpora?

3. Before participating in this workshop, were you familiar with didactic and linguistic concepts underlying corpus analysis, i.e. concepts such as

   ○ collocation

   ○ colligation

   ○ frequency

   ○ the importance of linguistic (e.g. regional, register, diachronic) variation

   Which of the concepts you were already familiar with did you acquire during your (under)graduate studies? Which ones have you encountered only recently, and how (self-study, in-service training, etc.)?

   ○ (under)graduate studies: ....

   ○ recently acquired: ....

4. In which way / to which end would you consider using corpora?

   ○ to develop exercises on the basis of concordances (collocations, typical patterns, …)

   ○ to generate word lists

   ○ other (please specify): ....

5. In developing your textbook, have you made use of corpora / corpus-analysis? If so, with regard to which specific part(s)?

   ○ when developing exercises

   ○ when generating word lists

   ○ when judging the suitability or difficulty of a texts (percentage of high and low frequency words)

   ○ when checking grammatical issues (recent changes, variation, whether prescritpive rules in traditional grammar books are borne out by actual corpus data)

   ○ other: ………………………………..

6. Do you feel that *learners* too can benefit from (analysing) corpus data, and that to this end textbooks should pay attention to (search) strategies, e.g. via google?

   ○ yes / no

7. Would you be interested in participating in possible future workshops and seminars on the use of corpora in the development of coursebook materials?

   ○ yes / no

8. Other remarks, suggestions, questions?

   ………………………………………………………………………………………..

Questionnaire 4 – Teacher trainers

1 Have you ever worked with (existing or self-compiled) corpora? If so, which ones?

…………………………………………………………………………………………

2 If you were already familiar with corpora, where did you first hear about corpora?

3 Before participating in this workshop, were you familiar with didactic and linguistic concepts underlying corpus analysis, i.e. concepts such as

- collocation
- colligation
- frequency
- the importance of linguistic (e.g. regional, register, diachronic) variation

Which of the concepts you were already familiar with did you acquire during your (under)graduate studies? Which ones have you encountered only recently, and how (self-study, in-service training, etc.)?

- (under)graduate studies: ....
- recently acquired: ....

4. In which way / to which end would you consider using corpora? Which is relevant to you personally?

- to develop exercises on the basis of concordances (collocations, typical patterns, …)
- to generate word lists
- other (please specify): ....

5 Have you made use of corpora / corpus-analysis in your own lesson plans? If so, with regard to which specific component(s)?

- when developing exercises
- when generating word lists
- when judging the suitability or difficulty of a texts (percentage of high and low frequency words)
- when checking grammatical issues (recent changes, variation, whether prescritpive rules in traditional grammar books are borne out by actual corpus data)
- other: ………………………………..

6. Have you ever introduced / discussed the use of corpora with your teacher trainees? If so, in which context (teaching vocabulary, learning strategies, ...) ………………………………………………………………………………………..

…………………………………………………………………………………………

7. Do you feel that *learners* too can benefit from (analysing) corpus data, and that to this end textbooks should pay attention to (search) strategies, e.g. via google?

- yes / no

8. Other remarks, suggestions, questions?

…………………………………………………………………………………………..

References

Frankenberg-Garcia, A. forthcoming "Raising Teachers' Awareness to Corpora". Accepted for publication in a volume with selected papers from the 7th TaLC conference in Paris.

Mukherjee, J. 2004. "Bridging the gap between applied corpus linguistics and the reality of English Language teaching in Germany" In *Applied Corpus Linguistics: A Multidimensional Perspective*, U.Connor & T.Upton (eds.). Amsterdam: Rodopi, 239-250.

Nesselhauf, N. "Learner corpora and their potential for language teaching" In *How to use corpora in language teaching*, J.McH.Sinclair (ed.). Amsterdam: Benjamins, 125-156.

O'Keeffe, A., McCarthy, M. and Carter, R. 2007. *From Corpus to Classroom. Language use and language teaching*. Cambridge: CUP.

Tribble, C. 2001. "Corpora and teaching: adjusting the gaze" Paper presented at the ICAME conference in Louvain, Belgium.

Manual http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html [Access date 11/12/2003]

CORPORA IN THE CLASSROOM AND BEYOND: ASPECTS OF CORPUS COMPETENCE

Rolf Kreyer[258]

Abstract

Over the last two decades corpus-linguistic findings have had a strong influence on the teaching of foreign languages, as can be seen in the increasing exploitation of corpus-linguistic findings for EFL teaching in the creation of textbooks, learner-dictionaries and grammars or corpus-based exercises. In addition, teachers start to find out about the usefulness of corpora for preparing lessons, for creating teaching materials or for marking exam's and students' writing. In the light of this increasing impact of corpus linguistics on EFL teaching, it is striking that teachers are still reluctant to use corpora in the classroom, more specifically, to give their students access to this vast resource of authentic language data. It is argued that this reluctance is due to a number of serious problems that arise if students are confronted with raw corpus data. The poster suggests a set of skills called 'corpus competence' as a way of addressing these problems (see Kreyer 2008 for a more detailed discussion). Each of these skills is described and suggestions are made as to possible ways of teaching corpus competence. It is argued that corpus competence will turn the corpus into a tool for non-institutionalized and self-responsible life-long learning, that will allow students to continually enhance their language skills in many different areas and become ever more competent users of their foreign target language.

Keywords: Data-driven learning, problems with unlimited access to corpora, interpretation of corpus data, corpus competence, corpus as a tool for life-long learning

Introduction

The last decade has witnessed an increase in the exploitation of corpus-linguistic findings for EFL teaching in the creation of textbooks (e.g. Grabowski & Mindt 1995), learner-dictionaries and grammars (e.g. Ungerer 1999) or corpus-based exercises (e.g. Johns 1991, Tribble & Jones 1997 and Kreyer 2007). At the same time, teachers have become more and more aware of the usefulness of corpora for preparing lessons or for marking exams and students' writing. This general acceptance of corpus-linguistic tools and findings in classroom-external EFL-contexts is faced by a considerable reluctance to use corpora in the classrooms; more specifically, teachers refrain from letting their students work with corpora: a survey among 248 English language teachers in German Secondary Schools (Mukherjee 2004) shows that while over 95% of the subjects regard corpus data as potentially useful in school, 83.9%, think that only teachers may profit from them. In addition, while over half of the teachers think that corpora might be used for teacher-centered activities, such as the creation of concordance based learning materials, corrections of class tests, or corpus-based word- or phrase lists, less than 12% think that corpora are suitable for use by students (Mukherjee 2004: 241). There still seems to be "a need to convince practising teachers to use corpora and concordances in the classroom" (Römer 2006: 129).

Problems in working with corpora

Reservations as the ones mentioned above are not unfounded; Sinclair's (2004: 2) warning that "a corpus is not a simple object, and [that] it is just as easy to derive nonsensical information from the evidence as insightful ones" cannot be overestimated. It is obvious, that serious problems may arise if students are confronted with 'unfiltered' authentic language data as, for instance, provided by concordancers. The problems are of three different kinds: (1) getting the relevant data, (2) awareness of language varieties and genres, and (3) interpreting frequencies.

As to the first problem, getting the right data, we have to be aware of the fact that while teachers usually have a particular use of an individual lexical item or construction in mind, corpora usually yield a large variety of

[258] Rolf Kreyer is an Assistant Professor of Modern English Linguistics in the department of English, American and Celtic Studies at the University of Bonn, Germany. His research interests include corpus linguistics, syntax, text linguistics, and cognitive linguistics. He is the author of Inversion in Modern Written English. Syntactic Complexity, Information Status and the Creative Writer, which was published in 2006 by Gunter Narr He has just finished a manuscript titled The Nature of Rules, Regularities and Units in Language. A Network Model of the Language System and of Language Use. This book is a cognitively oriented study on the nature of grammatical rules, corpus-linguistic concepts (e.g. collocations or n-grams) and cognitive schemas; it attempts to explain a vast array of different linguistic phenomena with the help of a unifying network model.

different instances of use. Imagine, for instance, students that are asked to ‘find’ the use of perfect aspect with *since* as a preposition or conjunction in raw corpus data. An obvious first step would be to search a corpus for all occurrences of *since* and analyse the concordance lines. The examples below, providing five random instances of the word in the British National Corpus, indicate that students may encounter great difficulties:

(1) The AC contact wasn't Shelby's first; he'd been hawking his sports car plans around since the late '50s and among the people he'd contacted with a view to building his hybrid had been his good friend Donald Healey. (BNC: A6W 1038)

(2) Since we had no home of our own, any official communication would go to Leslie's mother's address. (BNC: AMC 1333)

(3) Many hundreds of casts of these human fossils have since been made, and from them we can learn a good deal about the last appalling hours in the life of Pompeii. (BNC: ASR 347)

(4) But since then the evidence against oscillations has been mounting. (BNC: B71 782)

(5) The development of occupational pension schemes, and the rapid increase of owner-occupation since the 1950s, has given many elderly people far more economic security in old age than they ever had when they were reliant on the sale of their labour power. (BNC: CKP 861)

Of the five examples above, only tokens (1) and (4) might be immediately helpful to the learner. Example (3) is also a relevant instance but it needs experience to find that *since*, here, is understood as ‘since then’, with ‘then’ most probably referring anaphorically to a particular point in time mentioned in one of the previous sentences. The remaining two tokens show other uses of *since*. On the whole, data like these will make it difficult for students to ‘find’ the use of aspect after *since* as it is described in grammars or textbooks. What usually is regarded as the strong side of corpus work, i.e. the wealth of authentic data, turns against the inexperienced user: relevant and useful tokens are often hidden among a large number of instances that do not exemplify the phenomenon under scrutiny at all or are not very typical. This is due to phenomena like polysemy or homonymy (as in the case of causal *since* in example (2)) and the diversity of uses of a particular lexical item.

The second problem, the influence of language varieties and genre, becomes apparent if we compare spoken and written use of language. Illustrative in this respect is Mukherjee's (2002, pp. 88-95) discussion of *let me*. The expression is used very frequently in both spoken and written English, as a look at the British component of the International Corpus of English (ICE-GB) shows. It is important to note, though, that most of the instances in the written component of ICE-GB (84%) come from the ‘correspondence’ section of the corpus. Within the spoken component instances are distributed evenly across the categories ‘dialogue’, ‘monologue’ or ‘mixed’. That is, while at first sight the expression seems to be used both in written and spoken English alike, a closer look at the data reveals its confinement to one particular genre of written language, namely letters. This entails functional differences. While in letters *let me* is mainly used to formulate requests for action or for further information (examples (6) and (7)), in spoken language *let me* is used as a filler in such expressions as *let me think* or *let me see* or as a discourse organizer (Mukherjee, 2002, p. 92), as in example (8) below:

(6) … if you would let me know as soon as possible … (ICE-GB: w1b-018 #072)

(7) Let me know more details about your plans. (ICE-GB: w1b-015 #030)

(8) I think being Prince of Darkness is actually quite an attractive title isn't it <laughter> Probably more attractive than Minister for the Arts in some ways But let me just deal first with two of the general points made I mean first the politics of this (ICE-GB: s1b-022 #66-#70)

While *let me* constructions are appropriate for organizing spoken discourse it does not usually occur in this function in written language. Again, the learner who is not aware of genre differences may draw the wrong conclusions from corpus analysis.

The third problem concerns the interpretation of frequencies. To sketch out just one aspect, a given expression or construction might be frequent in a corpus because it is frequent in a particular genre or because it is one of the idiosyncrasies of a particular writer or speaker. High frequency itself does not guarantee idiomaticity and general applicability of a given expression or construction.

Aspects of Corpus competence

As should have become clear from what was said above, the use of corpora is not entirely unproblematic, and, consequently, students need to be taught how to work with corpora; they need to develop a corpus competence that allows them to benefit from the advantages of corpus linguistic methods while at the same time keeping the problematic aspects at bay. As Mauranen (2004) points out:

> […the learner] needs skills and guidance in dealing with the kind of data a corpus provides. […] this is learnable, but what I want to emphasise here is that corpus skills constitute a learning task in themselves, much in the way that many other subskills of learning do, such as group work skills. Once acquired, they facilitate learning greatly and need not be constantly refreshed. (Mauranen 2004, p. 99)

In the following, I sketch out what kinds of skills are needed to successfully work with a corpus even without the guidance of a teacher. These skills can be understood as answers to the problems mentioned above. For instance, the discussion of *since* has pointed at the problems of homonymy and polysemy. Students should be aware of this feature of languages and of the fact that a lexical search may yield tokens that instantiate phenomena other than those they were actually looking for.

Also, students should be taught some corpus-linguistic background knowledge. Highly relevant in this respect is awareness of genres and varieties and the idea of representativeness. Words, phrases and constructions are used differently in, say, spoken and written English and British and American English. Students should be aware of what a given corpus represents: for instance, they might be ill-advised to improve their idiomaticity in spoken British English by working with the BROWN corpus. Students should also know that corpora vary with regard to the degree of annotation. While a corpus like ICE-GB allows searching for syntactic constructions this is usually not possible with other corpora. Most search runs will be of a lexical kind and thus may be limited in certain respects. On the other hand, however, anything syntactic that is tied to a lexical item (like the use of present perfect after *since*) can be explored with the help of lexical search runs. Also, lexical searches can take the student a long way so that it seems reasonable to teach students how to get the maximum out of these searches rather than bothering them with more sophisticated corpora and corpus tools (e.g. ICE-CUP or SARA).

The most fundamental skill, therefore, probably is that of working with concordances. This includes sorting with regard to different positions to the left and right of the node and an awareness of the fact that patterned behaviour often extends over several words to the left or the right, as Sinclair's (1996) discussion of *naked eye* and his (1998) description of *budge* show.

Having identified relevant tokens with the help of concordance software, students are faced with the problem of interpreting the data. One central aspect, here, is the interpretation of frequencies. 1) High frequencies do not necessarily mean that a particular expression is appropriate in all situations of use: although high frequencies are generally a good indicator of idiomaticity, students should also be aware of the influence of genre: a given expression might, for instance, be very frequent in spoken English, but almost non-existent in written English. 2) Low frequencies do not necessarily indicate that the expression is odd or unidiomatic: a given expression might be rare in the whole corpus but tokens may cluster within one particular genre, such as written academic English. In this case, students should not discard the expression altogether but rather try to memorize it as a useful piece of language in particular situations. 3) A frequency of zero does not necessarily mean that the expression is ungrammatical or incorrect: students should first try to find a larger corpus and check there. If the larger corpus does not provide any instances either, this does not mean that the expression at issue is incorrect. At the most this indicates that the expression is rather rarely used and that it is likely to be unidiomatic. Students should know that a corpus can never give evidence as to ungrammaticality. This information can only be provided by native speakers or reference works.

The skills discussed so far will guarantee that students are able to use concordancers and are able to interpret the data that a concordance yields. The final skill that is needed is knowledge about how to get access to corpora and corpus software. Similarly to dictionaries, corpora and corpus software come in many shapes and forms, and, quite often, they are expensive and thus not affordable to students. It follows that as soon as students leave school or university, they will not have access to the (maybe expensive) corpora that hitherto had been available. Fortunately, the internet provides a number of corpus-linguistic resources, as a search run for 'corpus', 'corpora', 'corpus software', or 'concordancer' with the usual search engines, such as Google or Yahoo, shows. In an instant students will get access to a huge number of corpora and concordancers, the problem being how to choose from the many on offer. In this respect the other skills described above are very useful, i.e. knowledge about degrees of annotation of corpora, about concordancers and what to do with them, and finally about what kind of English the different corpora want to represent. Keeping these aspects in mind, students will be able to pick those corpora and those tools that are appropriate for the analyses they want to conduct.

Teaching corpus competence

As we have seen above, "corpus skills constitute a learning task in themselves" (Mauranen 2004: 99). It follows from that that corpus skills also constitute a teaching task. A possible teaching approach is sketched out in the following.

It seems reasonable to start off with pre-edited concordance lines which provide an optimal illustration of the phenomenon to be studied. These could be used to bring to the attention of the students the influence of genre and varieties and the need to interpret frequencies correctly. Students could then be confronted with the raw concordance data that underlie the edited concordance lines. This would give students an idea of the precision (or imprecision) of concordancers and also an impression of the different shapes and forms which a particular phenomenon can take in actual language use. This would be a good occasion to teach some basic linguistic concepts that are useful in the interpretation of raw concordance data. Working with raw data would also allow students to experience that within this (maybe overwhelming) wealth of data they can find the information they are actually looking for, since the raw concordance data was the basis for the pre-edited concordance lines.

Working with concordance lines should then be complemented by working with concordancers. A possible first task could be to replicate the data sets that had been used before. A later assignment might be to identify a number of patterns that a particular lexical item occurs in with the help of concordancing software. In this way, students could train the use of sorting procedures and the use of wildcards.

The next step would be to teach corpus-linguistic background, including aspects of corpus design and knowledge about individual corpora. On the basis of this newly gained knowledge students could go back to the concordance lines that have previously been analysed and assess the suitability or unsuitability of the database from the perspective of genre and variety. Also, students should experiment with identical search runs on different corpora and see how this changes the results. This would heighten their awareness of genre and variety influences.

At this stage, students will be independent of the teacher in terms of interpreting corpus data and of choosing the right corpus and formulating the right questions with the help of a concordancer. The final step to achieving full corpus competence is becoming independent of school or university software through becoming acquainted with the resources that the internet provides. On the basis of the skills obtained up to this point, students will be able to identify those programs and websites that are most suitable for their demands.

Conclusion

Corpus competence, as described in this poster will turn the corpus into a tool for non-institutionalized and self-responsible learning. Students equipped with this kind of competence will have at their disposal massive amounts of authentic language data that hitherto have not been accessible to them. They will be able to use these data to increase their language proficiency throughout their whole life, thus enhancing their language skills in many different areas and becoming ever more competent users of their foreign target language.

References

Grabowski, E. and **Mindt, D.** 1995. "A corpus-based learning list of irregular verbs in English." *ICAME Journal* 19: 5-22.

Johns, T. 1991. "Should you be persuaded: Two examples of data-driven learning materials." *English Language Research Journal* 4: 1-16.

Kreyer, R. 2007. "Das Korpus im Klassenzimmer." *PRAXIS des Fremdsprachlichen Unterrichts* 6: 17-21.

Kreyer, R. 2008. "Corpora in the classroom and beyond." In *Handbook of Research on Computer-Enhanced Language Acquisition and Learning*, F. Zhang & B. Barber (eds.). Hershey: Information Science Reference, 422-437.

Mauranen, A. 2004. "Spoken corpus for an ordinary learner." In *How to Use Corpora in Language Teaching*, J. Sinclair (ed.). Amsterdam: John Benjamins, 89-105.

Mukherjee, J. 2002. *Korpuslinguistik und Englischunterricht. Eine Einführung*. Frankfurt a. M.: Peter Lang.

Mukherjee, J. 2004. "Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany." In *Applied Corpus Linguistics: A Multidimensional Perspective*, U. Connor & T. A. Upton (eds.). Amsterdam: Rodopi, 239-250.

Römer, U. 2006. "Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments." *Zeitschrift für Anglistik und Amerikanistik* 54: 121-134.

Sinclair, J. 1996. "The search for units of meaning." *Textus* 9: 75-106.

Sinclair, J. 1998. "The lexical item." In *Contrastive Lexical Semantics*, E. Weigand (ed.). Amsterdam: John Benjamins, 1-24.

Sinclair, J. 2004. "Introduction." In *How to Use Corpora in Language Teaching*, J. Sinclair (ed.). Amsterdam: John Benjamins, 1-10.

Tribble, C., and **Jones, G.** 1997. *Concordances in the Classroom. A Resource Guide for Teachers.* Houston, TX: Athelstan.

Ungerer, F. 1999. *Englische Grammatik Heute*. Stuttgart: Klett.

USING RNC IN TEACHING RUSSIAN BUSINESS COMMUNICATION

Anna Levinzon[259]

Abstract

This paper reports on different approaches to organizing class activities based on using Russian National Corpus (RNC, see www.ruscorpora.ru), more precisely, on using its subcorpus of business and financial press.

Learning business communication means getting familiar not only with a specific linguistic register, which differs from literary Russian substantially and in many respects, but also with a specific cultural tradition of written communication. Although the most popular corpora-based method of L2 teaching consists of encouraging students to make their own research, this type of corpus consultation is unacceptable for our students. Being businessmen, they are interested not in linguistics but in increasing their own professional skills. Our goal here is to propose a 'plug and play' model for creating exercises. This model will enable teachers to explore many issues relating to learner access to a professional language and culture which are less time-consuming than conventional (non-corpus) methods.

As an example, we consider the semantic field "Financial markets and commerce" and explore 1) collocations of the verbs blokirovat' ("to block") and fiksirovat' ("to record, to stop"); 2) two lexical items -promoter and merchandiser- familiar to students from their L1; and 3) metaphors including concepts of 'container' and 'part of a territory.

Keywords: Russian National Corpus, business Russian, L2 acquisition

The Modern Stage of Development of Russian Business Language

Learning business communication means getting familiar not only with a specific linguistic register, which differs from standard language substantially and in many respects, but also with a specific cultural tradition of written communication. Preparing class materials for their business Russian lessons today is a real challenge for those teachers who are members of linguistic rather than professional business community. The difficulties arise first from the lack of available pedagogical materials which, in turn, is explained by the lack of researches based on the study of changes perceived in the area of Russian business communication in recent years. While in the Soviet period the linguistic register used in all genres of business documentation was full of bureaucratese, during the last two decades our society has been accepting new, primarily American, models. Statistics of performance of word *business* itself shows this evolution of social mind: according to the data from Russian National Corpus after the year 1990 it's frequency in written texts rises 62 times. "Americanisation" is a process common to business Russian in general and, in particular, to economic and financial language, the latter describing markets constructed on the pattern of USA.

The main traits of the new model are

- a less formal style in certain communicative domains (business letters, oral communication) and

- an abundance of American loanwords (consider, for instance, these found in business press examples: ***Tendenc****ia* ***expansi****i* ***teleindustri****i* ***progress****iruet (Television industry tends to progress)* or ***Menedžment kompanii loial****en* ***biznes partner****u (Company's management is loyal to its business partner*) (*Expert,2008*))

Our students with advanced English have thus to confront a rather unusual problem: they often have substantial difficulties with proper usage of lexical items familiar to them from their L1.

It is worthy of note that Russian business register is now experiencing a stage of formation and is not yet fully standardised. The existing teaching guides for Russian as well as for foreign readers don't give a complete insight in the problem of an adequate translation of American terminology and a necessity of such a translation. On the other hand every professional business community has it's own professional slang widespread in oral and informal written communication. The quantity of loanwords in such slang is usually extremely high. A teacher

[259] Anna Levinzon teaches rhetoric at State University Higher School of Economics, Moscow. She is a member of a team of Russian instructors in the University's School of Russian Studies; in years 2007 and 2008 organized Russian seminars for business students in the Center for International Fellowship, Washington. She is the author of number of articles exploring corpus-based methods of teaching Russian as L2.

unfamiliar with a real business practice can be unable to ascertain the appropriateness of usage of a given americanism in a standard business text.

Russian National Corpus as an Effective Teaching Tool

The case of usage of loanwords discussed previously illustrated that corpus data can perfectly serve pedagogical ends in the situation of the lack of standards. Being forced, in the absence of guidelines, to check his advises against the real language practice, a teacher can easily perform this task - здесь и далее by consulting a representative corpus of texts. The subcorpus of business and financial press (a part of a Russian National Corpus (RNC, see www.ruscorpora.ru)) consists of over 3670000 tokens – a volume of material ample for the needs of such a research. The subcorpus includes various types of texts representing modern standard (written) Russian: an appearance of a slang word is therefore rather unlikely. Another advantage of RNC is its ability to customize a corpus page by choosing a variety of genres, communicative situations, dates of creation etc.

It is a common knowledge that corpus is an effective tool enabling L2 learners to increase their language competence through a simulation of a linguistic research. Business Russian students are however on the whole incapable of this kind of activity: their principal aim is an acquisition of professional skills, whereas language learning plays only a small part and must be achieved without time-wasting. Teachers' goal in these specific circumstances is to provide their students all possible means of expressing themselves in the absence of rich linguistic resources – the goal which is usually achieved trough the methodology of lexical approach. If indeed collocations (and loanword collocations in particular) found in RNC are an ideal unit for teaching, the question arises as to which extent the procedure of compiling exercises may be the result of a creative process. We propose here a "plug and play" corpus-based model which enables the teacher to explore different issues relating to a professional language; there is evidence to suggest that such a model is in general less time-consuming than conventional (non-corpus) methods.

6 Steps to Create an Exercise

We suggest breaking our exercise-making job down into 5 steps.

Step1. Compose a curriculum.

During this preliminary stage it is necessary that the teacher communicate with his students and inquire into their professional interests. Set a tusk of writing an essay related to the description of students' previous professional activities or of the probable work placement in Russia may be helpful. While studying these essays the teacher specifies a content to be taught and tries to identify

- loanwords, used correctly and improperly; words and collocations that could be replaced by loanwords (loanword collocations) of the same meaning;
- widely used restricted set of metaphorical constructions which do/may effectively appear in the texts (we choose metaphors from all figures of speech because of their virtue of being highly productive and sticking to one's memory units) ;
- communicative situations in which the students are likely to find themselves in the course of their professional activities;

Thus operating the teacher finally designs kind of a program and define a set of linguistic items and tusks.

As an example we propose and discuss in what follows a topic "Financial markets and commerce", metaphorical constructions including concepts of 'container' and 'part of a territory', communicative situation of prohibition.

Step2. Ascertain a usage.

Teachers would be anxious to ensure learners not to use professional slang loan-terms in general business communication. A frequency and context analysis helps to determine the appropriateness of using professional lexical units in the standard business Russian. It will be recalled that the teacher doesn't possess himself a financial education and can't for this reason advise on the question.

Let's consider an example of two equally spread in the professional community names of a security: *obligacia* and *bond*. RNC gives us only one example of the usage of *bond*: *<Ih>Bogatstvo…sostoialo iz …bondov…i tomu* *podobnyh bumag (Miasnikov,2000) (<Their> wealth consisted of bonds and other securities)*. The analysis of the context shows that even in this case the author use the word as a descriptive-representational device: the wealth was received by deceit, *bond* sounds professional but covers a fraud.

Frequency analysis helps to solve another problem of usage: a usage of metaphorical constructions. The goal here is to dismiss some of the widespread in general written Russian variations as inappropriate in business communication.

Step3. Explore a morphological structure.

The significance of learning collocations for a successful acquisition of L2 is unquestionable, yet it may be not less helpful to analyse possible derivatives of the loanword and for this reason an analysis of the morphological structure is an indispensable step in creating exercises.

We explore hereafter a use of verb prefixes, a notable pitfall to all learners of Russian independently of their professional domain.

Step4. Select collocations.

Teacher logically wants to include in exercises those formulas most frequently used in standard business communication. Exploring one-two pages of examples presented by RNC gives an insight in occurrences of such collocations. In some cases it is advisable to compare a number of occurrences in the main corpus with a number of examples revealed by the customized subcorpus. Yet one important criterion for the selection – an ability of formula to be a shell for wide range of different meanings.

Step5. Analyze contexts.

Teachers need to consider the necessity to demonstrate a distinction between a Russian loanword meaning and the same word in English, when composing exercises. For these purposes they might

- construct a corpus-based synonymic row;

- specify rhetoric strategies;

- identify a particular communicative situation.

Step6. Choose an exercise form.

Business language learners often have no wish to enjoy their lessons as one enjoys an exciting game or a pleasant conversation. Hence the teacher should ensure the exercises are the variations which will be not the most interesting but the most useful to the learners. We suggest 7 most effective in our opinion forms:

1) Reading sets of contexts. Two tasks are possible afterwards:

   a. to guess the collocation meaning;

   b. to translate chunk-for-chunk.

2) Filling in blanks with an appropriate collocation/ part of a collocation or select contexts in which a given collocation could appear. After the teacher has selected a number of contexts in RNC he can adjust them to the students' level of understanding. Learners' proficiency with loanwords will increase if by the way of a variation the teacher proposes contexts in which the word could be used in English and can't in Russian.

3) Rewriting sentences using words/collocations given. The synonymic rows found in RNC at the previous stage could serve a material here.

4)Correcting mistakes. While preparing this exercise the teacher himself "spoils" the examples revealed by RNC.

5) Making a sentence using items given. We would suggest presenting in the set of items only one part of a collocation and thus oblige students to recall the whole.

6) Answering questions using items given. Apparently many affirmative sentences revealed by RNC are easily transformed into interrogations with interrogative words.

7) Writing a text. RNC presents a unique possibility of using as models extracts of different genres common in business communication. As the less time-consuming for the teacher and the most effective for the students

exercises we could recommend writing a business letter. First the teacher demonstrates found in RNC extracts containing desired teaching material, then the students are asked to copy the pattern.

We provide here three examples of implementation of our model as applied to the topic “Financial markets and commerce”. All the data presented unless otherwise qualified are revealed by the the RNC subcorpus of business and financial press.

Example1. Collocations of the verbs blokirovat’ and fiksirovat’

Step2

a) A translated expression *blokiruiuš*^{i}*ij paket akcij* (*blocking stake)* is widespread in financial communication. The analysis of its frequencies proves that the expression can’t be considered as a slang one: for 72 tokens of the verb 24 are found in the collocation.

b) Let’s consider two variations of the verb *fiksirovat*^{i} used by professional financial market players: *fiksanut*^{i}*sa (to close a position*) and *fikst inkam (fixed income*). RNC doesn’t present any of the words given. Hence the teacher must strictly recommend the students not to use them in any formal situation.

Step3

a) Our special attention here is directed to verb prefixes. RNC provides teacher a possibility of search of “a part of a word”: we assign a special morphological characteristic (“verb” in our case) to this part of word and designate the rest of it with a * sign (*blok*) . Blokirovati is commonly used with the prefixes za- (perfective aspect) and raz- (termination).

It is worth of note that the verb is a derivative from the noun blok (block, unit) popular in a political discourse, yet the reflexive verb with the same meaning blokirovatisa s (to cooperate with) must be classified as occasionalism (1 token in the whole body of RNC).

b) There is only one prefix used with the verb fixirovati: za- (perfective aspect).

It seems that the verb was formed after the same model as the previous one, nonetheless the noun fiks (earnings minus bonus, pure salary) exists only in financial slang.

Step 4

a) Corpus data permit us to select following collocations: blokiruiušiij paket akcij, blokirovati proekt/ reshenie/plany/popytky/rabotu (blocking stake, to prevent a project/a decision/plans/to block efforts/activity).

b) We observe two groups of collocations: fixirovat’ reshenie/predlojeni and fixirovati cenu/dohod/ob’em/marju (to record a decision/ suggestion and to terminate the growth of a price/income/volume/margin).

Step 5

The semantic of *blokirovat*^{i} is always connected with the communicative situation of prohibition and in many contexts has negative connotations: *SSA sumeli blokirovat*^{i} *proekt pod predlogom nehvatki sredstv(Zavarski, 1996) (USA have stopped the project under the pretence <false as we see from the further context> of the lack of funds).* As in English its synonym is *zamoraživat*^{i} *(to freeze).*

a)The meaning of Russian loanword *fixirovat*^{i} differs substantially from the semantic English *to fix*. RNC data permit us to construct two synonymic rows: *zapisyvat*^{i}, *otmechat*^{i} *(to record, to write down)* and *ostanovit*^{i}*, ne dat’ dvigat*^{i}*sa (to stop, to prevent movement).* This latter is also specific for the situation of prohibition: while *blokirovat’* often means unfriendly preventing of an action, *fixirovat*^{i} in many contexts prevents a hostile movement. The former meaning makes impossible the word-by-word translations of many English collocations with the verb *to fix*: *to fix a day*, for instance, must be translated as *naznačit*^{i} (but not *fixirovat*^{i}*) den’* (RNC gives 528 contexts of the first collocations and 0 of the second), because the authors rarely refer to writing the dates in a document.

Step 6

Exercises used to develop learners' knowledge of given lexical units are concentrated on the following :

- an incongruity of meanings of Russian and English verbs;
- positive or negative appraisal.

Example 2. Loanwords promouter and merčandaizer

The two words - *promoter* and *merchandiser* - are widely used in the professional communication as symbols of professional status. They are relatively new (20 contexts for both words before the year 2000 and 79 after).

Step 3

The correspondent nouns designating the professional activities sound as *promoušen* and *merčendaizing*. The second morphological structure is obviously less alien to Russian: RNC reveals 9761 nouns with the *–ing* ending and only 585 with the *–šen/šn* ending.

Step 4

Promouter and *merčandaizer* do not enter in any collocations – the fact that marks their special position in general Russian.

Step 5

The analysis reveals a frequent usage of *promouter* as a loanword that needs an explanation and might be specified by the everyday conversation noun *raskrutka (spin-off)*: *Posle takogo promoušena (raskrutki) devočku znala vs*^{i}*a Moskva (Kononova, 2002) (After this promotion (a spin-off) the girl was well-known in Moscow*). The usual rhetoric strategy underneath the contexts is irony: *desant promouterov (Diadik, 2002)(promoter assault), promouter Fedorov (Kostikova, 1997)* (a “foreign” professional status in contrast with a typically Russian name). *Merchandaizer* as RNC data studies reveal is a word with more neutral connotations, used generally without explanations.

Step 6

The exercises exploring these two nouns attempt to provide the ground for the adequate comprehension of the author’s intentions. The student himself must be prepared to use the two loanwords in a very restricted set of communicative situation; he is taught to avoid the misunderstanding of his language behaviour and to present, if necessary, sufficient explanations.

Example 3. Metaphorical constructions including concepts of ‘container’ and ‘part of a territory’

Step 2

RNC presents teachers a possibility of semantic search that can provide a huge database of metaphorical constructions. We have chosen two popular in general Russian metaphors including concepts of ‘container’ and ‘part of a territory’. The comparison of occurrences revealed by the whole corpus and by the subcorpus of business and financial press demonstrates some particularities in the representation of these wide metaphorical concepts in business Russian.

Step 4

The most common for general Russian metaphors of the kind are collocations of the words *mor*^{i}*e (sea)* - *mor*^{i}*e krovi/sveta/udači (sea of blood/light/fortune*); *istočnik (source) – istočnik radosti/ bed (source of joy/misfortune)* and *granicy (boundaries) – granicy svobody/ otvetstvennosty (boundaries of freedom/responsibility).*

The subcorpus of business and financial press offers the data of different kind. While *sources* and *boundaries* , though in different formulas, conserve their significance : *istočnik informacii/ finansov (source of* *information/funds), – granicy torgovli/ rosta (boundaries of commerce/growth), sea* shows itself extremely unpopular – for over 1 000 metaphorical constructions with *mor*^{i}*e* in RNC only 2 in the subcorpus of business and financial press. In addition to those the subcorpus reveals three more keywords characteristic for the business language: *kanal (channel) – kanal peredači deneg/ sv*^{i}*azi (money transfer/ communication channel), paket (portfolio) – paket akcii/ documentov (block of share, package of documents).*

Step 6

The exercises are based on the idea that an important part of written language acquisition is the ability to comprehend and produce metaphors taking into account the register of the communication

References

Aston, G. (1999). Corpus use and learning to translate. In: *Textus 12*, 289-314

Aston, Guy (2000): Corpora and language teaching. In: Burnard, Lou & McEnery, Tony (eds), 7- 17

Lewis, M. (1993). The lexical approach: The state of ELT and the way forward. Hove, England: Language Teaching Publications.

Sharov, S (2007). Centralizovanoe planirovanie ili stihia rynka. Moskva:Dialog.

Wichmann, Anne; Fligelstone, Steven; McEnery, Tony & Knowles, Gerry (eds) (1997): Teaching and Language Corpora. London: Longman.

THE CONCEPT OF "TEXT FACET" AS A MEANS TO ACHIEVE PEDAGOGICAL INDEXATION OF A TEXT BASE DEDICATED TO LANGUAGE TEACHING

Mathieu Loiseau

Georges Antoniadis

Claude Ponton[260]

Abstract

This communication is meant to present our project of pedagogically indexed text base. After introducing the notion of pedagogical indexation, which needs to be articulated around the teachers needs, we explain to which extent existing pedagogical resource description standards are inadequate to achieve pedagogical indexation for raw texts for language teaching. We then introduce the notions underlying the creation of our prototype through a fictional study case. The notions, which we introduce, are meant to be able to take into account the pedagogical context of the potential use of the text when giving a value to its pedagogical properties. These notions include text facet, view of a text according to a facet for a given pedagogical context, homogeneous text collection and text visualization. Through the use of these notions our prototype will allow teachers to query for texts depending on pedagogical criteria and provide them with assistance for the actual choice of the text.

Keywords: Computer Assisted Language Learning (CALL), Pedagogical Indexation, Natural Language Processing (NLP), Pedagogical Resource Description

Introduction

Text base and pedagogical indexation

Despite the popularity of the communicative approach (Levy 1997:123) and the increased use of authentic texts[261], there is no text base available that allows teachers to query in language didactics relevant terms. Teachers have adapted some of their practices to existing computer tools, such as in Data Driven Learning (DDL) (Johns 1991) or proposed methodology for "pedagogic mediation of corpora" (Braun 2005). All the same, tool-wise, some flaws of CALL systems identified in (Antoniadis *et al.* 2004) remain representative of the situation of language corpora for language teaching: if a teacher seeks to find a text in a corpus, systems will not allow him/her to express his/her query in terms of his/her set of problems: using pedagogical concepts.

Our project of pedagogically indexed text base directly stems from the previous observation. As part of this project, a prototype is being implemented. In order to present some of the concepts underlying its design we define the notion of pedagogical indexation as "*indexation performed following a documentary language describing the objects according to pedagogical criteria (relevant to didactics)".*(Loiseau *et al.* 2005).

Users' practices

In order to try to adapt the system to the actual teachers' practices, we decided to adopt an empirical approach: we performed a study in three parts. We initiated it with 8 interviews of language teachers of different experience, taught language and *computer literacy* (Bawden 2001). We used the information we gathered to prepare a short questionnaire destined to grasp how teachers handle authentic texts and the classification and research of texts. This questionnaire was answered by 133 teachers and allowed us to validate the hypothesis that a given text can

[260] Georges Antoniadis and Claude Ponton are both Maitre de conférence at the LIDILEM laboratory in Grenoble. They have worked in NLP in automated generation of text and recently focused on the added value NLP functions could provide in CALL systems, in particular with the MIRTO platform. They supervise the PhD thesis of Mathieu Loiseau, the object of which is the creation of the model and prototype being discussed in this article.

[261] in (Taylor, 1994) Taylor quotes various consistant definitions of "authentic text", among which Nunan's: "A rule of thumb for authentic here is any material which has not been specifically produced for the purposes of language teaching."

be used in a variety of pedagogical contexts[262] (Loiseau *et al.* 2008). We also concluded that teachers favor authentic texts and that they resort to specially constructed texts when they want to control their linguistic content (grammatical structures, vocabulary), especially with beginners groups. We then issued a longer questionnaire, meant to precise the information gathered in the first questionnaire and to isolate research criteria.

Connecting thread

We will expose our conclusions, confront them with existing pedagogical resource description standards and introduce some of the concepts underlying the design of the prototype to the light of a virtual case study. We will imagine the case of an English teacher, whom we will call Bert for the sake of not repeating "the teacher / his or her / him or her" throughout the article. Bert wants to work with a group of students on the preterit tense and seeks texts in order to prepare some activities around this grammatical notion. We will follow Bert's steps throughout the article and try to show the actual consequences of the different modeling options on his practices.

Inadequacy of pedagogical resource description standards

In our virtual case study Bert wants to work on the preterit and therefore is most likely to be looking for a text containing occurrences of preterit[263]. Considering which is the best and most natural way for a teacher to phrase his/her query is a problem which ought to be addressed in a later version of the prototype[264], we will focus here on how to design the system so that it can answer this query: "text containing occurrences of preterit forms of verbs".

In order to retrieve resources based on their pedagogical properties various standards have been developed. Among these standards we are going to consider here the case of Learning Object Metadata (LOM) (IEEE 2002)., which is representative of all the standards we have studied[265]. LOM proposes a set of data elements (more than seventy) meant to describe the properties of a "pedagogical object" that is to say "*any entity – digital or non-digital – that may be used for learning, education or training*". A text to be used in a language learning activity undoubtedly satisfies this description.

Now, let us imagine Bert using a text base, the objects of which are described with LOM data elements. A text containing preterit verb forms can be described as "Text adequate for the introduction of the preterit tense" using LOM data element 5.10, "Description". But the text might also be adequate to perform a phonetic exercise on compounds, to work on the lexical field of sports or any other use. Now, one can wonder whether it is possible to actually list every single potential uses of a text. From Bert's point of view, this means that a text described as adequate for work on the preterit tense will be so, but one that has not been described as adequate might be, all the same. The only way for Bert to figure out, is to actually read the articles, which cannot be considered a significant upgrade from his present practices.

Appropriately describing a text using LOM raises a concern of exhaustiveness, in that the index of the text would need to reference all the possible uses of the text. It is not because one annotator has considered the text fit for a given activity that it cannot be used for others. Indeed our second questionnaire not only confirms that a given text can be used in various pedagogical contexts, it also establishes that pedagogical properties of the text actually depend on one another. For instance the difficulty of the text (LOM descriptor 5.8, "difficulty") depends on what is to be made of it (LOM descriptor 5.10, "description") and with whom (LOM descriptor 5.7 "Typical age range"). LOM considers pedagogical properties as intrinsic to the resource described. In the case of raw resources, such as texts in the context of our work, an exhaustive description of the resource is bound to be extremely tedious, if feasible. We therefore need to focus on a different method of description, which would not require considering *a priori* all the possible combinations of properties of the text.

Text facets and views of a text according to a facet for a pedagogical context

Definitions

To be able to take into account the various parameters influencing the properties of the text we introduce the notion of text facet: "*a* text facet *is a property defined with a view to the text's pedagogical exploitation in*

[262] By "pedagogical context" we mean the didactical goals and all the characteristics of the audience (level, age, interests, etc.) and of the institution (track/diploma, material constraints, number of learners, etc.)

[263] We assume here he does not look for a text to be used in a rephrasing structural exercise of the "Change the verbs in the following text into past tense" type.

[264] Preterite *vs* preterit *vs* simple past *vs* past historic *vs* (\w*ed) *vs* VPret, etc.

[265] Among which are Dublin Core Metadata Elements, GEM, EdNA, SCORM, IMS Metadata

language teaching, accompanied by at least one mechanism to compute (automatically or not) the value of this property for any text depending on a given pedagogical context".

Let us come back to Bert's query, looking for a text containing occurrences of preterit. A useful facet for this query would be a facet that we will call "representative elements of a notion count". From now on, we will refer to this facet as F_{RepEt}. The mechanism involved to compute the value of this facet would regroup natural language processing (NLP) a morphological analyzer, a pattern matching program and a counter. Through his query, Bert specifies the pedagogical context by which to compute the value of the facet. In our case he wants to work on the preterit, the system will therefore compute for each text the number of representative elements of the notion "preterit" using the result of the morphological analysis. This value, computed for each text and depending on the pedagogical context is called a *view* of the text according to F_{RepEt} for the pedagogical context "preterit form of verbs".

Pedagogical context and constraints

With the views of the texts, Bert is now able to know whether each text contains occurrences of the preterit and how many. Given this information, he or she will obviously not be interested by certain texts (those not containing any occurrence for instance). Rather than letting him browse through all the texts contained in the text base, we could slightly modify F_{RepEt} in order to allow a more precise pedagogical context. Our study showed that depending on the kind of activity they want to perform with the text, teachers do not look for the same number of occurrences of the notion they wish to work on. Introducing the notion, for instance requires less occurrences than compiling a gap-filling structural exercise. It therefore seems relevant to let Bert constrain the value of F_{RepEt} through an extended pedagogical context changing his query to "texts containing at least 4 occurrences of preterit". This new constrained version of F_{RepEt} will be called F_{RepEtC}. The view of a text satisfying the condition according to F_{RepEtC} will return the number of occurrences of the structure. If the text does not satisfy the condition, its view according to F_{RepEtC} will be called empty. To a given pedagogical context, the system will yield all the texts with a corresponding non empty view.

Homogeneous text collection

At this point, Bert has been given access to a subset of texts satisfying his original query: "texts containing at least for occurrences of preterit". Still, depending on the number of text indexed and the variety of their content, he might be given a wide choice of texts. In such cases the teacher should be able to gradually refine his/her query using different facets until the number of texts is low enough for him/her to choose one. This ultimate choice cannot be performed by the system, for it involves notions that are very difficult to model and compute automatically (such as the theme of the text) or that are not yet computable (the interest the students might have in the text).

In order to allow the user to gradually refine his/her queries, we introduce the notion of view of a collection of texts[266]. The view of a collection of texts C_1 according to a facet F for a pedagogical context CP is a collection C_2 of texts containing all the texts of C_1 the view of which is not empty (according to F for CP). We call C_2 a "*homogeneous text collection*" in that all the texts it contains at least have in common the property of satisfying the constraints enunciated in CP.

The consequence is that Bert does not necessarily have to formulate a complete query and can specify it further, provided that the system contains enough texts satisfying the most simple version of his query; and this, without having to compute the views all over again. To follow up on our example, we can imagine that Bert's query yielded too many results. In order to narrow down the choices we introduce a new facet: F_{WC}, which counts the number of words contained in the text. As we said it, Bert wants to introduce the notion based on a comprehension activity. He wants the text to be interesting, and does not require it to be too dense in occurrences of the preterit. His students are still beginners; the texts therefore need not be too long. He is looking for texts of 250 words (with a tolerance of 25%) among those containing 4 occurrences or more of preterit.

In figure 1 below, C_C stands for complete collection, C_1 could be the view of C_C according to F_{RepEtC} for the pedagogical context "at least for occurrences of preterit", as requested by Bert. In this case, C_2 is the view of C_1 according to F_{WC} for the pedagogical context "250 words ± 25%". We will explain in the next paragraph, the notion of visualization.

[266] We use the term "text collection" in order to distance ourselves from the constraints of corpora: "Words such as collection and archive refer to sets of texts that do not need to be selected or do not need to be ordered or the selection and/or ordering do not need to be on linguistic criteria. They are therefore quite unlike corpora." (Sinclair 1996)

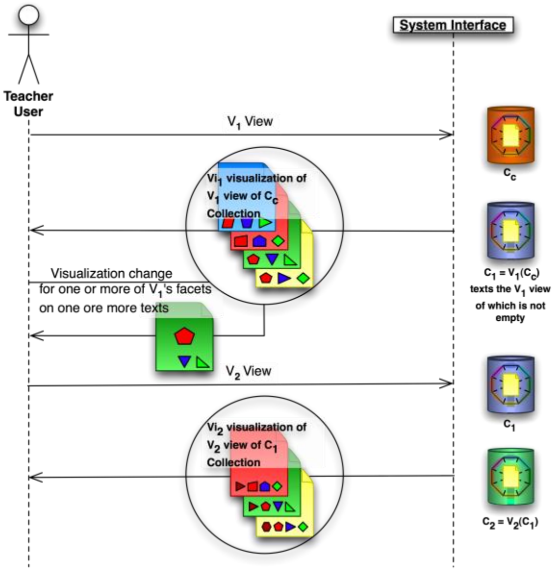
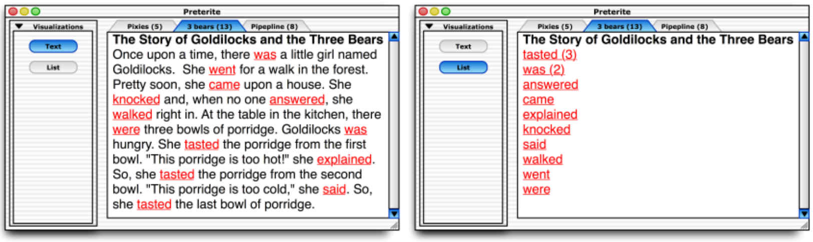


Figure 1 - Example of interaction between a language teacher and the system to refine his or her original query**.**

Qualitative access to the facets: visualizations

For the sake of our example, the system yields a dozen texts among which to choose. So far the system just acted as a filter based upon the pedagogical context submitted by the user. At this point, there are various criteria that the system as we have described it – a system implementing only two facets: F_{WC} and F_{RepEtC} – has used all the information it disposes of. The choice between the candidate texts can only be done by Bert himself, who knows what the center of interests of his students are, who can evaluate the adequacy of the text with the level of the students and what uses of the preterit he wants to display to his students. For the latter, the system is able to help: to be able to count the occurrences of preterit, it has annotated them and can re-use this annotation to offer Bert some assistance. For each facet, the system associates one or more graphical representations, which we call visualization. These visualizations can either use the view itself or the underlying information used to compute it. For instance, in the case of F_{RepEtC}, computation of the view required a morphological analysis of the text. The system can reuse it to highlight all the occurrences of preterit or just present a list of all the preterit forms in the texts, instead of showing the whole text (cf. figure 2).



The Story of Goldilocks and the Three Bears
Once upon a time, there was a little girl named Goldilocks. She went for a walk in the forest. Pretty soon, she came upon a house. She knocked and, when no one answered, she walked right in. At the table in the kitchen, there were three bowls of porridge. Goldilocks was hungry. She tasted the porridge from the first bowl. "This porridge is too hot!" she explained. So, she tasted the porridge from the second bowl. "This porridge is too cold," she said. So, she tasted the last bowl of porridge.

The Story of Goldilocks and the Three Bears
tasted (3)
was (2)
answered
came
explained
knocked
said
walked
went
were

Figure 2 - Example of two different visualizations of the same text for the same facet

Conclusion

We have tried to explain, through a very simple example using very few tools, how the concepts of text facet, view and visualizations could help to acknowledge the influence of the pedagogical context on the pedagogical properties of the text. The pedagogical resource description standards while adapted to the description of already pedagogically exploited resources, do not seem fit to describe raw resources, in particular texts, in the context of their use in language teaching. The simplicity of the facets presented suggest that such a system could offer higher pedagogical added value, should other information or tools be made available to it. This is why we resorted to a very modular architecture, meant to regroup all the functions used in a treatment unit called the prism. Each facet is thus associated to a treatment sequence, using the functions grouped in the prism.

Combination of facets

This modularity is meant not only to be able to integrate to the prototype various sources of information, such as annotated corpora, or NLP tools, but also to reuse or combine existing facets. We believe that this architecture is evolvable enough to improve pedagogical indexation of text for language teaching through an iterative process and the collaboration between teachers, didactics experts and computer scientists. Starting with the two very simple facets we have introduced here (F_{RepEtC} and F_{WC}), one can already start to create more evolved facets. Our study showed that the kind of activity the teachers meant to use the text in, influenced the number of representative elements of the notion at the center of the activity and on the length of the text. Based on our results, we could create a facet asking the teacher what he or she wants to do with the text and what kind of structures he or she is interested in confronting the students with. The system would then translate this information into threshold values and tolerance for F_{RepEtC} and F_{WC}. Of course, this new facet would rely on declared and not actual practices and would thus probably not be very powerful. Still the prototype could help gather information on actual practices to make the values more accurate. In turn, additional parameters could be taken into account, such as the level of the students, which also influences the length of texts depending on the kind of activity. The integration of the new parameters can, in the same way, be confronted to the teachers practices via the prototype to be fine tuned and so on. To be effective, iterative process should feed off research in language didactics, NLP and the conclusions which can be drawn from the use of the prototype.

References

Antoniadis, G., Échinard, S., Kraif, O., Lebarbé, T., Loiseau, M. and Ponton,C. 2004 “Nlp-based scripting for call activities.” Paper presented at the *Coling Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning*, Genève, August, 2004.

Bawden, D. 2001. “Information and digital literacies; a review of concepts” *Journal of documentation* 57/2: 218-259.

Braun, S. 2005. "From pedagogically relevant corpora to authentic language learning contents” *ReCALL* 17/1: 47-64.

I EEE LTSC WG1 2. 2002. *Final 1484.12.1 lom draft standard document*.

http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf [Access date 15/02/2008]

Johns, T. 1991. “Should you be persuaded: two examples of data-driven learning” In *Classroom Concordancing*, T. Johns and P. King (eds) English Language Research Journal 4: 1-16.

Levy, M. 1997. *Computer-Assisted Language Learning, context and conceptualization*. Oxford: Oxford University Press.

Loiseau, M., Antoniadis, G. and Ponton, C. 2005. "Pedagogical text indexation and exploitation for language teaching " In *Recent research developments in learning technologies*, A. M. Vilas *et al.* (eds.). Badajoz: FORMATEX, 984-994.

Loiseau, M., Antoniadis, G. and Ponton, C. 2008. "Model for pedagogical indexation of texts for language teaching" Paper to be presented at the *3rd International Conference on Software and Data Technologies* (ICSOFT 2008), Porto, July 5 - 8 , 2008.

Sinclair, J. 1996. “Preliminary recommendations on corpus typology” *EAGLES (Expert Advisory on Language Engineering standards)* http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz [Access date 05/30/2008]

Taylor, D. 1994. " Inauthentic Authenticity or Authentic Inauthenticity?" *Teaching English as a Second or Foreign Language* ½: A-1.

USING LINGUISTIC CORPORA IN TEACHING CZECH GENITIVES AFTER QUANTIFIERS TO ENGLISH NATIVE SPEAKERS

Michaela Martinková[267]

Abstract

When learning Czech genitives after quantifiers, English speakers find it difficult to choose between a singular and a plural form. Kresin instructs English learners of Czech to resort to the category of countability: 'words denoting quantity are used with the genitive singular of mass nouns and the genitive plural of count nouns' (Kresin 2000: 245). Countability, however, is not a grammatical category of the Czech noun and it is not treated as such in grammar books and dictionaries. Besides transfers from English, which are often unreliable, students can profit from the information in the dictionary that a particular noun is 'pomnožné' (a plurale tantum), or 'zpravidla množné' (usually used in the plural). But they will have difficulties finding out whether a particular Czech noun is a singulare tantum, let alone in which of its meanings. A large corpus of Czech thus often remains the only resource to consult whether a particular noun is used after an indefinite expression of quantity in the singular or in the plural genitive form. In the Czech national corpus, or more specifically SYN2005, 'několik' (a few) is only used with nouns in the genitive plural and 'trochu' (a little) mostly with nouns in the genitive singular. If the plural is used after 'trochu', the noun in question usually has an uncountable English equivalent, which presents a problem to English learners of Czech. Czech equivalents of 'much' and 'many' are used with both singular and plural genitive forms, sometimes even of the same noun. Meaning shifts that accompany the transition of a singulare tantum to a noun expressing a regular singular – plural contrast are often language and culture specific, and are therefore important for the instruction of Czech as a foreign language.

Keywords: Czech, corpus, genitive, quantifier, teaching

Introduction

The English quantifiers *many* and *much* translate into Czech as *mnoho, hodně,* and *moc*; *a few* and *several* as *několik*; *a little* as *trochu*; and both *few* and *little* as *málo*. *More* is *víc(e)*, and both *fewer* and *less* are *méně* or *míň*. *How much, how many* are *kolik*, and *this much, that many tolik*. All these quantifiers, and cardinal numerals from five up, take nouns in the genitive. In this report I will consider the distribution of singular and plural genitives after Czech indefinite expressions of quantity, since the choice presents enormous difficulties to English learners of Czech. Using the Czech National Corpus (CNC) I will investigate Kresin's thesis that 'words denoting quantity are used with the genitive singular of mass nouns and the genitive plural of count nouns' (2000: 245), and that '*trochu* is only used with mass nouns and *několik* is only used with count nouns' (2000: 246). With *mnoho*, *hodně* and *moc*, I will focus on shifts of meaning if both singular and plural forms of the same noun are possible, and mention some cross-linguistic differences between English and Czech.

Genitives after quantifiers in textbooks of Czech: singular, or plural?

Genitives are introduced to learners of Czech at the lower intermediate level, after nominatives and accusatives: Rešková and Pintarová (1998) introduce first genitive singular and in the following lesson genitive plural. Holá presents genitive singular only, noting that 'when cooking, buying and counting things you will need to use the genitive plural' (Holá 2005: 119), which, according to her, happens 'after numerals from five to infinity and all the quantifiers' (Holá 2005: 119). The fact that most indefinite expressions of quantity take both singular and

[267] Michaela Martinková is a senior lecturer at the Department of English and American Studies, Philosophical Faculty, Palacký University, Olomouc, Czech Republic. She got her Ph.D. from Palacký University, after a ten-month research stay at the Department of Linguistics, UC Berkeley. Her dissertation was a corpus-based study of spatial adjectives. She is interested in functional aspects of language description and her primary areas of interest are semantics, history of English, contrastive grammar, and corpus linguistics. She has taught courses of Czech for foreigners.

plural genitives is not recognized as a problem in textbooks of Czech written by Czechs at any level of language instruction. English speakers of Czech at the lower intermediate level, however, do wonder why they should say *hodně švestek* gen.pl. (many plums), but *hodně slivovice* gen.sg. (much plum brandy). Kresin (2000: 245) offers a rule which for speakers of English is immediately at hand: 'words denoting quantity are used with the genitive singular of mass nouns and the genitive plural of count nouns'. The problem with this rule is, however, that countability is not a grammatical category of the Czech noun and it is not treated as such in Czech grammar books or dictionaries. It is reflected within the category of number. The normative grammar of Czech *Mluvnice češtiny* (1986: 45) mentions uncountable nouns as those that do not enter the singular – plural opposition. In other words, uncountable nouns are not only those having only singular forms (singularia tantum) but also those having only plural forms (pluralia tantum). Rules about English cannot be directly applied to Czech either, since countability is a matter of conceptualization (Cruse 2000: 270) and crosslinguistic differences abound. Even Kresin admits that 'Czech and English sometimes differ in what each treats as a "count" or "mass" noun' (Kresin 2000: 245). She mentions, quite rightly for the level of intermediates, examples from the category of food: *těstoviny* (pl) vs *pasta*, *zelenina* (sg) – *vegetables*, *mrkev* (sg) – *carrots*, noticing at the same time that 'in Czech *money* is always in the plural (*peníze, mnoho peněz*)' (Kresin 2000: 246).

Genitives after quantifiers in the Czech national corpus: singular, or plural?

Of the corpora of the CNC available, I worked with SYN2005, a synchronic and arguably 'representative corpus of contemporary written Czech containing 100 million words' (CNC).[268] It is well annotated and so it allows searches not only by parts of speech, but even by more specific categories such as case, number and gender. Unfortunately, spoken corpora of the CNC, such as ORAL2000, PMK (corpus of Czech spoken in Prague) and BMK (corpus of Czech spoken in Brno), much more modest in size (1 million, 675,000, and 490,000 words respectively), are not annotated and thus do not allow a systematic search for genitive forms.

The case of 'několik' and 'trochu'

Několik (a few, several) and *trochu* (a little, some) are the only quantifiers which seem to select nouns according to countability. Kresin (2000: 246) claims that '*trochu* is only used with mass nouns and *několik* is only used with count nouns', and Komárek (2006: 62) argues that *několik* only expresses number, not amount. It is thus expected that *několik* will be used with nouns in the genitive plural and *trochu* with nouns in the genitive singular. This is partly disproved by my findings. It works with *několik*, not so much with *trochu*.

	noun in genitive singular		noun in genitive plural	
		one ADJ intervening		one ADJ intervening
trochu	2,706	512	219	39
několik	0	0	31,634	6,906

Table 1: Statistics of nouns in genitive singular and plural after *několik* and *trochu* in SYN2005

Několik is indeed used in SYN2005 only with nouns in genitive plural (all tokens with the singular were mistakes in spelling and annotation and had to be manually discarded). Kresin's thesis that it 'is only used with count nouns' (2000: 246) was confirmed as well. *Několik* can thus be considered a direct equivalent of *a few* and *several*.

But *trochu*, though used predominantly with the singular, does admit genitive plurals. Still, at the lower intermediate level, saying that *trochu* combines with nouns in genitive singular is perhaps appropriate. As expected, in most cases the nouns are singularia tantum. If nouns normally counted are used after *trochu,* there is a shift of meaning from individual to mass interpretation. Most of them are related to food: *trochu kachny* (a little bit of duck), *kuřete* (chicken), *candáta* (pikeperch), *banánu* (banana), but there are also others: *trochu rtěnky* (a

[268] http://ucnk.ff.cuni.cz/english/index.html [Access date 2/5/2008]

little bit of lipstick), *klarinetu* (clarinet music). Similar shifts of meaning are found in English and so they do not present a problem to a speaker of English.

As stated, the real problem for a non-native speaker of Czech is that Czech dictionaries do not say that a certain noun is only used in the singular, and in which of its meanings. Grammar books will not help either. *Mluvnice češtiny* (1986: 49) admits that the singularia tantum class is semantically fuzzy, mentioning only four types of nouns: 'proper nouns', 'some abstract nouns', 'mass nouns', and 'collective nouns'. There is thus not much a non-native speaker of Czech can rely on. Besides transfers from English, looking up a particular phrase as a concrete corpus query is often the only solution. Luckily, among words quantified by *trochu* most frequently, English and Czech are in accord, i.e. all English equivalents are uncountable nouns. The following table presents nouns that are quantified by *trochu* at least twenty times:

noun in genitive singular	English equivalent	number of tokens
vody	water	224
času	time	200
vína	wine	94
čaje	tea	85
mléka	milk	77
štěstí	luck	76
kávy	coffee	74
soli	salt	69
světla	light	56
vzduchu	air	47
jídla	food	45
klidu	peace	42
oleje	oil	42
krve	blood	39
mouky	flour	39
šťávy	juice	39
cukru	sugar	37
vývaru	broth	33
místa	space	32
rozumu	reason	29
octa	vinegar	27
pozornosti	attention	26
sněhu	snow	26
másla	butter	25
whisky	whisky	25
piva	beer	24

zábavy	fun	24
barvy	colour	22
lásky	love	20
polévky	soup	20
trpělivosti	patience	20

Table 2: The most frequent nouns in genitive singular following *trochu* in SYN2005

Among the less frequent nouns, however, the English equivalent of a Czech singulare tantum is not always a noun in the singular, typically from the category of food: *cukroví* (sweets), *čočka* (lentils), *hrách* (peas), *koření* (spices), *mrkev* (carrots), *pití* (drinks), *strouhanka* (breadcrumbs), *ovoce* (fruit(s)), *zelenina* (vegetables). The Czech noun *paprika* has a regular plural if it denotes a kind of vegetables (*400 g zelených paprik* – 400 grams of green peppers) but it is a singulare tantum if it denotes a spice (*trochu papriky* – a little red pepper). Also the noun *cibule* (onion) is often used in the singular where English has the plural form (*Musím koupit cibuli* (acc.sg.). – I have to buy onions.)

The fact that plural genitives do occur after *trochu* should be taken into account at a higher level of language instruction. In SYN2005, there are 219 tokens of *trochu* immediately followed by a plural genitive noun, and 39 tokens with an intervening adjective. Almost a half of these (102), however, is represented by the phrase *trochu peněz* (a little money). *Peníze*, a plurale tantum, cannot be counted. The noun is not commonly quantifiable by *několik* and cardinal numerals. Luckily, this time, a learner of Czech will find the information that a particular noun is a plurale tantum in the dictionary – the noun will be glossed as 'pomnožné'.

Many other plural nouns following *trochu* are Czech pluralia tantum with uncountable English equivalents. Some are other expressions denoting money (*trochu drobných* – small change), but typically they are from the category of food: *těstoviny* (pasta), *makarony* (macaroni), *povidla* (plum jam), *kvasnice* (yeast).

Interestingly, some plural genitive forms following *trochu* are those that according to *Mluvnice češtiny* (1986: 50) are used predominantly in the plural, i.e. the plural, not the singular, is the unmarked form. Many of these are in the dictionary glossed as 'zpravidla množné' (usually plural), unfortunately, by no means all of them. Arguably, these nouns denote objects which, in the default case, occur in large quantities, or objects composed of many parts: *trochu vlasů* (a little hair),[269] *trochu odpadků* (a little trash). If they denote abstract entities, these entities again either occur in large quantities, or recur repeatedly: *trochu informací* (a little information), *zkušeností* (a little experience). These nouns, again, are all uncountable in English and thus present a problem for English learners of Czech.

There are quite a few nouns of this type, however, whose English equivalents are nouns in the plural, often from the category of food: *trochu potravin* (foodstuffs), *keksů* (cookies), *starých sušenek* (old biscuits), *osmažených, vařených brambor* (fried, boiled potatoes), *brambůrků* (chips), *nakrájených mandlí* (chopped almonds), *rybích jiker* (eggs of fish), *studených fazolí* (cold beans), *čerstvých, dušených, sušených hub* (fresh, stewed, dried mushrooms), *žampionů* (champignons), *sušených, nasekaných bylinek* (dry, chopped herbs), *brusinek* (cranberries), *jahod* (strawberries), *ostružin* (blackberries), *rozinek* (raisins), and also *jablek* (apples), *vajec* (eggs), and *vajíček* (eggs dim.).

The case of 'mnoho', 'hodně', and *'moc'*

Mnoho, hodně and *moc* are equivalents of the English quantifiers *many* and *much*. As they all take nouns in both singular and plural genitive forms, the choice between them has nothing to do with the countability of the following noun. Concrete numbers are presented in the following table. [270]

[269] Interestingly, *grey hair* is a plurale tantum in Czech, i.e. the singular form does not even exist: *trochu šedin* (a little grey hair).

[270] Only tokens with a noun immediately following the quantifier were retrieved here. The reason is that *moc* and *hodně* also function as intensifiers of adjectives, i.e. they are not related to the noun in the genitive, and a lot of manual sorting would be ne cessary. Tags will not help, because the quantifiers in question are not tagged systematically in the CNC.

	nouns in genitive singular	nouns in genitive plural
mnoho	1,958	10,132
hodně	1,379	2,177
moc	1,137	1,026

Table 3: Statistics of nouns in genitive singular and plural after *mnoho*, *moc*, and *hodně* in SYN2005

Since *mnoho, hodně* and *moc* take nouns in both genitive forms, the same noun often occurs in both genitive forms. This is not surprising since, as *Mluvnice češtiny* (1986: 46) argues, the borderline between nouns expressing contrast between the singular and the plural, and those always used in one number form, is gappy. It is thus formally possible to form a potential plural for most of the Czech nouns. If this potential is made use of, such transitions are often accompanied by a change in meaning. As Cruse argues, 'mass nouns used as count nouns are usually to be interpreted in one or two ways, either as unit quantities of the continuous mass, or as different types or varieties' (Cruse 2000: 271).

i. *mass → a unit of*

The only institutionalized unit of a consumable liquid seems to be half a litre for beer: *hodně piva* (much beer) – *hodně piv* (many beers). But not only consumable liquids have conventionally agreed units: *hodně papíru* (much paper) – *hodně papírů* (many sheets of paper). Other examples include particular instances of abstract notions, mostly associated with human existence or activities: *moc života* (much life) – *hodně životů* (many lives), *hodně světla* (much light) – *hodně světel* (many lights), *hodně stínu* (much shade) – *moc stínů* (many shadows), *hodně místa* (much space) – *hodně míst* (many places), *hodně reklamy* (much advertising) – *hodně reklam* (many ads), *hodně práce* (much work) – *hodně prací* (many works), *mnoho krásy* (much beauty) – *mnoho krás* (many beautiful things), *mnoho lásky* (much love) – *mnoho lásek* (many loves), *mnoho stresu* (much stress) – *mnoho stresů* (many stressful situations), *moc příležitosti* (much opportunity) – *moc příležitostí* (many opportunities).

ii. *mass → a kind of*

If the noun *víno* (wine) occurs with an indefinite expression of quantity in genitive plural, it never means a unit, but always a variety. This is perhaps because for *víno* there is no conventionally agreed unit: *moc vína* (much wine) – *mnoho vín* (many wines). Other transitions include: *hodně vitamínu C* (much vitamin C) – *hodně vitamínů* (many vitamins), *mnoho cukru* (much sugar) – *mnoho cukrů* (many sugars), *hodně tuku* (much fat) – *hodně tuků* (many fats), *hodně barvy* (much colour) – *hodně barev* (many colours), *hodně tekutiny* (much liquid) – *hodně tekutin* (many liquids), *hodně jídla* (much food) – *hodně jídel* (many meals), *hodně materiálu* (much material) – *hodně materiálů* (many materials), *hodně sportu* (much sport) – *moc sportů* (many sports), *hodně disciplíny* (much discipline) – *hodně disciplín* (many disciplines), *moc zájmu* (much interest) – *hodně zájmů* (many interests), *mnoho pravdy* (much truth) – *mnoho pravd* (many truths).

Some of the tokens, however, do not fit either of the two categories. What all of them share is the fact that the English equivalent of the Czech phrase with the genitive plural is quite hard to find. I suggest that the difference between the genitive singular and the genitive plural is one of intensity, which corresponds to a specific usage of the plural form: *hodně síly* (much power) – *hodně sil* (great pains), *hodně škody* (much damage) – *hodně škod* (great damage), *hodně úspěchu* (much success) – *hodně úspěchů* (much success), *moc naděje* (much hope) – *moc nadějí* (great hopes). The noun *smysl* (sense) is a singulare tantum, but its opposite *nesmysl* (nonsense) readily occurs in the plural (*hodně nesmyslů*).

The most radical meaning shift of all is *hodně prachu* (much dust) – *hodně prachů* (much dough).

As to the choice between *mnoho*, *moc* and *hodně*, if countability of the noun that follows is not what influences it, English learners of Czech ask what it actually is. In SYN2005 *moc* is the least frequent of the three, and *mnoho* is by far the most. This is counter-intuitive for a native speaker of Czech. As early as 1957 Poldauf (1957: 73) glosses *mnoho* as 'knižní' (literary) and while some textbooks of Czech still present it as a direct equivalent of *many* and *much* (Kresin 2000), in others (Holá 2005) it is not even mentioned. To a Czech native speaker, *mnoho,* especially when used with nouns in the genitive singular, sounds very formal (even in SYN2005, *mnoho* is five times less frequent with the genitive singular than with the genitive plural) and in my classes of Czech I instruct my students to avoid it. *Moc*, on the other hand, is in the dictionary *Slovník spisovné češtiny pro školu a veřejnost* glossed as 'hovorové' (colloquial).[271] The difference thus seems to be by and large a stylistic one and since SYN2005 contains only written texts, *mnoho* predominates there.

Conclusions

Countability is not a grammatical category of the Czech noun. Nouns are not glossed as countable or uncountable in Czech dictionaries and no respective information is to be found in grammars either. An English learner of Czech can thus only profit from the fact that a plurale tantum is glossed as 'pomnožné' in the dictionary. With singularia tantum, however, English speakers can only rely on transfers from English. Since these are not always reliable, what is left to the learner is to resort to a corpus of Czech to look up a particular phrase in question. *Několik* is only used with nouns in genitive plural, *trochu* mostly with nouns in genitive singular. If the plural occurs, it is very often because the noun is a plurale tantum, or a noun which typically occurs in the plural. These nouns often have uncountable English equivalents. *Mnoho, hodně* and *moc* are all used with both singular and plural genitive forms, sometimes even of the same noun. Concrete examples of singularization are sometimes accompanied by language and culture dependent changes of meaning. The choice between *mnoho, hodně* and *moc* does not depend on the semantics of the following noun, but is largely a stylistic one. *Mnoho* is five times more frequent in the plural than in the singular, but seems to be limited to written texts and learners of Czech should perhaps be discouraged from using it.

The lack of reliable information in relevant literature means that Czech teachers of Czech need a large corpus of Czech to verify their intuitions. The corpus supplies minimal pairs of singular and plural genitive forms which one would otherwise not notice, or concrete numbers proving a tendency of some quantifiers to prefer certain nouns or nouns in certain number forms.

References

Cruse, A. 2000. *Meaning in Language. An Introduction to Semantics and Pragmatics.* Oxford: Oxford University Press.

Czech National Corpus – SYN2005. Institute of the Czech National Corpus, Prague 2005. Accessible at WWW: <http://ucnk.ff.cuni.cz>.

Holá, L. 2005. *New Czech Step by Step*. Praha: Akropolis.

Komárek, M. 2006. *Příspěvky k české morfologii.* Olomouc: Periplum.

Kresin, S., Kořánová, I., Subak-Kaspar, H. and **Kašpar**, **F**. 2000. *Čeština hrou. Czech for fun*. New York: McGraw-Hill Primis Custom Publishing.

Mluvnice češtiny (2). Tvarosloví. 1986. Praha: Academia.

Poldauf, **I.** 1957. "Vyjadřování kvantity v češtině." *Slovo a slovesnost* 18: 71-85.

Rešková, I. and **Pintarová, M**. 1995. *Communicative Czech (Elementary Czech).* Praha: Ústav jazykové a odborné přípravy Univerzity Karlovy.

Slovník spisovné češtiny pro školu a veřejnost. 2007. Praha: Academia.

[271] Moc also seems to be limited to negative contexts, a valuable observation supplied by the principle of serendipity during this research.

HOW TO DEFINE AN ENRICHED CORPUS FROM A PEDAGOGICAL PERSPECTIVE? THE MEDI@TIC CASE

Maribel Montero Perez[272]
Hans Paulussen[273]
Nathalie Faidherbe[274]
Piet Desmet[275]

Abstract

The project described in this paper is Medi@tic, a web-based database for authentic high quality video-materials, available for teaching and learning of French and Dutch. The aim of the Medi@tic project is to develop an enriched corpus of learning objects, especially conceived for foreign language learning and teaching. The pedagogical enrichment of the project can be situated at different levels, each contributing to a specific pedagogical approach of the learning objects. The enrichment of the Medi@tic case is composed of three steps, namely the intake and selection of high quality data, the way of processing the material and the metadata that are at the basis of the web interface.

Medi@tic consists currently of French and Dutch learning objects, especially video materials, and plans to include also documents in other languages. Next to the multilingual character of Medi@tic, the project aims to create a multimodal repository with not only video, but also audio and written authentic learning objects. All videos integrated in the corpus can be viewed with or without the corresponding transcript. This allows users to develop reading and listening comprehension skills at the same time. The transcripts are also at the basis of another enrichment process in which these texts will be annotated. This will allow foreign language teachers to search texts or video fragments in function of words or specific linguistic structure. All Medi@tic learning objects have corresponding metadata in order to facilitate a search on the web interface, especially designed in function of the metadata criteria.

Thanks to all learning objects integrated up to now, and the future developments we have in store, Medi@tic will become a multilingual and multimodal enriched corpus for language learning and teaching.

Keywords: Enrichment, learning objects, transcript, metadata, annotation

Introduction

In the domain of corpora, the concept of enrichment can cover different elements, each contributing to a better pedagogical approach of the corpus material. In this paper, we focus on the Medi@tic project, a web-based

[272] Maribel Montero Perez studied Romance languages and literature, French and Spanish, at the K.U.Leuven Campus Kortrijk and K.U.Leuven (2003-2007). She graduated in july 2007 with a MA thesis in the domain of French linguistics. Since September 2007, she works at the K.U.Leuven Campus Kortrijk. She worked for the Lingu@tic project and is presently involved in the DPC-project (Dutch Parallel Corpus), a multilingual annotated corpus with Dutch as a pivotal language. She is involved in the compilation and the linguistic angle of the Dutch and French data that will be integrated in the corpus

[273] Hans Paulussen is a senior researcher at the University of Leuven, Campus Kortrijk. He has been involved in a number of computational linguistic projects on CALL, corpus compilation and tagging. He has many years of experience in foreign language teaching (Dutch, English and French) and research in computational linguistics. He wrote a PhD thesis on the contrastive analysis of prepositions and particles in English, French and Dutch within the cognitive linguistic framework. Part of this project consisted in building the Namur Corpus: a trilingual parallel corpus of English, French and Dutch fiction and non-fiction. He is presently involved in the compilation of the Dutch Parallel Corpus (DPC), a multilingual corpus which will be useful for research in language technology, linguistics, translation studies and language teaching.

[274] After a degree in English literature, Nathalie Faidherbe graduated, in 2006, from the University of Lille 3, France, with a MA in the domain of language didactics, especially French as a foreign language, and the use of multimedia for language learning/teaching. She then developed modules of French on a multimedia learning environment for foreigners studying in French universities. Since March 2007 Nathalie has been working at the K.U.Leuven Campus Kortrijk, on the Franco-Belgian Lingu@tic project as an educational content developer, for the electronic learning environment Franel, and for the Mediatic database which aims at providing teachers of French and Dutch with free authentic material to be used in the classroom.

[275] Piet Desmet is full professor of French and Applied linguistics and Foreign Language Methodology at the K.U.Leuven (Belgium). His research mainly focuses on French and Applied Linguistics, with a particular interest in Computer Assisted Language Learning. He coordinates different CALL research projects, is coordinator of /*ALT, Research Centre on CALL* <http://www.kulak.ac.be/ALT/>/ as well as of the Research Centre /ITEC, *Interdisciplinary research on Technology, Education & Communication*/.

database of authentic video-materials especially conceived for the teaching and learning of foreign languages (Desmet 2006b).

The first section of this paper gives a general presentation of Medi@tic, explaining its origin and development. The concept of enrichment is the key-word of the second section, in which we describe all functionalities of the project and the future developments we have in store in order to enrich and improve the quality of the present corpus.

General presentation

Medi@tic is a product of the Lingu@tic project. The essence of this Franco-Belgian project is to allow people living in the North of France (Nord - Pas-de-Calais) and those living in the Dutch-speaking part of Belgium (West-Flanders) to get to know each other's languages and cultures. Therefore, the Lingu@tic project aims at providing language teachers and learners of both regions with educational tools and material in order to help them reach that goal.

The first tool that was developed by our research team was Franel, an electronic language learning environment for Dutch and French as foreign languages (Desmet 2006a). Franel is based on the educational use of authentic audiovisual material (Desmet 2005). Franel is therefore a learning tool. In addition to Franel, our team is also developing a teaching tool called Medi@tic, which is the focus of this paper.

As a matter of fact, Medi@tic is an on-line database containing authentic audiovisual material meant to be used by teachers of Dutch and French as foreign languages on both sides of the Franco-Belgian border. The videos available within the Medi@tic database – and also those used in the Franel environment – are provided by Belgian and French regional television channels, namely WTV, Notele and C9, which guarantees the authenticity and quality of the videos. Those three channels are partners in the Lingu@tic project which means that the videos can be used freely, regardless of copyright.

Medi@tic as an enriched corpus from a technical and pedagogical point of view

In the following section, Medi@tic will be described in three steps. In a first point, we describe the intake and the selection of the data that will be integrated in the corpus. Subsequently, we detail the enrichment that will be added to the written texts and the transcripts of the audio-visual learning objects. In a last point, we insist on the presence of metadata in order to create a pedagogically well-designed web interface. These three levels emphasize both the technical and the pedagogical objectives of our project and make them more concrete.

Data selection

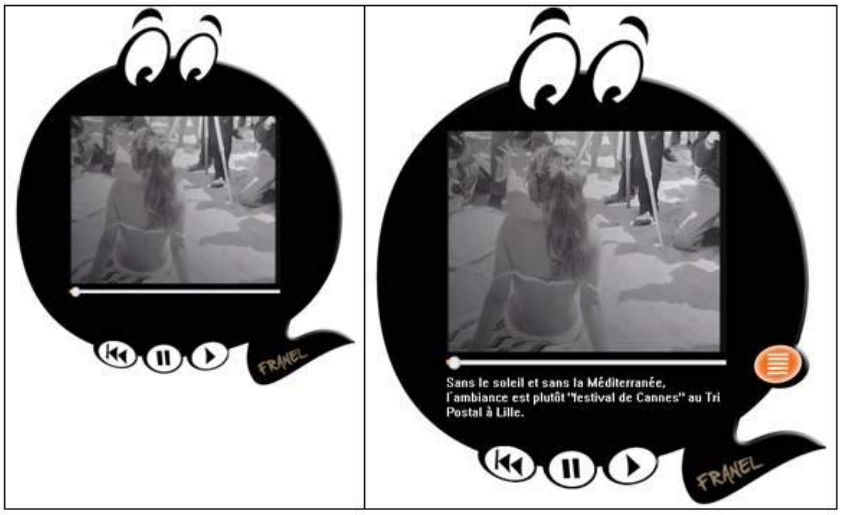
The selection of high quality material is a first important action for the pedagogical enrichment of the project. Up to now, Medi@tic offers authentic video material from regional broadcasting houses. The video material not only stimulates the listening comprehension of the students, the authenticity of the provided material also motivates the language learners. Nevertheless, language teaching covers different competences. It is therefore crucial to incorporate not only video material as is. Indeed, our team is currently working on the integration of audio material and written documents as well. This will be made possible thanks to an agreement with one of the prominent publishing firms in Belgium, specialised in the edition of magazines in French and in Dutch. This new partnership allows us to offer a much wider range of authentic material to the users of Medi@tic in the near future.

The written monolingual Dutch and French journalistic texts will be provided in XML-files. However, since authentic lay-out is crucial for the pedagogical exploitation of the textual material, all texts will be presented as a rich pdf-version with images and lay-out. Still, XML-files are necessary in order to search the texts for words or linguistic structures that are interesting for teachers and language learners eager to gain more in-depth knowledge. For example, teachers can search articles, on the basis of a keyword of the video fragment, in order to develop the same theme as presented in the video.

As shown in the preceding paragraphs, Medi@tic will become a multimodal database of learning objects by providing not only video material, but also audio fragments and textual material from high quality journalistic magazines. Next to the multimodal aspect, we want to insist on the plan to provide not only Dutch and French learning objects, but also German and English material. From a pedagogical point of view, this means that Medi@tic can be integrated in all language courses. In Flanders for example, English, French and German are taught as foreign languages. Nevertheless, it has to be stressed that language learners do not have the same competences in these three foreign languages. That is why Medi@tic will integrate fragments of different levels also because the knowledge of a language evolves throughout the process of language learning.

Exploitation of transcripts

As presented in the first section, especially video materials have been integrated up to now and can be viewed thanks to the Medi@tic web interface. After a (extensive) search on the web interface (cf. Infra), the user can view a video by clicking on the appropriate icon. Two video players are available: either with or without transcription. In this part, more detailed information on the enrichment of the transcripts and their exploitation is given.



Sans le soleil et sans la Méditerranée, l'ambiance est plutôt "festival de Cannes" au Tri Postal à Lille.

Figure 1: The video players without and with transcript

The availability of the transcript presents many advantages. The possibility of activating a transcript can help language learners to develop their reading and listening skills at the same time. Moreover, the transcripts help learners with limited comprehension skills to improve their competences in a foreign language.

From a technical point of view, different elements concerning the transcripts will be detailed below. In this section, we stress the fact that not only the data itself, but especially the technological processing of the data creates an enormous pedagogical added value in this second enrichment process.

The video transcripts are stored in XML files compatible with the MAGPie DTD, a validation schema developed for aligning captions to the video source, based on time offsets used as reference point. The MAGPie captioning editor[276] makes editing the alignment of video transcripts an easy task. Next to creating captioning files readily available for online video players, such as QuickTime, Windows Media, and Real, MAGPie also creates the XML file from the aligned data. Although the XML file is initially used only for the alignment of video extracts, it is an important data file which can be explored further for other applications which make life easier for language learners.

Thanks to a validated XML format, the transcription files can easily be transformed into other file formats, including selection filtering. A simple XSLT filter can help to render the XML file in a more legible format, more suitable for the language learner. Such an XSLT filter is capable of rendering XML into HTML or PDF. One could, for example, use coloring to differentiate the different speakers in the dialogue, and the voice over. Rendition can be done for the whole XML file, or for specific selections within the XML file. In the case of XML files containing both the transcription and translation of a video extract, one could, for example, select only the transcription part, or select only the translation part for certain words found in the transcription part.

[276] MAGPie (Media Access Generator) has been developed by the Carl and Ruth Shapiro Family National Center for Accessible Media (NCAM) . See: http://ncam.wgbh.org/webaccess/magpie/

A possible application consists in creating a lexicon on the basis of the words found in the transcripts. In order to build for example a French lexicon, an XSLT script can filter first of all the French captions, each selected caption line preceded by its time index or offset. The output of such a filter is shown in figure 1. A second script can then transform the output into a concordance file, where each line starts with a word, followed by all time offsets of the transcription lines containing that word. The table below shows such an output, where time references are separated by a pipe symbol. This example is only a basic selection filter, which requires further processing, including filtering stop words, tagging and lemmatizing the words. This simple example nevertheless shows that a validated XML file can be the source for further automatic processing of the data.

00:00:04.0000	La cuisine du Nord ne se prend pas le chou sauf pour les recettes,
00:00:06.6400	comme cette entrée festive, l'aumônière de Saint-Jacques.
00:00:09.6800	Plus qu'une introduction à un réveillon aux saveurs d'ici, c'est une protestation :
00:00:14.1800	la grande cuisine du Nord veut exister et se faire reconnaître.
00:00:18.0300	« On n'est pas assez reconnu, je trouve, dans le nord, on va toujours dans le sud, comment dire, chercher les talents.
00:00:24.8900	On va nous dire qu'il y a beaucoup plus de talents dans le sud alors que c'est pas vrai,
00:00:28.8400	dans le Nord on a des talents.
00:00:31.9000	On ne va pas assez fouiller dans le nord en fait pour chercher les bonnes tables,
00:00:35.6500	on s'arrête toujours dans le bas de la France, quoi ».
00:00:38.3000	Régionalisme au fourneau, peut-être un peu, curiosité aussi,
00:00:42.0800	celle de composer et de proposer de nouvelles saveurs par le travail de produits régionaux.
00:00:47.2800	« Je pratique les Saints-Jacques dans mon restaurant, d'une façon différente
00:00:51.1200	mais je prends les Saints-Jacques qui viennent du Nord-Pas de Calais, quoi ».

Table 1 : Video sequence aligned transcriptions

A	00:00:09.6800
A	00:00:24.8900|00:00:28.8400|00:00:59.9000|00:01:12.7000
Alors	00:00:24.8900
assez	00:00:18.0300|00:00:31.9000
Au	00:00:38.3000
aussi	00:00:38.3000
Aux	00:00:09.6800|00:01:05.8100
Bas	00:00:35.6500
beaucoup	00:00:24.8900|00:00:55.0000|00:00:59.9000
bonnes	00:00:31.9000
calais	00:00:51.1200
Ce	00:01:05.8100
Celle	00:00:42.0800

c’est	00:00:09.6800|00:00:24.8900
cette	00:00:06.6400|00:01:05.8100|00:01:12.7000
chefs	00:01:16.0000
chercher	00:00:18.0300|00:00:31.9000
Chou	00:00:04.0000
choux	00:01:05.8100
Club	00:01:16.0000
comme	00:00:06.6400
comment	00:00:18.0300
composer	00:00:42.0800
cuisine	00:00:04.0000|00:00:14.1800|00:01:12.7000
curiosité	00:00:38.3000
Dans	00:00:18.0300|00:00:18.0300|00:00:24.8900|00:00:28.8400|00:00:31.9000|00:00:35.6500|00:00:47.2800|00:01:16.0000|00:01:19.9500

Table 2: Concordance based on time offsets

The lexicon built can be used to automatically select video extracts and evaluate the level of difficulty of the text. Instead of showing the video film as a whole, one could start a video extract from the time offset indicated in the word index list. This requires a video player which can be programmed to start playing at an offset which lies beyond the usual starting point of the video. One should also consider defining a context frame to start the video a little before the actual offset. Once the lexicon is built, one can also analyse texts and measure the lexical density of the text, so that texts can be graded according to their lexical complexity.

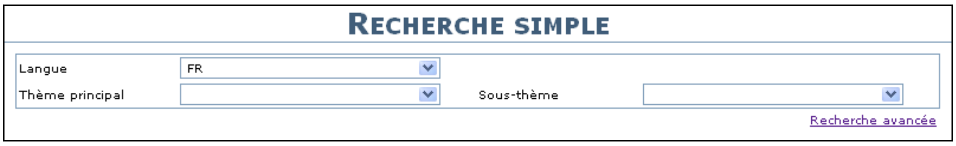
The words found in the transcription files can also be used to help the annotators classify the texts. Up till now, annotators skim through the video extracts and the aligned transcription, in order to identify the topic of the video extract. However, the lexicon created from the transcript file can help the annotator in choosing the right topic. By disregarding stop words and low frequency words, and presenting only potential topic words, the annotator no longer has to go through the whole video. The possible topic words are situated in the f2 word set: the set of words which follow the highest frequency words (f1), and which have a higher frequency in the text, than expected in a general text.

As soon as one has classified a considerable number of texts in the MEDIATIC database, these texts can also be used as reference set for the classification of new texts. Suppose the set of f2 words in a new text is very similar to the f2 lists of a number of texts already classified, there is a high probability that the new text will be of similar text type.

Although data stored in XML format makes further processing an easy task as far as computer processing is concerned, a number of ambiguities may require extra processing. A case in point is the structure of aligned transcriptions. Initially, the caption files focus on the alignment of video extracts and the text spoken. However, the original alignment disregards sentence boundaries. Now and then, a sentence starts in one sequence, and continues in another. Similar problems involve the fact that two speakers speak in the same sequence. In order to analyse linguistically coherent text chunks, the sentences will have to be annotated correctly. This kind of linguistic annotation was not foreseen in the initial alignment files.

The metadata and the Web Interface

The last step in the enrichment process has to be situated at the level of the web interface. Medi@tic works as a search engine which allows the user to explore a selection of documents available in the Medi@tic database, based on a certain number of criteria. In the “basic” search window, the user selects a language (Dutch or French) and a theme, such as economics, business, health, daily life, the environment and more. A “sub-theme” can also be selected in order to filter the results.



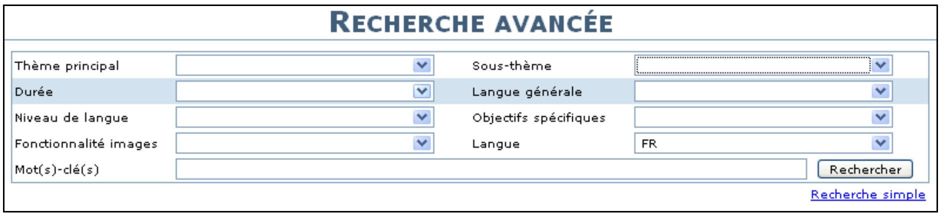
RECHERCHE SIMPLE

Langue	FR		
Thème principal		Sous-thème	

Recherche avancée

Figure 2: The “basic” search window

Thanks to the “advanced” search window, the user may enter a greater number of search criteria such as the length of the video clip, the language level (i.e. as described in the *Common European Framework of Reference*), the functionality of the images (i.e. contribution to comprehension) or the type of language (i.e. general or for specific purposes, e.g. business). The user can also search for a document by entering one or several keywords.



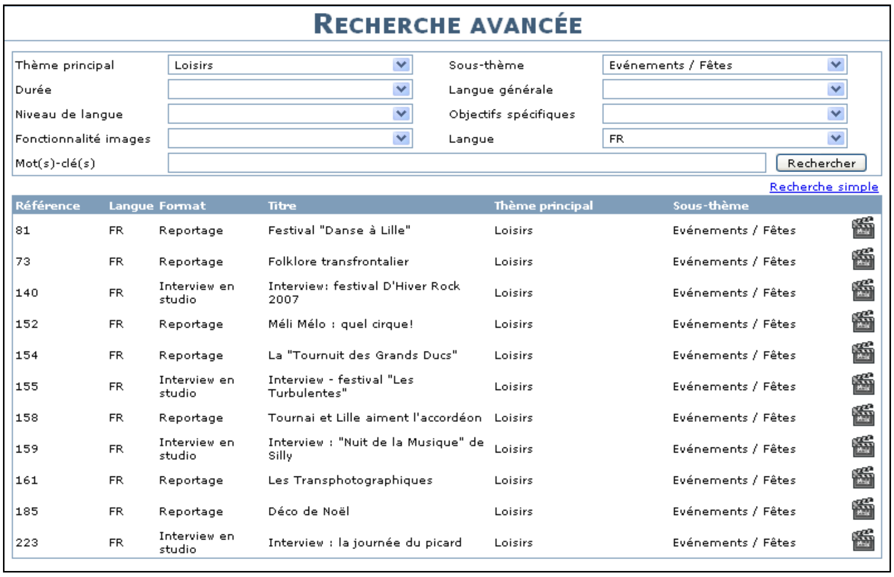
RECHERCHE AVANCÉE

Thème principal		Sous-thème	
Durée		Langue générale	
Niveau de langue		Objectifs spécifiques	
Fonctionnalité images		Langue	FR
Mot(s)-clé(s)		Rechercher	

Recherche simple

Figure 3: The “advanced” search window

The list of results is then displayed, showing, for each result, the title and the type of document (e.g. for videos: news coverage or studio interview). The video icons on the right lead directly to a preview of the document.



RECHERCHE AVANCÉE

Thème principal	Loisirs	Sous-thème	Evénements / Fêtes
Durée		Langue générale	
Niveau de langue		Objectifs spécifiques	
Fonctionnalité images		Langue	FR
Mot(s)-clé(s)		Rechercher	

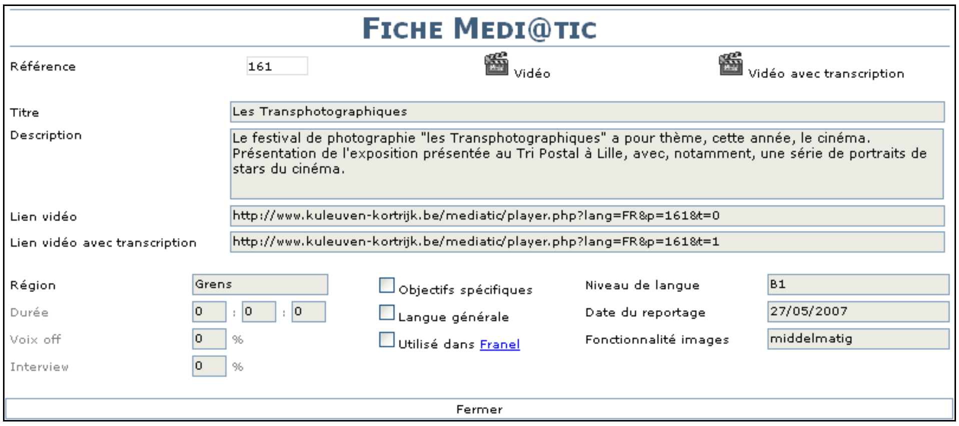
Recherche simple

Référence	Langue	Format	Titre	Thème principal	Sous-thème
81	FR	Reportage	Festival "Danse à Lille"	Loisirs	Evénements / Fêtes
73	FR	Reportage	Folklore transfrontalier	Loisirs	Evénements / Fêtes
140	FR	Interview en studio	Interview: festival D'Hiver Rock 2007	Loisirs	Evénements / Fêtes
152	FR	Reportage	Méli Mélo : quel cirque!	Loisirs	Evénements / Fêtes
154	FR	Reportage	La "Tournuit des Grands Ducs"	Loisirs	Evénements / Fêtes
155	FR	Interview en studio	Interview - festival "Les Turbulentes"	Loisirs	Evénements / Fêtes
158	FR	Reportage	Tournai et Lille aiment l'accordéon	Loisirs	Evénements / Fêtes
159	FR	Interview en studio	Interview : "Nuit de la Musique" de Silly	Loisirs	Evénements / Fêtes
161	FR	Reportage	Les Transphotographiques	Loisirs	Evénements / Fêtes
185	FR	Reportage	Déco de Noël	Loisirs	Evénements / Fêtes
223	FR	Interview en studio	Interview : la journée du picard	Loisirs	Evénements / Fêtes

Figure 4: The display of results

For each learning object, a Medi@tic record has been conceived with all corresponding metadata (cf. Infra). The links available in the “Medi@tic record” make it easier to integrate the chosen document(s) in other CALL (Computer Assisted Language Learning) applications. For example, in order to use the document in the classroom, the teacher can copy the appropriate link displayed in the “Medi@tic record” and paste it on a webpage or any other document. The student or any other user will then be able to directly open the video player by simply clicking on the link.

In the Medi@tic records, users can also find specific metadata concerning the document they have selected: a short description of what the document is about, the date of broadcasting, the availability of the transcription of the document, the voice over/speaker ratio.



FICHE MEDI@TIC

Référence 161 — Vidéo — Vidéo avec transcription

Titre: Les Transphotographiques

Description: Le festival de photographie "les Transphotographiques" a pour thème, cette année, le cinéma. Présentation de l'exposition présentée au Tri Postal à Lille, avec, notamment, une série de portraits de stars du cinéma.

Lien vidéo: http://www.kuleuven-kortrijk.be/mediatic/player.php?lang=FR&p=161&t=0

Lien vidéo avec transcription: http://www.kuleuven-kortrijk.be/mediatic/player.php?lang=FR&p=161&t=1

Région: Grens

Durée: 0 : 0 : 0

Voix off: 0 %

Interview: 0 %

□ Objectifs spécifiques

□ Langue générale

□ Utilisé dans Franel

Niveau de langue: B1

Date du reportage: 27/05/2007

Fonctionnalité images: middelmatig

Figure 5: The “Medi@tic record”

From the preceding paragraphs, it results that the metadata are crucial for the Medi@tic web interface. The metadata also allow us to refer to another project of our research team, namely the DPC-project (Dutch Parallel Corpus). The project aims to create an annotated parallel corpus (Dutch, English and French) within the framework of the STEVIN[277] programme of the Dutch Language Union (Paulussen 2006). STEVIN is a long term programme aimed at stimulating the development of language and speech technology in Flanders and in the Netherlands in order to consolidate the position of Dutch in the modern information and communication society. The DPC-project contains published material obtained from quality-assured sources. Each text integrated in DPC will consist of a wide range of metadata. Since our team has gained a lot of experience and knowhow of the importance of metadata thanks to DPC, it was chosen to structure the Medi@tic-corpus in a similar way. All documents integrated in both DPC and Medi@tic are presented through the help of a web interface, structured and designed in function of the metadata criteria in order to be user-friendly and easily usable in language courses.

Conclusion

Starting from a database especially built for the Lingu@tic[278] project, the Medi@tic database is now available on the web for the language teacher who wants to use freely available video material in language classes: www.kuleuven-kortrijk.be/mediatic. The importance of authentic material for language learning and teaching has inspired the team to continue the development of Medi@tic. The aim to create a multimodal, multilingual database, available thanks to a web interface has been explained in detail in this paper with a special attention for the enrichment of the learning objects. As shown in the paper, the enrichment can be situated at three different levels. The first level corresponds to the intake of high quality data, such as video, audio and textual material. All material is provided in original lay-out, which means that the learning objects are perfect for any further pedagogical exploitation. The exploitation of transcripts is a second step in the enrichment process of the Medi@tic material. From a pedagogical perspective, the processing of the data presents an enormous added value for the corpus. Not only a lexicon can be created, it is also possible to classify the texts thanks to the annotators. Finally, all objects integrated in the corpus have a list of metadata. These metadata are at the basis of the creation of a pedagogically well-designed and user-friendly web interface. As shown in this paper, a web interface gives an appropriate overview of the pedagogically enriched source material integrated in the presented database.

[277] The acronym STEVIN stands for Speech and Language Technological Essential Data and Facilities in Dutch.
[278] For more information on the project: www.kuleuven-kortrijk.be/linguatic.

References

Desmet, P. – **Héroguel, A.** 2005. “Les enjeux de la création d’un environnement d’apprentissage électronique axé sur la compréhension orale à l’aide du système auteur IDIOMA-TIC”. /ALSIC (Apprentissage des Langues et Systèmes d’Information et de Communication)/ 8. http://alsic.u-strasbg.fr/v08/desmet/alsic_v08_12-poi4.htm

Desmet, P. – **Eggermont, C**. 2006a. “FRANEL: Un environnement électronique d’apprentissage du français qui intègre des matériaux audio-visuels et qui est à la portée de tous ». /Cahiers F. Revue de didactique français langue étrangère/ 7. 39-54

Desmet, P. 2006b. "L’apprentissage/enseignement des langues à l’ère du numérique: tendances récentes et défis". /Revue française de linguistique appliquée/ 11. 119-138.

Paulussen, H. – **Macken, L.** – **Trushkina, J.** – **Desmet, P.** – **Vandeweghe, W.** 2006. "Dutch Parallel Corpus: a multifunctional and multilingual corpus". *Cahiers de l'Institut de Linguistique de Louvain*, CILL, Louvain-La-Neuve, 32.1-4 (2006), 295-312.

CONSTRUCTING A LARGE-SCALE ENGLISH-PERSIAN PARALLEL CORPUS

Tayebeh Mosavi Miangah[279]

Abstract

In recent years the exploitation of large text corpora in solving various kinds of linguistic problems, including those of translation, is commonplace. Yet a large-scale English-Persian corpus is still unavailable, because of certain difficulties and amount of work required to overcome them.

The project reported here is an attempt to constitute an English-Persian parallel corpus composed of digital texts and Web documents containing little or no noise. The Internet is useful because translations of existing texts are often published on the web. The task is to find parallel pages in English and Persian, to judge their translation quality, and to download and align them. The corpus so created is of course open; that is, more material can be added as the need arises.

One of the main activities associated with building such a corpus is to develop software for parallel concordancing, in which a user can enter a search string in one language and see all the citations for that string in it and corresponding sentences in the target language. Our intention is to construct general translation memory software using the present English-Persian parallel corpus.

Keywords: alignment, concordancing, parallel corpus, translation memory

Introduction

A corpus is simply defined as a large collection of linguistic evidence, mainly naturally occurring data either written texts or a transcription of recorded speech. Corpora can be exploited for a range of research purposes in a number of disciplines. In recent years large monolingual, comparable and parallel corpora have played a crucial role in solving problems of computational linguistics, such as part-of-speech tagging (Brill, 1995), word sense disambiguation (Mosavi Miangah and Delavar Khalafi, 2005), language teaching (Aston, 2000; Leech, 1997; Nesselhauf, 2004), phrase recognition (Cutting, et al., 1992), information retrieval (Braschler, & Schauble, 2000) and statistical machine translation (Brown et al., 1990). Corpus-based linguistics has provided an accurate description of languages, and these descriptions of structure and use have many applications in theoretical linguistics, translation tasks and language teaching. There are different kinds of corpora for different kinds of applications. Parallel corpora have texts in one language with the corresponding translations in some other language or languages.

In this paper we present our work on constructing and using English-Persian parallel corpora to support research in fields such as English-Persian bilingual lexicography, developing translation memory software, English-Persian cross-language information retrieval, and statistically-based machine translation from English into Persian. This corpus is extendable: more and more parallel sentences in the languages may be added, and it will be provided free to those interested in language and translation matters, especially translation trainees. One of the main activities associated with building such a corpus is developing software for parallel concordancing, in which a user can enter a search string in one language and see all citations for that string in the search language as well as corresponding sentences in the target language. Aligned bilingual corpora have in fact proved useful in many tasks, including machine translation (Brown *et al.*, 1990; Sadler, 1989), sense disambiguation (Brown *et al.*, 1991a; Dagan *et al.*, 1991; Gale *et al.*, 1992), cross-language information retrieval (Davis and Dunning, 1995; Landauer and Littman, 1990; Oard, 1997) and bilingual lexicography (Klavans and Tzoukermann, 1990; Warwick and Russell, 1990).

Our underlying hypothesis is that exploiting the present English-Persian parallel corpus in building a translation memory system - a database of previously translated units (sentences, phrases and words), along with a set of other translation tools - will result in improvements to the translation process, in speed, consistency and quality.

[279] The author is an assistant professor of applied and mathematical linguistics in English Language Department of Payame Noor University of Yazd in Iran. She has got her Ph.D. from Russian Academy of Sciences in Moscow in 2002. She is currently researching on English-Persian machine translation and related issues such as part of speech tagging, word sense disambiguation, automatic lemmatization and morphological analysis, corpus-based approaches in language processing and the like. Compiling two linguistic corpora – a monolingual Persian corpus as well as a bilingual English-Persian parallel corpus, she is going to design a statistical machine translation system in the time to come.

A typical, and of course the earliest, parallel corpus is the Canadian Hansard corpus, consisting of transcripts of debates from the Canadian Parliament in the country's official languages, English and French. However, in recent years there has been an increase in the number of parallel corpora for various language pairs and of various sizes. For example, an English-Norwegian Parallel Corpus (ENPC) has been created at the University of Oslo (Johansson, 1997), and the Translation Corpus of English and German has been compiled at the Technical University of Chemnitz-Zwickau in Germany. The Japanese NTT Communication Science Laboratories have carried out research into tagging and aligning several collections of Japanese and English texts, while the Thai Internet Education Project has collected and studied on Thai/English parallel texts and developed a toolkit to work on them. (May Fan' and Xu Xunfeng, 2002).

Bilingual corpora for high density languages such as English or French are very extensive, and the results are encouraging because of the easy accessibility of the texts in these languages in digital form, including Websites. However, when a low or medium density language such as Persian is one of the languages involved in a bilingual corpus, the problem is much more difficult because of the shortage of digitally stored materials and detectable parallel pages on the World Wide Web.

The collection of parallel texts

Unlike Resnik (1998), who used an automatic method for extracting parallel material from the Web, we tried to do the task manually. We tried to gather pages from the Web that were potential translations of each other by searching documents in one language which have links containing the name of another language. For instance, if an English web page contains a link such as "Persian page" or "Persian version", the page associated with this link is taken to be a potential translation of that English page.

The Internet is a cumulative source of language data potentially available to everyone, everywhere, for every purpose and in great quantity. However, there are many problems in extracting parallel corpora in English and Persian from the Web, such as the following. As we mentioned earlier, the number of parallel English-Persian pages on the Internet is relatively small. This problem is compounded by the fact that some of these pages are not downloadable because of their special formats, particularly in the Persian side. That is, not all Persian translations of English pages can be copied or downloaded. Some of them are images and others are written in a special type of PDF[280] which cannot be converted into Text format. For example, in the present experiment, about ten percent of the pages obtained from the Web in Persian for which an English translation was available could not be entered into the bilingual corpus for this reason. The other problem encountered when extracting parallel pages from the Web was that some texts in one language were not exact translations of the other language.

Despite these problems, we collected as many well-matched texts in English and Persian as possible. Most Web pages extracted for this purpose had HTML[281] format. However, some of them were PDF, in which case they had to be converted to text format, which in turn has its own difficulties.

In sum, although the availability of bilingual texts involving Persian is subject to some limitations due to the low density of this language around the world and the unavailability of texts in some specific genres and domains, we succeeded in collecting a relatively large number of texts, totaling over 4,860,000 words in English and Persian.

Sources for mining data

As we have seen, over 4,860,000 words of English and corresponding Persian texts have been collected and included in the corpus. The vast majority of these texts was collected from the Internet and covers a variety of domains, such as current affairs, literature, interviews, instruction manuals, religion, pedagogy, and offline digital material. Table 1 presents the different types of texts and their absolute and relative sizes in our English-Persian parallel corpus.

[280]- Portable Document Format
[281] - HyperText Markup Language

Text types	No. of files	words (thousand)	Percent
news	165	499	10.25
reports	102	710	14.58
articles	53	266	5.46
literature	372	1104	22.68
interviews	54	97	1.99
courses	42	65	1.34
offline digital material	265	973	19.98
manuals	142	132	2.71
religious texts	298	1023	21.01
Total	**1493**	**4869**	**100**

Table 1: Distribution of text types in the English-Persian parallel corpus

Corpus preparation

Raw texts extracted from the above sources must be preprocessed to enter the corpus. The preprocessing database is composed of about 1500 files, each containing at least 57 words and at most 30,699 words. Downloading, format conversion, and text normalization are among the labor-intensive steps in preparing the corpus. Some HTML pages which were irrelevant, as well as all figures, tables and pictures, were removed from the texts before they entered into the corpus. Moreover, some parallel pages in PDF format were not convertible to Text format and hence were discarded from the corpus. In some cases where a sentence or part of a sentence was not translated, untranslated parts were deleted. Moreover we tried to manipulate the corresponding paragraphs such that they consisted of corresponding sentences. The resulting movements of sentences or of parts thereof to other sentences in the translations was another problem here. It was resolved by personal checking and manual correction before the texts were entered into the corpus. After verification, they were uniformly encoded into XML[282] format (using XML encoding tools), in order for the corpus to be application-independent and easier to exchange via the Internet. The programming language was ASP, VBSCRIPT, using the databank ACCESS. Persian is of course a language among with rich morphology. We have a program for stemming morphological variants of Persian words (Mosavi Miangah, 2006), but we have not incorporated this program into our corpus because it requires revisions. This will certainly be done eventually.

Aligning the parallel corpus

The very first requirement for any bilingual corpus is alignment. Alignment can be done at the paragraph, sentence, or word level. In this experiment the alignment of sentences was done entirely manually. Although we could have used automatic methods, we preferred to align sentences manually in order for the accuracy to reach 100%. The present corpus is intended to be used in tasks (among which building a translation memory system) for which high accuracy in aligning is crucial.

The corpus software has several components. One enables us to add more sentences with their translations. No limitation was placed on the length of sentences. They vary in length from two to over fifty words. Two other components are an option for editing previous records and one for deleting unwanted records (previously entered) from the corpus. The system has two types of search, simple and advanced. The simple search enables the user to enter a word, a phrase or a sentence in English or Persian as a query. The user then has three choices depending on whether the desired results must correspond exactly or approximately, or both result types are desired. In the first case the result will correspond exactly to the query. In the second case, it contains similar expressions (in terms of graphic forms) to these in the query. In the third case, every word of the given expression occurring in corpus sentences will be shown. It should be noted that the searches are carried out simultaneously in the English and

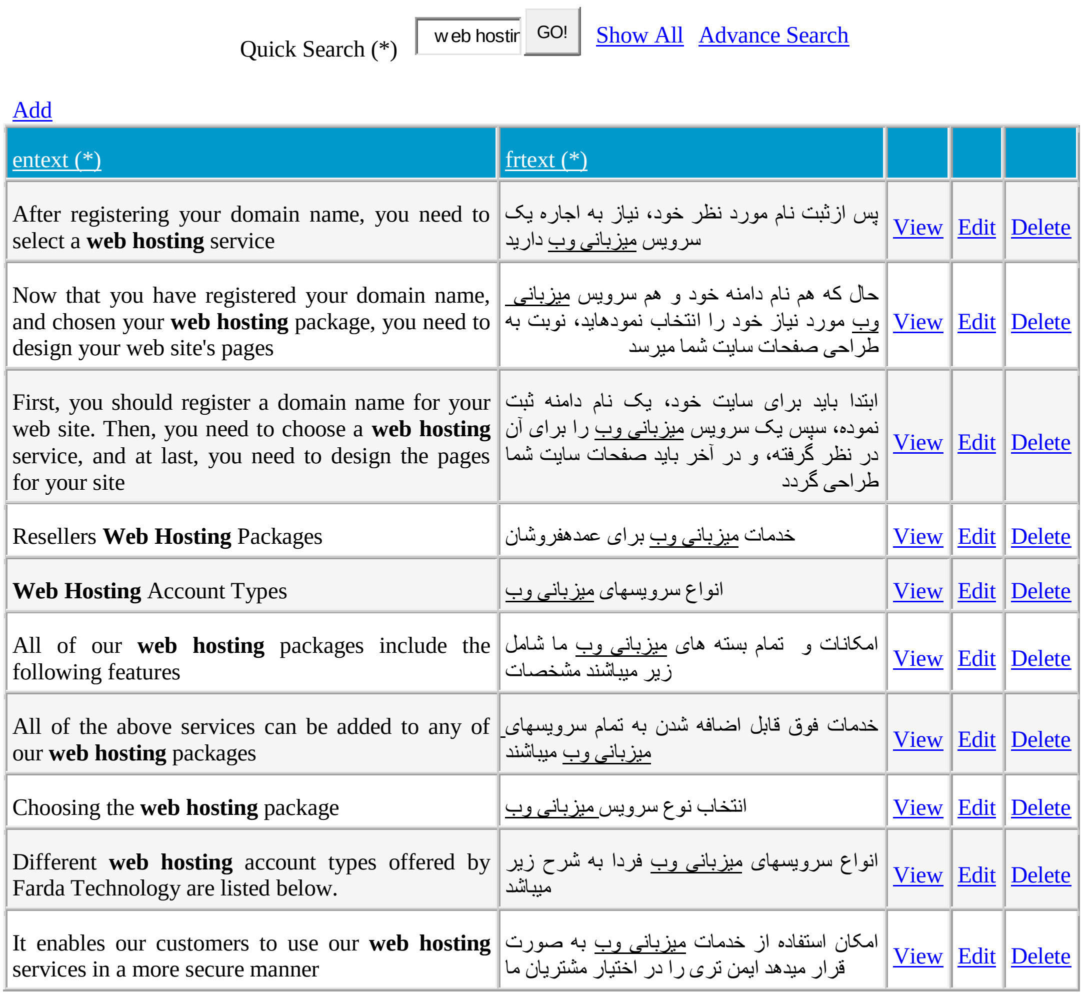
[282] - Extendible Mark-up Language

Persian texts. In the advanced search, the user can search any expression separately in the English or Persian corpus, and the results will be shown only in the selected language. In this sense the corpus acts as two monolingual corpora.

Finding translation equivalents

One of the main applications of parallel corpora is to find different possible equivalents of certain words or collocations. That is, aligned translation units are simply displayed on the screen, offering the translator a range of similar contexts from a corpus of past translations. In some cases we may refer to a parallel corpus to verify the equivalent(s) provided by bilingual dictionaries, since it is believed that parallel corpora provide information that bilingual dictionaries do not usually contain. Using a bilingual dictionary for selecting a translation equivalent, the translator will decide on the appropriateness of different possible equivalents, based on their definitions or the few examples given by the dictionary, while a parallel corpus offers the best possible translation equivalent, based on real-world evidence gained from past translations. It is said that dictionaries take a synthetic approach to lexical meaning (by means of definition), while parallel corpora take an analytic approach (by means of multiple contexts).

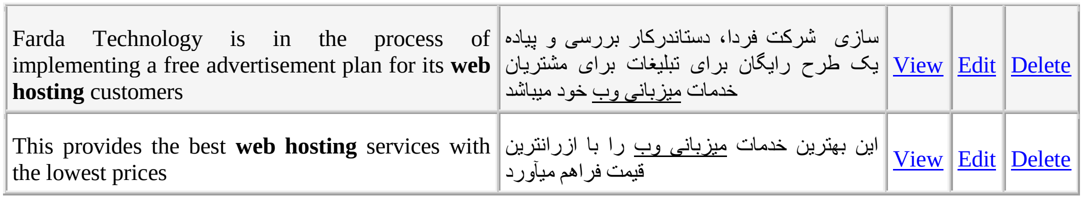
Finding appropriate and natural equivalents for different types of collocations is a difficult task, especially in a non-native language, and parallel corpora can be of great help in this respect. In some cases the aim is to confirm the translation equivalent of a collocation which is the same as the majority of occurrences. As Figure 1 shows, the most frequently occurring equivalent for the collocation “web hosting” is میزبانی وب , with a frequency of 12 in the corpus.



Quick Search (*) web hostir GO! Show All Advance Search

Add

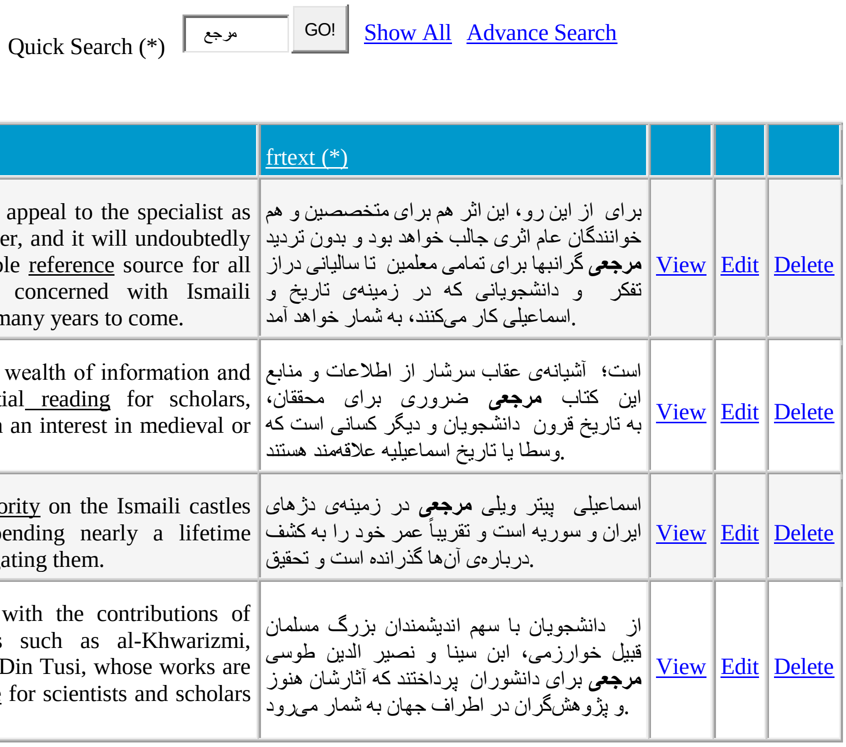
entext (*)	frtext (*)			
After registering your domain name, you need to select a **web hosting** service	پس ازثبت نام مورد نظر خود، نیاز به اجاره یک سرویس میزباني وب دارید	View	Edit	Delete
Now that you have registered your domain name, and chosen your **web hosting** package, you need to design your web site's pages	حال که هم نام دامنه خود و هم سرویس میزباني وب مورد نیاز خود را انتخاب نمودهاید، نوبت به طراحی صفحات سایت شما میرسد	View	Edit	Delete
First, you should register a domain name for your web site. Then, you need to choose a **web hosting** service, and at last, you need to design the pages for your site	ابتدا باید برای سایت خود، یک نام دامنه ثبت نموده، سپس یک سرویس میزباني وب را برای آن در نظر گرفته، و در آخر باید صفحات سایت شما طراحی گردد	View	Edit	Delete
Resellers **Web Hosting** Packages	خدمات میزباني وب برای عمدهفروشان	View	Edit	Delete
Web Hosting Account Types	انواع سرویسهای میزباني وب	View	Edit	Delete
All of our **web hosting** packages include the following features	امکانات و تمام بسته های میزباني وب ما شامل زیر میباشند مشخصات	View	Edit	Delete
All of the above services can be added to any of our **web hosting** packages	خدمات فوق قابل اضافه شدن به تمام سرویسهای میزباني وب میباشند	View	Edit	Delete
Choosing the **web hosting** package	انتخاب نوع سرویس میزباني وب	View	Edit	Delete
Different **web hosting** account types offered by Farda Technology are listed below.	انواع سرویسهای میزباني وب فردا به شرح زیر میباشد	View	Edit	Delete
It enables our customers to use our **web hosting** services in a more secure manner	امکان استفاده از خدمات میزباني وب به صورت قرار میدهد ایمن تری را در اختیار مشتریان ما	View	Edit	Delete



Farda Technology is in the process of implementing a free advertisement plan for its **web hosting** customers	سازی شرکت فردا، دستاندرکار بررسی و پیاده یک طرح رایگان برای تبلیغات برای مشتریان خدمات میزبانی وب خود میباشد	View	Edit	Delete
This provides the best **web hosting** services with the lowest prices	این بهترین خدمات میزبانی وب را با ازرانترین قیمت فراهم میآورد	View	Edit	Delete

Figure 1: Display of some lines generated by our corpus for the search word "web hosting"

However, there are other cases in which the translator needs to see all possible equivalents of a certain expression in each language, and then make the best decision, based on similar or identical contexts in which the expression is found. Four different English translations of the word مرجع have been displayed in Figure 2.



Quick Search (*) مرجع GO! Show All Advance Search

Add

entext (*)	frtext (*)			
As such, this work will appeal to the specialist as well as the general reader, and it will undoubtedly prove to be an invaluable reference source for all teachers and students concerned with Ismaili history and thought for many years to come.	برای از این رو، این اثر هم برای متخصصین و هم خوانندگان عام اثری جالب خواهد بود و بدون تردید **مرجعی** گرانبها برای تمامی معلمین تا سالیانی دراز تفکر و دانشجویانی که در زمینه‌ی تاریخ و اسماعیلی کار می‌کنند، به شمار خواهد آمد.	View	Edit	Delete
Eagle's Nest contains a wealth of information and resources; it is essential reading for scholars, students and others with an interest in medieval or Ismaili history.	است؛ آشیانه‌ی عقاب سرشار از اطلاعات و منابع این کتاب **مرجعی** ضروری برای محققان، به تاریخ قرون دانشجویان و دیگر کسانی است که وسطا یا تاریخ اسماعیلیه علاقه‌مند هستند.	View	Edit	Delete
Peter Willey is an authority on the Ismaili castles of Iran and Syria, spending nearly a lifetime discovering and investigating them.	اسماعیلی پیتر ویلی **مرجعی** در زمینه‌ی دژهای ایران و سوریه است و تقریباً عمر خود را به کشف درباره‌ی آن‌ها گذرانده است و تحقیق.	View	Edit	Delete
The students engaged with the contributions of great Muslim thinkers such as al-Khwarizmi, Avicenna, and Nasir al-Din Tusi, whose works are still a point of reference for scientists and scholars around the world.	از دانشجویان با سهم اندیشمندان بزرگ مسلمان قبیل خوارزمی، ابن سینا و نصیر الدین طوسی **مرجعی** برای دانشوران پرداختند که آثارشان هنوز و پژوهش‌گران در اطراف جهان به شمار می‌رود.	View	Edit	Delete

Figure2: Display of some lines generated by our corpus for the search word " مرجع "

This shows the actual translations of the same expression chosen by translators according to the context, concrete data which is not found directly in bilingual dictionary.

Conclusion

In this paper we have described a general method for collecting, building, and aligning a parallel corpus for English and Persian. The corpus so created is open; that is, more material can be added as the need arises. Naturally, the richer the corpus is in terms of the volume of data and its variety, the more useful it will be for solving linguistic problems. This is a work in progress and there is great room for enhancing the potential of the corpus. New ways may be found to obtain more parallel texts. As we have mentioned, the present corpus will be aligned at word level in the near future, making the corpus a database for a translation-memory system. At that stage the corpus will be of even greater help to translators between English and Persian. A parallel concordance tool is a future goal in this direction.

Acknowledgements

The research project reported here is partly supported by the Research Affairs Office of Payame Noor University of Yazd in Iran.

References

Aston, G. 1997. “Enriching the learning environment: Corpora in ELT”. In: A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (51-64). New York: Addison Wesley Longman.

Braschler, M. and Schauble, P. 2000. “Using corpus-based approaches in a system for multilingual information retrieval”. *Information Retrieval*, 3, P. 273-284.

Brill, E. 1995. “Unsupervised learning of disambiguation rules for part of speech tagging”. In *2nd Workshop on Large Corpora*, Boston, USA.

Brown, P., Cocke, S., Della Pietra, V., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R. & Roosin, P. 1990. “A Statistical Approach to Machine Translation”. *Computational Linguistics* 16:2, P. 79-85.

Chris Callison-Burch and Miles Osborne. 2003. “Bootstrapping Parallel Corpora”. In *NAACL workshop "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond".*

Cutting, D.; Kupiec, J.; Peterson, J. and Sibun. P. 1992. “A practical part of speech tagger”. In *proceeding of 3rd Conference on Applied Computational Linguistics*, Trento, Italy, P. 133-140.

Davis M. and Dunning, T. 1995. “A TREC evaluation of query translation methods for multi-lingual text retrieval”. In *4th Text Retrieval Conference (TREC-4).* NIST.

Johansson, S., 1997.: “Using the English Norwegian parallel corpus—a corpus of contrastive analysis and translation studies”. In: Lewandowska-Tomaszczyk, B., Melia, J. (Eds.). PALC ’97 Practical Applications in Language Corpora. Lodz University Press, P. 282–296.

Landauer, T. K. and Littman, M. L. 1990. “Fully automatic cross-language document retrieval using latent semantic indexing”. In *Proceedings of the 6th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, P. 31-38, Uw Center for the New OED and Text Research, Waterloo, Ontario.

Leech, G. 1997. “Teaching and language corpora: A convergence”. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (1-23). New York: Addison Wesley Longman

May Fan' and Xu Xunfeng. 2002. “An evaluation of an online bilingual corpus for the self-learning of legal English” , *System* Volume 30, Issue 1, P. 47-63.

Mosavi Miangah, T. and Delavar Khalafi, A. 2005. “Word sense disambiguation using target language corpus in a machine translation system”. *Literary and Linguistic Computing*, 20(2), P. 237-249.

Mosavi Miangah, T. 2006. “Automatic lemmatization of Persian words”. *Journal of Quantitative Linguistics*, 13(1), P. 1-15.

Nesselhauf, N. 2004. “Learner corpora and their potential for language teaching”. In J. McH. Sinclair (Ed.), *How to use corpora in language teaching* (125-152). Amsterdam: Benjamins.

Oard, D. W. 1997. “Cross-language text retrieval research in the USA”. In *3rd DELOS Workshop.* European Research Consortium for Informatics and Mathematics.

Resnik, PH. 1998. “Parallel strands: A preliminary investigation into mining the web for bilingual text”. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529,* Langhorne, PA, October, P. 28-31.

Resnik, PH. 1999. Mining the web for bilingual text”. *Proc. of 37th Meeting of the ACL.* Maryland. P. 527-534.

Sun Lee, Lin Du, Yufang Sun, Jin Youbin. 1999. “Sentence Alignment of English-Chinese Complex Bilingual Corpora”. In *Proceeding of the workshop MAL'99*, P. 135-139.

Sun, Lee, Song Xue, Weimin Qu, Xiaofeng Wang, Yufang Sun. 2002. “Constructing of a Large-Scale Chinese-English Parallel Corpus”**.** *Proceedings of the 3rd workshop on Asian language resources and international standardization* - Volume 12, P. 1 – 8.

FLT MEETS SLA RESEARCH: THE FORM/FUNCTION SPLIT IN THE ANNOTATION OF LEARNER CORPORA

Stefano Rastelli[283]

Francesca Frontini[284]

Abstract

Our work[285] *explores the advantages of adopting a strict form-to-function perspective when annotating learner corpora. Hopefully, such a perspective provides both Foreign Language Teaching (FLT) and Second Language Acquisition (SLA) researchers with insights not relating to learners' errors, but to some systematic features of interlanguage (IL). A split between forms and functions (or categories) is desirable in order to avoid both the "closeness fallacy" and the "comparative fallacy". In fact - especially in basic learner varieties - forms (or "functors") may precede functions and in their turn, functions may show up in unexpected forms. In the computer-aided error analysis (CEA) tradition, all items produced by learners are traced to a grid of error tags, which is based on the categories of the target language (TL). In a different way, we believe it is preferable to account for IL features in terms of "virtual" TL categories. For this purpose, a preliminary project-study for the tagging of L2 Italian (PIL2) has been completed at the University of Pavia. The project concluded that it is possible to use a tree-tagger designed for L1 Italian also for learner data on condition that the tagging system retrieves separately four levels of annotation: (a) the information about how a word is actually spelled / uttered by learners; (b) its position in the sentence; (c) the virtual categories attributed to that form on the basis of formal resemblance with TL items; (d) the level of confidence in recognizing both the category and the lemma. The aim of PIL2 project is not to disclose areas where learners show under-use or overuses of linguistic features nor to know which errors learners commit more. Using a tree-tagger designed for L1 Italian on data of learner Italian may reveal unexpected IL phenomena and allows us to see how the functions of the TL are gradually acquired by learners.*

Keywords: Interlanguage, learner corpora, error-tagging, comparative fallacy, L2 Italian.

Is error tagging really inherent to learner corpora?

Far from neglecting or minimizing the tremendous importance of error tagging, especially for teaching purposes and for lexicography, we would like to propose a different way of pursuing the annotation of learner corpora. Our proposal gives up error tagging and consequently our answer to the question posed in the title of this paragraph (that is taken from Díaz-Negrillo and Fernández-Domínguez, 2006:84) is assumed to be "no". Error tagging should neither be considered as the Pillars of Hercules, beyond which the world ends, nor the only means available to teachers of becoming aware of learners' performance. The reason in twofold. First of all, it is possible that the nature of errors does not make them the best candidate possible for SLA research (see next paragraph). Secondly, researchers have yet to agree about general error-taxonomy, the standardization of error tagset still a long way from being at hand (Tono, 2003:801). According to Díaz-Negrillo and Fernández-Domínguez (2006:89), the number of tags in different error-tagging projects varies from 31 to 100. As far as the layers of analysis are concerned, phonetic, pragmatic and discourse errors are treated rarely and inconsistently, while the textual dimension do not seem to be considered at all (see also Rastelli, 2007: 99).

[283] Stefano Rastelli is post-doc research fellow at the University of Pavia where he also works as Italian language coordinator. He has been teaching Italian as a foreign language since 1988. His main areas of interest are: Second Language Acquisition (the acquisition of tense-Aspect system), Syntactic Theory, Corpus Linguistics and Foreign Language Teaching.

[284] Francesca Frontini is a Ph.D student at the University of Pavia. She is currently dealing with measuring the performances of stocastic algorithms on learner corpora. Her main areas of interest are: Computational Linguistics, Corpus Linguistics, Second Language Acquisition.

[285] Stefano Rastelli wrote the first five paragraphs and the conclusions while Francesca Frontini wrote the sixth, the seventh and the eighth paragraph.

The error tagging and the "comparative/closeness fallacy"

A Chinese beginner student of L2 Italian, describing a house suddenly catching fire, says: *la casa di loro c'è fuoco* [lit. "The house of them there is fire"]. None of the items of this sentence taken individually is wrong, nor is it straightforward to pinpoint the source of the ill-formedness. Despite the fact that this scene is clear, it is not enough in order to label the possible errors unambiguously because there are at least three ways to correct the "wrong" sentence. Far from being an exception in learner data, sentences like the one above show that - unfortunately for us - many interesting IL features are not proper "errors", that is, they do not show up as "incorrect forms" each having one or more correct equivalent in a native speakers' mind. First of all in learner data it is not always possible even to isolate the form responsible for the sentence becoming incorrect or to define what this form, once singled out, stands for (that is, which is its "correct version" in the TL provided that it has just one, see Rastelli, 2007). Secondly, errors are often seen as token-based, whilst they often entail (or are embedded in) other errors (this problem has been recently addressed by adopting a multi-level standoff annotation, see Lüdeling et al., 2005). Finally, especially in basic varieties, learners often produce not just "lacking", "wrong" or "mispelled" items, but rather "impossible" ones (the issue of the existence of different layers of "grammaticality" is partially addressed also in Foster, 2007:131). Here "impossible" is meant as unclassifiable and unpredictable. "Unclassifiable" is a combination of a number of *per se* well-formed items, that a native-speaker perceives as being wrong as a whole, despite not knowing the precise rule being violated. "Unpredictable" is a combination of characters whose nature is not capturable by using a pre-fabricated, closed set of errors, no matters its size. It has been pointed out that the practice of error tagging rests on native speaker's intuition. The elaboration of an error manual is usually meant to avoid or at least minimize taggers' subjectivity when dealing with deviant phenomena. While, in everyday life judgements, subjectivity is not necessarily a flaw, when it plays a decisive role in annotation of learner corpora it is at risk of committing "comparative fallacy" and "closeness fallacy", as far as these two concepts are intended by Huebner (1979), Bley-Vroman (1983), Klein and Perdue (1992), Cook (1997), Lakshmanan and Selinker (2001) (see also a special issue of TESOL &Applied Linguistics, 2004). The comparative fallacy emerges when a researcher studies the systematic character of one language by comparing it to another or (as often happens) to the TL. The "closeness fallacy" occurs "in cases where an utterance produced bore a superficial resemblance to a TL form, whereas it was in fact organised along different principles" (Klein & Perdue, 1992: 333). The comparative fallacy represents an attitude, while the closeness fallacy the most likely case of its practical application, that is, when the TL coincides with the language of the researcher. Failure to avoid the comparative fallacy will result in "incorrect or misleading assessments of the systematicity of the learner's language". Bley-Vroman's criticism (1983: 2) applies also "to any study in which errors are tabulated [...] or to any system of classification of IL production based on such notions as *omission*, *substitution* or the like". The logic of "correct-incorrect" binary choice which is so peculiar to errors, hides the fact that the surface contrast in IL may be determined by no single factor, but by a multiplicity of interacting principles, some of which unknown (8). For all these reasons, it is the analysis of unexpected and "spurious" items sorted out by the system also in non obligatory contexts that is likely to reveal the systematicity of some IL features. Since using error-tags means to get exactly what one expects and to hide developing and provisional non target-like learner grammars, in our project it was decided to find an alterative way to run queries on learner corpora. Since this query system should have been TL rule-oriented and not TL rule-governed, it was thought that the best way to deal with learner data without error tagging would have to focus on some kind of xml treatment of the outcome of a Treetagger designed for L1 Italian.

"Unexpected" data and patterned queries

The fact that, according to our view, "unpredictable data" is so important for SLA research does not mean that we should give up using TL categories and that all queries on the learner data should be carried out randomly. Also "unexpected/unpredictable" data should be looked for systematically when testing a hypothesis about developing learner grammars. The following example is taken from the Pavia Corpus. A Chinese beginner student of L2 Italian, when asked to report about his education, said: *Cinese fato media* ("Chinese done middle [school]" that is assumed to mean: "In China I attended the middle school"). A few days later, when asked about holidays, the same learner said that: *Sì, in Cina festa pasqua anche* ("Yes, in China holiday Easter too", that is assumed to mean: "Yes, in China there are Easter Holidays as well"). Following the bracketed and provisional interpretation and under an error-driven perspective, only in the first sentence is the learner blurring the distinction between the category of adjectives ("Chinese") and the category of nouns (here placed into a locative expression "in China"). We thus could label this as an "error" following the appropriate category of FRIDA tagset. It would belong to the subset of errors named "class" <CLA> (exchange of class) and to the higher set of Grammar <G> errors (Granger, 2003: 4). If we adopt a different perspective, we might compare the two items *cinese* and *Cina* in order to test the hypothesis that the learner in question is not lacking a rule, nor is he/she wild-guessing or even backsliding in his/her developmental path, but simply that she's/he's applying some kind of rule that affects both occurrences. We don't know this rule yet nor can we easily figure out what kind of rule it is. Using any tag based on binary opposition (correct vs. incorrect) would be misleading. The solution is to sort out all "virtual adjectives" and "virtual nouns" (for a detailed meaning of "virtual", see next paragraph) containing similar strings of characters (in our case, *c-i-n* or the like) in different positions of the sentence. By repeating this query pattern throughout the sentences in the corpus, we might find out that "virtual" adjectives (like *cinese*) rather than "virtual nouns" (like *Cina*) are likely to be placed to the initial place, at the left periphery of the sentence (the typical topic-position in Chinese) and that this preferably happens when a noun (like *media*) occurs somewhere rightwards. Or we might find out that the differences in suffixation that we expect to be between adjectives and nouns (*-ese* vs. *-a* or zero-suffix) are systematically blurred when there is what we interpret as being a locative expression. If either of these combinations of facts recurs systematically in the corpus, then the grammar of the learner might contain a rule of the kind "position of items counts more than their eventual suffix" or "items in locative expressions agree, regardless their category". If, on the contrary, these combinations do not recur systematically, it is likely that the learner's grammar does not contain such rules or that our interpretation of the learner's sentences was wrong under some respects. Whatever the answer, since this procedure prevents researcher's interpretation from affecting the annotation of the sentence, sooner or later other unexpected linguistic features will surface from the corpus and new hypotheses will be made available to be systematically tested out on data.

TL Rule-motivated vs. Form-motivated, "virtual" categories

As Nicholls (2003: 572) pointed out, error tags are not an end in themselves, "but rather act as a bookmark" for queries, that is, they should give the researcher the information they are looking for. Contrarily, our point is that error tags are likely to commit comparative/closeness fallacy and to obstacle - instead of allowing - the retrieval of important IL phenomena because what they are likely to annotate is taggers' TL-governed interpretation (often just one among other possible interpretations), not the structural value of the item in the IL. In everyday experience, human interpretation is called into action to unpredictable extent when trying to make sense of learners' utterances. We can include it in the annotation consistently or completely exclude it from annotation at the cost of losing usability in the query system. The solution provided is a compromise between transparency of data and usability. On one hand we decided to exclude all interpretation based on taggers' judgements, on the other hand we encoded all interpretations based on automatic and successful matching between the item in question and all TL items. In our view, this would prevent running the risk of "ontologizing" errors, that is, to treat them as if they were really psychological *realia*, sort of holes or gaps existing in learners' mind. Functional interpretation is thus excluded and "virtual", formal-motivated (TL-oriented) tags substitute rule-motivated (TL-governed) tags by allowing different levels of annotation, as will be shown in next two paragraphs.

When a L1 tagger is run on a learner corpus

The key idea is to use a L1 tagger on the L2 corpus as a means of detecting virtual categories corresponding to each L2 item. In our opinion, far from being a step back, this would help minimize the risk of comparative fallacy and gain deep insights into learners' IL. Using a strictly formal definition we can identify a category by lexical root, by morphology or by context. There are formal hints that must be taken into account in order to recognize, say a verb in a sentence like "Loro andavano a scuola" (They went to school): post-pronominal position, a verbal root like "and-" (go), verbal inflection "-avano" (3 person plural imperfect). In L1 the criteria normally converge and tend to be redundant. Rule-based taggers for instance generally rely on morphology and lemma in conjunction, so they will only recognize known lemmas with the right morphology attached. In IL, on the contrary, not all criteria are always satisfied at the same time. So ideally we need a much more flexible tagger that takes into account all hints and expresses a possible tagging together with its level of confidence. We chose to use Treetagger (Schmid, 1994), with the standard tagset and the standard training for Italian L1 and obtained encouraging results. Being built on a probabilistic algorithm, Treetagger will recognize, say, a verb by the presence of either a verbal position, a verbal root or a verbal morphology. These levels are independent: the tagger recognizes a verbal ending even if this is attached to an unknown lemma. Therefore, once each word is analyzed, the tagger issues a tag, a lemma (which can be <unknown>) and a confidence probability, which is determined by the convergence of the different hints. A verbal tag with lemma <unknown> and a low level of confidence means that the lexical criteria failed and that the tagging was performed on the basis of position and (possibly) morphology.

Annotation sample

Once the annotation per category, lemma and probability is translated in xml tags, queries can be performed on the corpus, mixing the virtual categories level with positional information and formal data at the source level (via regular expressions matching). The tagset at word level is defined as follows:

<token> – grammatical word

attributes: **tag** – part of speech; **lemma**; **prob** – Treetagger confidence level

Here is a sample of annotated text:

"è un bambino che in la camera sua ha un cane e una rana..." (it's a child that in his room has a dog and a frog).

```
<token tag="VER:pres" lemma="essere" prob="1.000000">è</token>
<token tag="DET:indef" lemma="un" prob="0.998249">un</token>
<token tag="NOM" lemma="bambino" prob="1.000000">bambino</token>
<token tag="PRO:rela" lemma="che" prob="0.594519">che</token>
<token tag="PRE" lemma="in" prob="1.000000">in</token>
<token tag="DET:def" lemma="il" prob="0.999939">la</token>
<token tag="NOM" lemma="camera" prob="1.000000">camera</token>
<token tag="PRO:poss" lemma="suo" prob="1.000000">sua</token>
<token tag="VER:pres" lemma="avere|riavere" prob="1.000000">ha</token>
<token tag="DET:indef" lemma="un" prob="0.998249">un</token>
<token tag="NOM" lemma="cane" prob="1.000000">cane</token>
<token tag="CON" lemma="e" prob="1.000000">e</token>
<token tag="DET:indef" lemma="una" prob="1.000000">una</token>
<token tag="NOM" lemma="rana" prob="0.694963">rana</token>
<token tag="ADV" lemma="dentro" prob="0.830941">dentro</token>
<token tag="PRE" lemma="di" prob="1.000000">di</token>
<token tag="DET:indef" lemma="un" prob="0.997119">un</token>
<token tag="NOM" lemma="barattolo" prob="1.000000">barattolo</token>
<token tag="SENT" lemma="." prob="1.000000">.</token>
```

Basic queries

We give here just one example of how to query the tagged corpus in order to find IL features (including the so-called "errors") without any need of error tags, just by using the following information from Treetagger: (a) (form-motivated) virtual categories; (b) the level of confidence in the tagging and in recognizing the lemma; (c) strings and positional context. Note how here that the possible weakness in analysing IL with a TL tagger, with all recognition problems involved, turn out to become an advantage for the end user. Let's imagine we want to investigate the *transition from indiscriminate to selective verbal suffixation*, this being our starting hypothesis on the learner developing grammar. Here are some useful and very simple queries, using first lemma information, then adding confidence level information and finally position:

Query 1: search tokens with lemma <unknown> that have been tagged as verbs (at this stage the level of confidence is ignored). The query outputs contexts such as:

Query 1: search tokens with lemma <unknown> that have been tagged as verbs (at this stage the level of confidence is ignored). The query outputs contexts such as:

(1.a) il ragazzo **pienere** su la roccia per gritare
the boy <unknown>-verb:infinite on the rock to cry

(1.b) ogni giorno **conoscia** dieci persone
every day (he) meets? ten people

In (1.a) the system recognizes something that could resemble the infinite suffix "-ere" even if it is attached to an unknown stem. Maybe the learner is trying to categorise the token as verb by using verbal morphology: if this is the case, the tagger recognizes it. In (1.b) both root and morphological agreement are target like, but the lemma is not recognised.

Query 2: search all verbs with lemma NOT <unknown> which have been tagged with confidence less then 1.0. This captures all virtual verbs that have been recognised by Treetagger with some degree of uncertainty, like:

(2.a)	quando	si	**sveglia**		il	bambino
	when	(refl)	wakes	up	the	child

(2.b)	salì	a	la	cime	de	una	rocca.	Continua	**chiamandola**
	climbed	to	the	top	of	a	rock.	keeps	calling (ger+clit.)

(2.c)	è	sotto	una	nave	che	si	**sta**	costruggendo
	is	under	a	ship	that	(imp.)	is being	built

Here we get a broader spectrum of phenomena, some of them unexpected and really interesting. We have target like sentences (2.a) in which a form that presents a categorial ambiguity in isolation (sveglia_noun, “alarm-clock” vs sveglia_verb3ps, “wake up”) is correctly disambiguated by context; well formed items in unexpected and possibly non target-like contexts, as in (2.b), where the presence of the verb “continuare” normally requires “a”+infinitive; ill-formed items like “costruggendo”, which apparently stems from the root of the TL verb “costruendo” ("build") in (2.c). Note that these contexts are retrieved without previously tagging them with any error category on purpose.

Query 3: search all sequences of token 1 and token 2 such as token 1 is a virtual verb with confidence < 1 (some degree of uncertainty) and with lemma <unknown> and token 2 is a virtual verb of any kind.

(3.a)	e	corri	corri	corri	il	bambino	sulla	testa
	and	run	run	run	the	child	on the	head

(3.b)	ho	dovuto	parlare	l'	inglese
	had	must	speak	the	English

(3.c)	e	quando	il	furgone era	andato	
	and	when	the	truck	was	gone

Here too a variety of phenomena is present in the output: conversational traits such a repetitions and false starts (3.a); target like compound verbs and verbal periphrasis (3.b, modal) in what one may judge being appropriate or inappropriate context. Again, since our point is that a certain amount of spurious results is proof of the absence of comparative fallacy, also the transparency of the data is thus being respected. Queries like these should be run on portions of the corpus divided by level (and by learner) in order to study the evolution of the phenomena in object.

Future Developments

Using XSL-transformations on XML allows us not only to query the corpus, but also to add further tags “online”. This can be implemented to allow the researchers to assign their own further levels of annotations, like tagging functions related to the sistematicity they might have found in the IL. These tags could be later combined with the others to perform "patterned queries", that restrict the search in a more fine grained and specific way without using any error tag.

References

Bley-Vroman, R. 1983. The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning, 33*: 1-17.

Díaz-Negrillo, A., Fernández-Domínguez, J., 2006. Error tagging system for learner corpora. *RESLA, 19*: 83-102.

Cook, V., 1997, Monolingua Bias in Second Language Acquisition Research. *Revista Canaria de Estudios Ingleses, 34*: 35-50.

Foster, J., 2007. Treebanks gone bad. Parser evaluation and retraining using a trebank of ungrammatical sentences. *International Journal on Document Analysis and Recognition, 10*: 129-145.

Granger, S., 2003, Error-tagged learner corpora and CALL: a promising synergy. *CALICO 20/3*: 465-480.

Huebner, T., 1979, Order-of-Acquisition vs. dynamic paradigm: A comparison of method in interlanguage research, *TESOL Quarterly*, *13*: 21-28

Klein, W., Perdue, C., 1992, “Utterance structure”. In *Adult language acquisition: cross-linguistic perspectives. Vol.2: The results*, C.Perdue (ed.), Cambridge, Cambridge University Press.

Lakshmanan, U., Selinker, L. 2001. Analysing interlanguage: How do we know what learners know? *Second Language Research*, *17*: 393-420.

Lüdeling, A., Walter, M., Kroymann, E., Adolphs, P. 2005. "Multi-level error annotation in Learner Corpora". Paper presented at the *Corpus Linguistics 2005 Conference*, Birmingham, U.K., www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc [access date 25/04/2008]

Nicholls, D., 2003, “The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT”. In *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster, United Kingdom.

Rastelli, S., 2007. Going beyond errors: position and tendency tags in a learner corpus”. In *Language Resources and Linguistic Theory*, A. Sansò (ed.), Milano, Franco Angeli, 96-109.

Schmid, H., 1994, “Probabilistic Part-Of-Speech Tagging Using Decision Trees”. *Proceedings of International Conference on New Methods in Language Processing,* Manchester, United Kingdom.

Tono Y., 2003. “Learner corpora: design, development and applications”. In *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003). Technical Papers 16*, D. Archer, P.Rayson, A.Wilson, T.McEnery (eds.), University Centre for Computer Corpus Research on Language, Lancaster, 800-809.

“AND WE’RE TALKING ABOUT ALL THE CRUCIALLY IMPORTANT THINGS”: EXPLORING THE PROGRESSIVE IN BULGARIAN AND GERMAN EFL WRITING

Svetla Rogatcheva[286]

Abstract

Certain areas of the English tense and aspect system are notoriously difficult for non-native speakers of English from a variety of mother-tongue backgrounds (cf. Hinkel 2004, Swan and Smith 1987). One such area is the progressive aspect - even advanced learners of English are often reported to use the progressive in a non-targetlike manner, failing to recognize its conventionalized meanings and functions in speech or writing. A small number of studies have so far employed a corpus-based approach to the use of the progressive in advanced EFL learners’ production and have shown differences between native and EFL learners’ patterns of use (e.g. Lenko-Szymanska 2007, Axelsson and Hahn 2001, Virtanen 1997).

The present study is a work-in-progress report which aims to explore advanced EFL learners’ use of the progressive in writing from a contrastive perspective. The research is based on the Bulgarian and German components of the International Corpus of Learner English (ICLE), which are compared to a control corpus – the Louvain Corpus of Native English Essays (LOCNESS) within the Contrastive Interlanguage Analysis framework (Granger 1996, Granger et al. 2002).

A preliminary analysis of small subcorpora of these components reveals remarkable similarities between the patterns of use of the progressive in the German learner subcorpus and the American subcorpus, as well as between those in the Bulgarian learner subcorpus and the British subcorpus. The goal of the present study is to exemplify some of the preliminary findings and to suggest possible explanations with respect to learner-related variables like the native language influence and the target language exposure.

Keywords: progressive aspect, learner corpus, Contrastive Interlanguage Analysis

Introduction

The progressive aspect is a feature of English grammar that has been explored in various diachronic and synchronic contexts in terms of its frequency of use, meaning variation and discourse functions. Still, second language use of the progressive has been sparsely examined, mostly in relation to lexical aspect within the framework of the Aspect Hypothesis (Wible and Huang 2003, Andersson 1996). A small number of studies have employed a corpus-based approach to second language use of the progressive in authentic learner writing, focussing on advanced EFL learners of Polish, German, Swedish and Finnish mother-tongue backgrounds (Lenko-Szymanska 2007, Axelsson and Hahn 2001, Virtanen 1997). These studies demonstrate that even advanced learners of English fail to use the progressive in a native manner, although some learner populations seem to be more successful than others. The aim of the present study is to continue the corpus-based research tradition and analyse the use of the progressive in argumentative writing produced by advanced EFL learners of very different mother-tongue backgrounds – Bulgarian and German.

The progressive aspect

From a semantic point of view, the progressive aspect in English is a category of the verb phrase which “designates an event or state of affairs in progress, or continuing, at the time indicated by the rest of the verb phrase” (Biber et al. 1999: 460). Quirk et al. define three basic meaning components of the progressive – duration, limited duration and incomplete action (Quirk et al. 1985: 198), whereas Mindt’s recent *Empirical Grammar of*

[286] Svetla Rogatcheva is a research assistant and a PhD student at the department of English Language and Linguistics at the University of Giessen, Germany. Her dissertation project deals with the use of tense and aspect markers in the argumentative writing of advanced Bulgarian and German EFL learners in contrast to non-professional native writing. She received a BA in English Language, Literature and Culture from the University of Sofia, Bulgaria in 2002 and an MA in English Linguistics from the University of Bayreuth, Germany in 2005. Her research interests and undergraduate teaching include second language acquisition, corpus linguistics, contrastive linguistics and morphology.

the English Verb System lists as many as nine different meanings of the progressive, ranging from temporal meanings like incompletion or temporariness to non-progressive meanings like highlighting, emotion or politeness (Mindt 2000). The most prominent meanings of the progressive as defined by Mindt are incompletion, temporariness and habit, whereas less frequent meanings include highlighting, prediction or politeness. Overlaps of two or more meanings in a single verb phrase are also common and therefore difficult to delineate (Mindt 2000: 256). These corpus findings are in line with Comrie's observation that "it may well be that English is developing from a restricted use of the progressive, always with progressive meaning, to this more extended meaning range" (Comrie 1976: 39).

Other aspects of the progressive concern its frequency of occurrence and distribution across different registers and varieties of English. The progressive is defined as an 'infrequent phenomenon' that occurs in less than 5% of all verb phrases in present-day English (Quirk et al. 1985: 198, Biber et al. 1999: 461). In terms of register distribution, the progressive is a feature of spoken rather than written English, as it is most frequent in conversation and least frequent in academic writing. In terms of regional variation, the progressive aspect is favoured by American English rather than British English in the approximate ratio of 4:3 (Biber et al. 1999: 461-462). Nevertheless, a considerable increase in use of the progressive aspect has been observed for both British and American English over the past few decades (Mair and Hundt 1995). Mair and Hundt account for this increase in terms of the gradual 'colloquialisation' of British and American news writing, where the progressive functions as a stylistic device, bridging the gap between spoken and written language (Mair and Hundt 1995: 117).

A third problem concerns the influence of the learners' mother tongue and other learner-related variables on the use of the progressive in second language argumentative writing. Most research on second language aspect use has neglected "what learners know from their native language, and what exactly their learning task is in acquiring a second language, including areas of typological difference where they may have the most difficulties" (Slabakova 2002: 185). Both Bulgarian and German as native languages lack the progressive aspect as a grammatical category and employ other grammatical and lexical means to convey a state or event in progress. The progressive aspect is absent in standard German; however, there are a few constructions indicating duration or incompletion consisting of the verb 'to be' and an adverbial phrase, which show certain features of grammaticalisation (Andersson 1989). In Bulgarian, the progressive is subsumed under a more general category of imperfectivity present in all Slavonic languages, which is expressed through imperfective verb inflections and an imperfect past tense (Scatton 2000). Thus, learning to use the progressive idiomatically in English writing requires not only acquiring the temporal and non-temporal meanings of the English progressive, but also acquiring their conventionalised uses which are appropriate for an academic context in the English-speaking world (cf. Hinkel 2004).

The study

The present study is a work-in-progress report based on learner data extracted from the International Corpus of Learner English (ICLE) and non-professional native data extracted from the Louvain Corpus of Native English Essays (LOCNESS) (cf. Granger et al. 2002). ICLE is one of the major written learner corpora which provides "an empirical resource for large-scale comparative studies of the interlanguage of advanced EFL learners with significantly different native language backgrounds" (Pravec 2002: 83). For the present study, four subcorpora of ICLE and LOCNESS were manually extracted: two subcorpora of the Bulgarian and German components of ICLE and two subcorpora of the American and British components of LOCNESS. The four subcorpora are fairly small, but carefully matched with regard to their argumentative nature, essay topics and size. The design of the subcorpora is presented in Table 1.

Subcorpus	**Words**	**Essays**
1.BUCLE_1	18,752	37
2.GICLE_1	19,009	47
3.BRSUR3	19,019	33
4.USSCU2	18,630	17
Total	**75,410**	**134**

Table 1. Subcorpora design

The four subcorpora were tagged for parts of speech on the basis of the CLAWS7 tagset with the help of the Wmatrix online tool (Rayson 2007). Wmatrix is an online tool which automatically calculates verb tag frequencies, thus allowing for a manual extraction of the finite verb phrases in each subcorpus. The number of finite verb phrases was calculated manually on the basis of the finite verb phrases model by Leech and Svartvik (Leech and Svartvik 1974: 73). Subsequently, concordances were run (Scott 2004) for all instances of the *–ing* verb tags in all four subcorpora; the *–ing* verb tags with a nominal or adjectival function were manually discarded from the count.

Quantitative results

Two types of measures were used to compare the frequencies of use of the progressive for the learner and native subcorpora: a normalised frequency per 1,000 words and a relative frequency of the finite progressive verb phrases in relation to all finite verb phrases. First of all, the normalised frequencies of the progressive were calculated for the British and American subcorpora of LOCNESS in view of the general differences between British and American English in terms of preference for the progressive. These are shown in Table 2.

aspect	**USSCU2**	**N/1,000**	**BRSUR3**	**N/1,000**	**LL**
progressive	81	4.3	37	1.9	**+17.73**

Table 2. Frequencies of the progressive in the British and American subcorpora of LOCNESS with significance values

The raw and normalised frequencies of the progressive in the American and British subcorpora of LOCNESS are compared with each other with statistical significance values and a '+' sign indicating overuse of the progressive in the first corpus relative to the second corpus (Rayson 2007). Similar to previous findings, the comparison shows that the progressive is more than twice as frequent in the American subcorpus as in the British subcorpus, the difference being highly significant (p < 0.0001). Therefore, it is important to distinguish between the American and British subcorpora in terms of the frequency of use of the progressive and compare them individually to non-native frequencies of use.

Table 3. illustrates a comparison between the frequencies of the progressive in the non-native subcorpora with the British and American subcorpora. The normalised frequency for the progressive in the German subcorpus is three times higher than that in the Bulgarian subcorpus and identical with the normalised frequency of the progressive for the American subcorpus. In contrast, the Bulgarian normalised frequency is slightly lower than the British normalised frequency and only a third of the American normalised frequency. The statistical significance test shows that there are highly significant differences between the Bulgarian subcorpus and the American subcorpus, as well as between the German subcorpus and the British subcorpus (p < 0.0001). German EFL learners overuse the progressive considerably compared to British novice writers, whereas Bulgarian EFL learners underuse the progressive significantly compared to American novice writers. There are no significant differences between the German and American subcorpora or between the Bulgarian and British subcorpora.

NNS corpora	**N progressives**	**N/1,000**	**NS corpus**	**N progressives**	**N/1,000**	**LL**
BUCLE_1	27	1.4	BRSUR3	37	1.9	-1.43
			USSCU2	**81**	**4.3**	**-28.61**
GICLE_1	81	4.3	**BRSUR3**	**37**	**1.9**	**+16.83**
			USSCU2	81	4.3	-0.02

Table 3. Frequencies of the progressive in the NNS and NS subcorpora with significance values

The second frequency measure consists of relating the number of finite verb phrases marked for the progressive aspect to the total number of finite verb phrases across the learner and native corpora. The percentages of the progressive verb phrases are presented in Diagram 1.

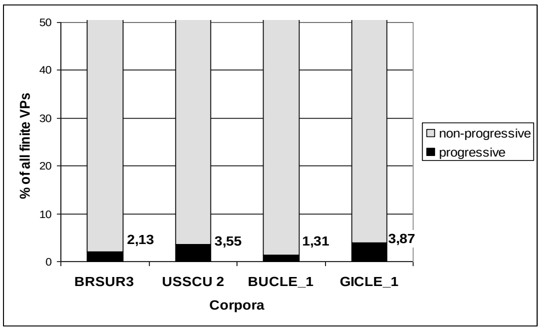


Diagram 1. Percentages of the progressive VPs in the NNS and NS subcorpora

The German subcorpus has the highest percentage of verb phrases marked for the progressive aspect of all subcorpora – 3,87%, whereas the Bulgarian subcorpus has the lowest percentage – 1,31%. Notably, the statistical significance test shows differences between the native and learner corpora at different significance levels: thus Bulgarian EFL learners radically underuse progressive verb phrases compared to American novice writers ($p < 0.0001$), whereas German EFL learners overuse progressive verb phrases considerably compared to British writers ($p < 0.01$). There are no significant differences between the German and the American subcorpus; however, there are significant differences between the Bulgarian and the British subcorpus. Bulgarian EFL learners underuse verb phrases marked for the progressive aspect significantly compared to British writers ($p < 0.05$).

Qualitative results

In the qualitative part of this study, the non-native use of the progressive aspect is evaluated by the author and a native speaker of American English in terms of the acceptability of the progressive within the temporal and discourse context of the essays. The native informant was asked to judge all finite progressive verb phrases for their grammaticality and idiomaticity and indicate whether a non-progressive verb phrase should have been used instead of a progressive. Only one example of inappropriate use of the progressive was found in the Bulgarian subcorpus:

1. You no sooner buy a new product than you **are thinking** about its replacement. ICLE-BG-SUN-0003.1

In this example, the use of the progressive was considered inappropriate due to the habitual nature of the action, where the native choice would have been the simple present tense. Considerably more instances of ungrammatical or unidiomatic use of the progressive were found in the German subcorpus; however, some of the instances were judged as more acceptable than others. Thus in the second and third example the progressive was judged as ungrammatical in terms of the habitual making of everyday situations of dreaming and watching television, where the native choice would have been the simple present.

2. When I **am dreaming** about past times I see little villages surrounded by dark woods. ICLE-FR-ULG-0002.1

3. The credibility of television seems to be unshakeable, especially when the pictures **are touching** us emotionally. ICLE-FR-ULG-0007.2

A great number of progressives without a strict temporal meaning were judged as unidiomatic and marginally acceptable. To illustrate, in the fourth example the writer refers to a habitual action of the mother’s drinking and children going their own ways, whereas in the fifth example the writer is not pleased with people calling them several times a week. These two examples convey the author’s personal feelings and negative attitude towards these circumstances and could be classified as an emotional use of the progressive; however, these uses were judged as inappropriate by my American informant.

4. The inspector is divorced, the families are ruined, the father has a mistress, the mother **is drinking**, the children **are going** their own ways. ICLE-FR-ULG-0007.2

5. I'm anxious about people who **are calling** me three times a week although I don't want to chat with them. ICLE-GE-AUG-0024.1

Other progressive cases were evaluated as unidiomatic in the context of the temporal framework of the whole essay, where the native choice would have been again the simple present. Thus in the sixth example the writer is describing a lively situation of being together with a friend discussing important happenings during the last day.

Such uses of the progressive have been previously analysed in terms of the writers' wish to convey immediacy of the situation and make it more concrete and vivid for the reader (cf. Axelsson and Hahn 2001: 26).

6. **I'm lying** on my bed, a pot of tea and a plate of biscuits are next to me and **we're talking** about all the crucially important things that happened the last 24 hours. ICLE-GE-AUG-0026.1

Undoubtedly, a strict classification of the progressive verb phrases into grammatical and ungrammatical or idiomatic and unidiomatic is difficult to carry out, partly because of the overlap of meanings in one progressive verb phrase and partly because native informants experience problems categorising non-native use of the progressive with certainty.

Concluding remarks

This study focused on some preliminary findings on non-native use of the progressive aspect in argumentative writing produced by advanced Bulgarian and German EFL learners. The quantitative results show that there are significant differences between non-native and native use in terms of the normalised and relative frequencies of occurrence of the progressive. Several findings are of particular interest and need further investigation: on the one hand, the differences between the British and American patterns of use of the progressive, and on the other, the similarities between progressive use in the German learner subcorpus and the American native subcorpus, and between those in the Bulgarian learner subcorpus and the British native subcorpus. Bulgarian EFL learners seem to follow British English patterns of use of the progressive in argumentative writing, even though a slight underuse was established in comparison with the British relative frequencies. German EFL learners seem to follow American patterns of use of the progressive in argumentative writing, as both the normalised and relative frequencies are identical or highly similar for the two subcorpora. Moreover, the German subcorpus features a similar number of progressive verb phrases with contracted forms like the American subcorpus, whereas the Bulgarian and the British subcorpora are similar in that the first one has one contracted form in a progressive verb phrase and the latter none.

These findings can be interpreted in two ways: first, the higher use of the progressive verb phrases with contracted forms in the American subcorpus indicates a higher degree of informality and a colloquial character of the American argumentative essays in contrast to the British ones (Leech and Smith 2006). Second, the similarities between the progressive use in the learner subcorpora and either the British or the American subcorpora suggest that advanced Bulgarian and German EFL learners may have been exposed to a different target language influence, since British English is the target norm in classroom teaching in both Bulgaria and Germany (Granger et al. 2002). However, more than half of the German EFL learners in GICLE have had at least one month of target language exposure in an English-speaking country in contrast to less than 10% of the Bulgarian EFL learners, which might explain the stronger conversational bias of the German essays. Unfortunately, the countries of target exposure in ICLE are unclear (Granger et al. 2002).

The qualitative results are even less straightforward in terms of classifying non-native use of the progressive as unidiomatic or inappropriate in the context of argumentative writing. German EFL learners' use of the progressive in contexts where native speakers would have opted for the simple form may be due to the progressive being employed as an attitudinal stylistic device rather than as an aspectual marker, as well as the more narrative nature of some of the German essays, which do not follow a strict argumentative pattern. Furthermore, one needs to consider the reliability of a native speaker's judgements and the possibility of involving several native informants as independent judges for an error-tagging project. However, this is a project for future research.

References

Andersson, S. 1989. "On the Generalization of Progressive Constructions: "Ich bin (das Buch) am Lesen"-status and usage in three varieties of German." In *Proceedings of the Second Scandinavian Symposium on Aspectology*, L. Larsson (ed.) Uppsala: Almqvist & Wiksell, 95-105.

Axelsson, M. and **Hahn, A.** 2001. "The use of the progressive in Swedish and German advanced learner English – a corpus-based study." *ICAME journal* 25: 5-30.

Biber, D. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Comrie, B. 1976. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems.* Cambridge: Cambridge University Press.

Granger S., Dagneaux E. and **Meunier F.** 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. 1996. "From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora." In *Languages in Contrast*, K. Aijmer and B. Altenberg (eds.). Lund: Lund University Press, 37-51.

Hinkel, E. 2004. "Tense, aspect and the passive voice in L1 and L2 academic texts." *Language Teaching Research* 8/1: 5-29.

Leech, G. and **Smith, N.** 2006. "Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English". In *The Changing Face of Corpus Linguistics*, A. Renouf and A. Kehoe (eds.). Amsterdam: Rodopi, 185-204.

Lenko-Szymánska, A. 2007. "Past progressive or simple past?: The acquisition of progressive aspect by Polish advanced learners of English." In *Corpora in the Foreign Language Classroom*, E. Hidalgo, L. Quereda and J. Santana (eds.). Amsterdam: Rodopi, 253-267.

Mair, C. and **Hundt, M.** 1995. "Why is the Progressive becoming More Frequent in English? A Corpus-Based Investigation of language Change in Progress." In *Zeitschrift für Anglistik und Amerikanistik*, 43/2: 111–122.

Mindt, D. 2000. *An Empirical Grammar of the English Verb System*. Berlin: Cornelsen.

Quirk, R. et al. (1985). *A Comprehensive Grammar of the English Language*. London etc.: Longman.

Pravec, N. A. 2002. "Survey of learner corpora." *ICAME journal* 26: 81-114.

Rayson, P. 2007. *Wmatrix: a Web-Based Corpus Processing Environment*. Lancaster University Computing Department. Lancaster: Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix [Access date 01/05/2008]

Scatton, E. A. 2000. "Bulgarian" In *The Slavonic Languages*, B. Comrie (ed.), London etc.: Routledge, 188-249.

Scott, M. 2004. *Wordsmith tools version 4.0.* Oxford: Oxford University Press.

Slabakova, R. 2002. "Recent research on the acquisition of aspect: an embarrassment of riches?" *Second Language Research* 18: 172-188.

Swan, M. and **Smith, B.** eds. 1987. *Learner English: A Teacher's Guide to Interference and Other Problems.* Cambridge: Cambridge University Press.

Virtanen, T. 1997. "The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English." In *Corpus-based Studies in English*, M. Ljung. (ed.). Amsterdam etc.: Rodopi, 299-309.

Wible, D. and **Huang, P.Y.** 2003. "Using learner corpora to examine L2 acquisition of tense-aspect morphology." In *Proceedings of the Corpus Linguistics 2003 Conference*, D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). Lancaster: UCREL**,** 889-898.

LEARNER CORPUS AND ENGLISH AS A FOREIGN LANGUAGE – AN EXPERIMENT WITH PUBLIC SCHOOL STUDENTS IN BRAZIL

Lílian Figueiró Teixeira[287]

Rove Luiza de Oliveira Chishman[288]

Abstract

One of the greatest challenges for a foreign language teacher is to make the classes meaningful and interesting for the students. A way of doing this is by changing the focus of the class, providing student-centered tasks in which the student is responsible for his learning and deriving some conclusions about the language for himself. In this paper, we present the application of some tasks in which the students of a public school in Brazil had to deal with corpus in order to improve their knowledge of English. For this experiment, the teacher organized a learner corpus composed of the students' compositions and some tasks based on concordances were conducted. The linguistic phenomenon chosen was the adjective order, since its use in English is different from the students' native language, Portuguese.

The main purposes of this study were to check to what extent a teacher can use some Corpus Linguistics resources as a matter of analyzing his students' compositions and to discover what the students' reaction to concordance based tasks is. For the corpus compiling, four groups of students from the first and second years of high school had to write letters and descriptive texts. We analyzed the most frequent adjectives found in the corpus: old, brown, favorite and a noun grade as a modifier. For most of the tasks the students had to compare some concordance lines from the learner corpus to some lines written by native speakers. During the application of the tasks, we could perceive that they were motivated in analyzing their own productions. They realized that the tasks were hard work, but that they also improved their language skills.

Keywords: Corpus linguistic, learner corpus, adjective order in English, foreign language learning, data-driven learning

A foreign language class

In this paper we report on a very different foreign language class, at least for the students who took part in these English classes. It was unusual for them, because they were used to listening to the teacher's grammatical explanations, accepting the rules and doing lots of structural exercises. For the first time, these students analyzed the real use of language, having contact with corpus, that "is a body of text assembled according to explicit design criteria for a specific purpose" (Atkins and Clear apud Granger 1998: 7). The area that compiles and explores corpora is called Corpus Linguistics.

Our main motivation was the fact that the teacher could recognize that these students, who are considered beginners, had some difficulty in learning the adjective order in English. We found occurrences such as "his actor favorite" and "a girl very special" in their writings. Our first impression was that it could be due to interference, so we decided to compile a corpus with the students' texts and analyze the occurrences using the Corpus Linguistic computational tools. Besides analyzing how these learners apply the adjective order in their texts, we also proposed some pedagogical tasks in which they had to do research and get to their own conclusions regarding language from the occurrences found in a corpus. This kind of task is usually planned for intermediate or advanced learners. Having beginner students learning in a more consciousness way is still a challenge for the professionals that support studying corpus in a foreign language class.

[287] Lílian Figueiró Teixeira is a teacher of English in a public school. She is a Master's student in the Applied Linguistics program at Vale do Rio dos Sinos University. Her dissertation is about parallel corpus and the semantics of noun compounds. She has publications about learner corpus, data-driven learning, semantic annotation and ontologies.

[288] Rove Chishman is a Professor in the PPG-LA program at Vale do Rio dos Sinos University (UNISINOS) and has been working in the research line called Language, Technology and Education. Her research interests are about computational linguistics, enphatizing semantic computing and corpus linguistics. She has published in computational linguistics journals, with papers about ontolog ies semantic lexicons.

Data Driven Learning

The Data Driven Learning (DDL) approach is proposed by Tim Johns as a way of teaching English grammar through real use of language and according to Berber Sardinha (2004) is the most consistent method that makes use of corpus. The authentic use of language that is found in a corpus is displayed by a concordance, which consists of a set of occurrences lines for a search word. The computer plays a central role for language learning and the studying is made through research oriented by the teacher. The student is the researcher who is supposed to create generalizations about the language by analyzing the data found in the concordances.

All the tasks are planned according to the corpus data, which can consist of texts produced by native speakers or by the students themselves. In Granger and Tribble (1998), they state that DDL has been used with native speaker corpora with the objective of making the learner reflect about the way people really speak or write in the target language. Instead of asking the teacher about a grammar, the students look for the uses in the corpus and then work out the generalizations by themselves, therefore developing the discovery learning in an inductive way. Considering this, we can say that the classes are student centered, but the teacher plays a very important role, as mediator and facilitator.

For Hadley (2005), it is important to present a set of consciousness-raising tasks from the first contact that the learners have with the method. Besides having to introduce the main concepts to them in a simple way, such as corpus and concordance, they must develop the identification, classification and generalization techniques. Berber Sardinha (op cit) states that identification is the first contact of the students with the concordance, when they identify some patterns or recurrence aspects. Immediately, they will classify this aspect according to their interpretation. As a concluding task, they elaborate generalizations about the occurrences. This sequence is not closed, the student can start by a generalization that he/she has found in the grammar, check if it really occurs in the data, and may come up with a new generalization.

We considered the DDL approach for the development of the pedagogical tasks, but our starting point was a corpus composed by students' compositions, which is called Learner Corpus, a topic that will be introduced in the next subsection.

Learner Corpus

Learner corpora are "systematic computerized collections of texts produced by language learners" (Nesselhauf 2004: 127). There are practical reasons why someone, as a teacher, decides to compile a corpus with his/her students' productions. Instead of just checking the writings, marking the students mistakes' in red and giving it back to the them, we can use these texts as a pedagogical instrument. The teacher can collect the students' writings in order to discover in which language aspect they have some difficulty and afterwards provide some specific practice. It is also possible to store their productions, made at different moments, and check how much they have improved. These texts must be computerized, since through some tools we can carry out automatic searches and visualize the concordances. If the students do not have access to computers at school, they can type them at home and give the floppy disc to the teacher or the teacher can type each composition and store it in a file. Another application for learner corpus is the project called "International Corpus of Learner English" (ICLE) whose objective is to collect texts written by English learners from different countries. They are intermediate and advanced students that are asked to write essays about some given topics.

In order to check how the learner really makes use of the language, we should follow some criteria for the selection of the texts that will constitute our learner corpus. According to Nesselhauf (op cit) it is not possible to find natural texts in the foreign language environment, but we should provide conditions in which their texts are the most natural possible. So we should propose writing tasks under very low control to obtain our data for the corpus, and discard writings guided by picture descriptions or translations. Considering the size of the learner corpus, it is not necessary to be very big. The most important point, for Ferreira (2003), is the corpus authenticity and that its organization can provide the data we need for our purposes.

Other criteria must be considered when compiling a learner corpus, such as the following ones suggested by Granger (op cit) concerning the language and the learner. We should specify if what we collect are samples of oral or written language, its discursive genre, the task subject and the setting. Among the criteria related to the learners, we have to set their age, gender (male or female), their mother language, level of knowledge of the language and the learning context, as foreign of mother language.

In our experiment, a learner corpus was organized by the teacher with the objective of checking how the students organize the adjectives in English and also to provide some data to serve as a basis for the elaboration of some pedagogical tasks.

Adjective order

By consulting a grammar book for beginners (Murphy 1998), the basic instruction regarding the use of the adjectives in English concerns its order in relation to the noun: the adjective is placed before the noun in the attributive position. If there is more than one adjective describing the same noun, different authors propose different alternatives for organizing these adjectives. While Celse-Murcia and Larsen-Freeman (1999) suggest that the order is the following: determiner, opinion, size, shape, condition, age, color, origin and noun, for Alexander (2002), the most appropriate order is : modifiers (as big and little), quality/opinion, size/age/shape/temperature/taste, etc, nationality, the ones formed by past participle and nouns. These grammar books suggest an adjective order without considering the real use of language. In a corpus based grammar book, as the Collings Cobuild (Sinclair, 1990), the adjectives are organized in qualifiers and classifiers, and their order suggestion is confirmed through corpus occurrences, it being: determiner, post-determiner, emphasizer, qualifier, color, classifier, noun modifier and noun. If there is a noun functioning as a modifier, such as in “cat food” or “vinyl coat”, it is placed right before the noun it is modifying.

In Portuguese, the most common attributive position of the adjective is after the noun. When there is more than one adjective, Neves (2000) states that the order is the following: qualifier, noun and classifier. Sometimes a noun is used as a modifier too, for example “bomba relógio”. Her grammar, just like Sinclair’s is also based in corpus data, and both classify the adjectives in the same way. The qualifiers identify the qualities of someone or something and they can be graduated and intensified. An example would be “sad”, since we can say that someone is very sad or that someone is sadder than another person. As examples of classifiers, we have “financial” and “Cambodian”, since they identify someone or something as a member of a group. We can say that “financial interests” are a kind of interests, and this type of adjective is not graduated or intensified.

Comparing the use of the two languages, it seems that very simple explanations about the adjective order in English would be enough for the Brazilian students to learn how to use it. When we read their written productions, we see that the students that took part in this experiment seem to have many doubts about the adjective order. The analysis results are displayed in the next section.

Pedagogical tasks

This experiment was conducted in a public school in Rio Grande do Sul, Brazil. Since fifth grade of the elementary school, the students have foreign language classes, and in most of the schools, English is the subject. The priority according to the legislation is to develop the reading skill in these classes. So working with corpus based tasks collaborates to this purpose. The application was divided into two different moments, first the learners had to write some compositions and later they performed a series of tasks based on the corpus compiled from their texts.

The groups that took part in this experiment were composed of four classes, 106 students from the first and second grade at high school. They were all teenagers who studied during the afternoon and only had formal instruction of English at school. Their English classes had a time duration of around one hour and a half hours per week. In the next subsections each step of this application is explained in more details.

Compiling a Learner Corpus

As a motivation for the writing task, the teacher proposed a pen friend project, in which the students had to write letters to students from another class with the objective of making new friends. After receiving an answer, they had to write a report showing their first impression about the new friends. Even though they had been studying English for more then four years at school, they can still be considered beginners, since they are not used to writing texts or speaking in the foreign language. These two genres were chosen because they had just learned how to describe themselves and other people. During the classes, the teacher could perceive the students’ doubts mainly about how to use the adjective to describe someone’s eyes or hair.

The students did not have access to computers at school, so the teacher typed all the compositions and stored them in a txt file, since this format is a requirement for the concordance tool. The students’ names were substituted by XXX in order not to embarrass them during the exposition of the concordance lines. Most of the participants were less than 18 years old, so their parents had to sign a consent term in order to allow the utilization of the texts for the research. This consent reduced the number of texts that could be included in the learner corpus. In the figure below, some data about the compiled learner corpus is summarized:

Number of letters:	49
Number of descriptions:	52
Total number of texts:	101
Number of types:	732
Number of tokens:	8,629

Learner Corpus Data

Data analyses

For the analyses we made use of two specific tools, a wordlist and a concordancer. These tools are free and available at the LAEL website, at http://lael.pucsp.br. After checking the wordlist and the number of occurrences for each word, we selected the three most frequent adjectives, which are: old (91 occurrences), brown (72 occurrences) and favorite (62 occurrences). Another interesting aspect that we could identify in the corpus data is the use of a noun modifier, grade (16 occurrences). Since it was the only case of a noun functioning as a modifier, we decide to include it in the analyses.

As suggested by Nesselhauf (op cit), we decided to compare the learner corpus occurrences to some concordance lines produced by native speakers. We made use of the concordances available on-line from the British National Corpus[289] (BNC) and the Bank of English[290] (Collins Cobuild project). The Cobuild website provides 40 concordance lines and the BNC, 50 lines; both visualizations are free and at random.

By comparing the concordance lines written by the students with the ones produced by native speakers, we could identify some subuses and overuses. Ferreira (op cit) explains these two concepts. For subuse, the items that the learner uses less frequently than the native speaker, some occurrences are the compound adjective to inform someone's age, just like "a 41-year-old teacher", and the noun modifier, such as "old submarine movie". The overuse is the opposite, the learner uses some items in a bigger proportion than the native speaker does. In the learner corpus, most of the time (65%) when the students described someone's hair, they used more than two adjectives, while in the native speaker lines it happened only 10% of the cases at the Collins concordance lines. At the BNC sample, it was not possible to find more than two adjectives modifying the same noun. There are some cases of interference from their native language, since they keep the same adjective order as Portuguese. Some examples are: "eyes brown", "my band favorite" and "the favorite movie his".

Tasks application

After the analysis made by the authors of this paper, we elaborated some tasks to be applied with the same students that produced the texts. These tasks were conducted by their teacher over the period of a month. The activities consisted of: questions about the uses in the learner corpus, tasks in which the students had to compare their uses with the ones provided by the concordance lines from the native speaker corpora, questions to stimulate the creation of generalizations about the language and some extra exercises to practice the linguistic phenomenon. Since there were no computers available at school, all the tasks were printed and handed to the students. Most of the time, they worked in pairs and could consult their books and notebooks. After each task, the pairs revealed their conclusions to the big group. It made them feel more confident about their answers. In the first class, the teacher presented the concordance lines from the learner corpus for "old" (see below) and explained to them how to read the concordances. We decided to follow Bertóli-Dutra's suggestions (2002) avoiding to use the term "concordance" with them, since it could confuse them. Instead of it, we referred to the concordances as "sentence fragments". They were also oriented to start reading the lines from the middle to the left or to the right. It is important to start with the keyword, which is in bold, and check what the words around it are.

1. what ' s your name ? how **old** are you ? what do you like
2. what do you look like ? how **old** is he ? what ' s your
3.is your favorite singer ? how **old** are your ? what ' s your
4. º year and you how mach year **old** , that year this , where in
5. eyes . i am a sixteen yeares **old** , i am 1st year student .

[289] Available at: <http://www.natcorp.ox.ac.uk/>
[290] Available at: <http://www.collins.co.uk/corpus/CorpusSearch.aspx>

6.udent and i ' m sixteen years **old** . i like music and food .
7.eyes . she has sixteen years **old** and 1st years student .
8.cause she have sixteen years **old** , she likes music and
9. thin . she is fifteen years **old** . she has long stratraught
10.is xxx , he is sisteen years **old** , he have black hair and bleck
11.   , she his fiveteen years **old** , and she study in the

Some concordance lines for "old", learner corpus data

According to Hadley's report (op cit), the concordance lines should be simplified, since the students can get a little confused when the concordance is very big and there are many new words. We decided to select only some lines, avoiding repetitions of the same uses, and we also grouped the uses that were similar. Even by doing this and informing the students that it was not necessary to understand every single word in the concordance lines, they said that the tasks were a little confusing because there were many words they did not understand. On the other hand, they were really motivated about analyzing their texts through the concordances. Without being asked to check the sentences, the students had some fun looking for the mistakes and imagining who would have written them. When asked about the positive points in the experiment, they stated that they could learn more about the language and also appreciated the moments in which they interacted with the classmates.

Impressions

We realized that it is possible to analyze the language through the Corpus Linguistic resources and that the DDL tasks proved to be positive for beginner students, since they were more motivated and more independent in relation to their learning. In order to develop the learner's researcher posture, we believe that more applications like this should be proposed. The more the students get used to these resources; less controlled tasks can be planned for them. In the future the students may be able to manipulate the concordances directly. We also point out that the DDL tasks were only a part of the classes, in a good class there should be different methods which can stimulate various students' skills.

References

Alexander, L. G. 2002. *Longman English Grammar.* New York: Longman.

Berber Sardinha, A. P. 2004. *Lingüística de Corpus.* São Paulo: Manole.

Bertóli-Dutra, **P.** 2002. *Explorando a lingüística de corpus e letras de música na produção de atividades pedagógicas.* São Paulo: PUC-SP. Master's dissertation, LAEL, Pontifícia Universidade Católica de São Paulo.

Celce-Murcia, M. and **Larsen-Freeman, D**. 1999. *The grammar book.* 2.ed. Boston: Heinle & Heinle.

Ferreira, W. R. 2003. *Deslexicalização no inglês de alunos brasileiros: um estudo baseado em corpora de aprendizes.* São Paulo: PUC-SP. Master's dissertation, LAEL, Pontifícia Universidade Católica de São Paulo.

Granger, S. (org.). 1998. *Learner English on computer.* New York: Longman.

Granger, S. and **Tribble, C.** 1998. "Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning." In *Learner English on computer,* S. Granger (org.). New York: Longman, 199-209.

Hadley, G. 2005. *Sensing the winds of change: an introduction to Data-Driven learning.* http://web.archive.org/web/20030404025654/web.bham.ac.uk/johnstf/winds.htm [Access date 02/04/2005]

Murphy, R. 1998. *English Grammar in Use.* 2.ed. Cambridge: Cambridge.

Nesselhauf, N. 2004. "Learner corpora and their potential for language teaching." In *How to Use Corpora in Language Teaching,* J. M. Sinclair. Amsterdam/Philadelphia: John Benjamins Publishing Company, 125-152.

Neves, M. H. de M. 2000. *Gramática de usos do português.* São Paulo: UNESP.

Sinclair, J. 1990. *Collins Cobuild English grammar.* Londres: Collins.

IMPATIENCE IS A VIRTUE: STUDENTS AND TEACHERS INTERACT WITH CORPUS DATA – NOW

James Thomas[291]

Abstract

The corpus-in-language-education fraternity is waiting for empirical evidence that will support their instincts that learners gain important knowledge, skills and experience through direct involvement with corpora. This paper argues that much of the language pedagogy research that has already seen the widespread adoption of what are now standard practices in ELT can be used to support these instincts. Too much attention has been paid to linguistics at the expense of pedagogy.

This paper describes a wide range of activity types that are pedagogically sound, piloted and acclaimed by a wide range of students and teachers, though admittedly not in vast numbers. There is no doubt that all of the activities and their over-arching approaches could be enhanced and improved, and they need tailoring for different circumstances, something at which teachers are generally adept..

The essential point is that waiting for research is tantamount to marking time. There are some motivated students and teachers equipped with open, enquiring minds and some classrooms equipped with budgets, hardware and software. There is, of course, no intention to convert anyone forcefully. Nor is there any need. Robert Heinlein's sage advice comes to mind: "Never try to teach a pig to sing; it wastes your time and it annoys the pig."

Keywords: ELT methodology, task-based learning, beginner to advanced, corpus building, academic writing

Introduction

This paper describes several ongoing and inter-related attempts to develop teaching procedures for using corpora and concordancers as a tool for EFL courses including the training of pre- and in-service teachers in corpus use. It might even be described as research though not in the quantitative, qualitative, statistical research paradigms in vogue today. The author is an ELT teacher trainer in general methodology as well as in ICT4ELT, and has long been a practising advocate of data-driven learning (DDL) and other corpus-based pedagogical procedures. The attempts described in this paper involve working with elementary adult students, with ELT teacher trainees and with post-graduate students at the Faculty of Informatics (FIMU) where he is based. One idea permeating this paper concerns the need for classroom corpus application to be considered in the light of current thinking on classroom practices, loosely known as *methodology*.

FIMU happens to be the home of the Sketch Engine (Kilgarriff and Rychly, 2004), through which most of these procedures are being piloted, partly thanks to our free access, but in no small part, thanks to the software itself. Being above all a lexicography tool, it is actually well-suited to a view of language that is lexically, rather than structurally based. After all, it is from the newly-weds, lexis and syntax, that the core linguistic aspects of much modern language pedagogy have sprung. See for example, the ELT authors, Michael Lewis (The Lexical Approach), Scott Thornbury (Natural Grammar), Dave Willis (Cobuild English course) and Paul Nation (Learning Vocabulary in another Language).

The web-based Sketch Engine returns concordance pages, lists of collocates, comparative collocates, and bi and tri grams. Importantly, it also generates *word sketches* for nouns, verbs and adjectives – these show the collocates of words categorised under their colligations – lexico-grammar *par excellence*. Within the Sketch Engine package, corpora can be created from one's own resources using Corpus Builder or automatically compiled from the word wide web using WebBootCaT.

Affiliated software, (Kilgarriff, et al 2008), will even generate a page with one algorithmically chosen illustrative sentence for each salient colligation-collocation of the search word. As the authors say, "This can be seen as a

[291] James Thomas has an MA in applied linguistics from Macquarie University, Australia. He is based at the Faculty of Informatics, Masaryk University, Brno in the Czech Republic, working on various corpus projects, database programming and teacher training in ICT for ELT as well as corpus use.

halfway house between confronting learners directly with concordances, and using them indirectly for dictionary-making." (p.6)

However, not being designed for pedagogical purposes, the Sketch Engine itself still needs more features useful to teachers and learners. For example, there are no options for creating printouts with line numbering or increased vertical spacing. Showing complete sentences in KWIC is not possible. Nor is it possible to select specific lines. A useful feature for teachers and students would be the ability to save searches for later retrieval, as would logging and recording student searches. Replacing the node with a lined gap would assist in making exercises. Most of these features, and many more, are available at the pedagogically conceived Compleat Lexical Tutor site (Cobb e.g .1999) where, unfortunately, search options are few and the corpora are small.

Bothersome is the hunch that those interested in proselytizing the direct use of corpus data in the classroom are working primarily from a monotheistic standpoint of linguistics, rather than incorporating it into a multi-faceted model which includes lessons from second language acquisition and an understanding of the school environments in which the new language teaching procedures are to be deployed.

Thus, the pedagogical contributions and implications of Vygotsky, Bloom, Task-Based Learning, Communicative Language Teaching, the Lexical Approach and a host of others that constitute contemporary language teaching need to brought to the fore in order to align corpus based procedures with current classroom practices. As long ago as 1998 Thornbury made the point that teachers are unlikely to be interested in a set of pedagogical principles per se: it is only when the same principles can be applied to classroom situations that their worth is evident.

Thus, a long term primary aim of my classroom experimentation is to develop a pedagogically and linguistically sound teaching approach that can be demonstrated to pre- and in-service teachers.

With none of my students Luddites, the hardware and software in place, and the courses flexible, the conditions are ideal for experimenting with new approaches. At their own pace, the trainees develop confidence in performing searches and selecting and using the relevant analytical tools; they interpret data on the fly, even if not always successfully – more about which below. And they use their results in the teaching resources they create.

A demonstration lesson, i.e., teaching regular students in front of trainees, is a more effective way of training future teachers in a new approach than having them read about it in a book or journal. As ELT trainees they will eventually experiment with their own applications of DDL: as they are not necessarily sworn disciples, their feedback is much anticipated.

It is clearly not the intention of this paper to demonstrate, let alone "prove", that superior language acquisition results from corpus consultation. In his survey of research papers, Boulton (BAAL 2007) finds that on average, only two studies which report some kind of evaluation of DDL beyond the researcher's opinion have been published each year since 1991. Given this dearth of research, the current application of corpus tools described here is targeted rather at the consultation of corpora for a variety of ends for both students and teachers. These include:

1. answering set questions
2. training in corpus consultation
3. creating "word profiles"
4. selecting best sentences
5. developing greater linguistic awareness

And in the specific case of the post-graduate students, observing the use of taught language features of academic writing and terminology in a home-grown field specific corpus. Three of these are discussed below.

Answering set questions

A group of adult elementary students, ICT school teachers in fact, recently searched the Cobuild Corpus Sampler (CCS) to sort out the uses of *much, many, a lot of.* The session began with eliciting what some of them already knew and we proceeded to verify this information with the data available. When exceptions were encountered, the "rules" were modified. The end product was their contribution to a wiki of their findings for the others to comment on and refine.

The students had no introduction to corpora per se, how to perform searches or what to do with the resulting concordances. But they managed to look at the data and answer leading questions, which led them to the desired results.

Other questions that can be answered similarly include:

1. Gap filling exercises in texts books can be checked with corpus data. For example:

   By the time we ____ to the station, the train had already left.

   arrived reach get find

2. What differences appear between *not too* and *not very*?
3. Which *sports* are played at/on/in *fields, courses, pitches, courts*?
4. Which prepositions are used after certain words?
5. What verb forms follow verbs such as *keep, manage, allow*?
6. Compare the use of *in the end* and *at the end*.
7. What do you notice about the lists of collocates of *take care of, look after, care for*?
8. What do the contexts of *fat, plump* and *overweight* reveal about the usage of these three words?
9. Determine which of the meanings of *manage* is meant in each sentence. How did you arrive at your answers?
10. Is *like* in *very like* a verb? How is this verb usually intensified?
11. If *mushroom* is a verb, what types of things are its subject? Think about the mushroom metaphor.
12. Find ten parts of the body which can be used as verbs. How are they used as verbs?

Working in groups, students may then be asked to select the most suitable illustrative sentences, and then justify their choice. In Moodle, classes create a Glossary of words and phrases, and include these sentences. Sometimes they come to the front and present their group's findings. They make mini PowerPoint presentations and present them to the class. As these tasks demonstrate, there is no need for a corpus of easy English, or too much pre-processing of corpora: it is the task which is adjusted, not the language, as David Nunan (2004) has it.

Many communicative language teachers behave somewhat like the Australian male bowerbirds, famous for elaborately decorating their nests with *objets trouves* so as to attract a mate. Such teachers find that the rich palette of available approaches offers them procedures which they can apply to their own situations in the hope of achieving their goal of developing independent learners capable of framing questions which they can in turn answer themselves.

> Writing of future lexicographers, Hanks (forthcoming) says: …
>
> instead of asking, How many meanings does this word have, and how shall I define them?" the lexicographer will start by asking, "How is this word used, how can I group them into patterns, and what is the meaning of each pattern?

The process of sorting and grouping has long been used in vocabulary teaching. To elevate it to Hanks' level in the classroom is not beyond many of the intelligent, enquiring, computer-literate students we teach today. Students are not unfamiliar with problem solving or dealing with data. If nothing else, such an activity becomes a superb detective game for the curious student! And what they gain in the process, is an understanding of how language works that equips them to make better choices.

Hyunsook Yoon (2008) reports finding that corpus use helped her six students solve writing and language problems as well promoting their perceptions of lexico-grammar and language awareness. Furthermore, the students assumed more responsibility for their writing. One likes to believe that guided interaction with language data will have an enduring impact on learners but the reality is that casual encounters rarely lead to life-long relationships, no matter how glamorous or gratifying the encounter may have been. This applies no less to looking up words in a dictionary than it does to a corpus search.

Some of the above corpus activity questions are typical of those that arise in the process of writing and speaking, as some of my Think Aloud Protocol teaching activities have shown. These involve pairs of student working on a writing task in which the writers express aloud the quandaries they are resolving as they write; the other student simply notes them for later whole class discussion. Some of the above questions derive from such activities and corpus-based investigations ensue.

As the following table illustrates, it is possible that unexpected word combinations do occur. It is the significance of significance that the learners need to grasp. While this is well-known to linguists, learners need to experience it for themselves. Before seeing this table, students receive a copy without any numbers and are asked to tick the cells where they expect to find matches. This is a standard prediction activity common in many an ELT lesson. They then perform the searches and record their findings. Students do not need to look very far to realise why these unlikely combinations do in fact exist, but in doing so they gain a fundamental understanding of the probabilistic nature of language. They grasp the significance of significance.

Time words from the BNC: the numbers across the top and in the left column are the number of times that each word or phrase occurs. The numbers in the body of the table represent the co-occurrences.

	morning **20,020**	**afternoon** **8,027**	**evening** **13,223**	**night** **34,976**
last **70,573**	11	21	64	8475
yesterday **19,344**	345	196	90	4
this **454,440**	4082	1703	1083	85
tomorrow **8,893**	411	100	70	447
in the **516,719**	3691	971	1197	585
next **42,221**	1234	31	57	86
at **466,110**	22	2	20	3034

Table 1

Word Profiles

Just as a spy builds up a dossier on people and organizations in order to determine their typical behaviour and even predict their future behaviour, a language learner can do likewise for words and phrases. Like the spy, numerous thoughtfully observed encounters of many elements need to be recorded. And these records I refer to as Word Profiles.

The number of elements increases with ever more linguistic research. Where once the meaning, spelling and pronunciation of a word sufficed, the checklist now includes collocates, phrases, register and connotations. It may well include semantic prosody, domain, as well as semantic relations such as synonyms, hyponyms, antonyms. And of course it must contain colligation, for what learner of English could use the word *priority*, for example, without knowing such frames as this?

s.o./s.thg takes priority over s.o./s.thg.

Learners need to know what they need to know (sic) about a word in order to use it. Without providing “knowing a word” training, an important aspect of the “role of the teacher”, learners do not usually know what they need to know. Increments in the depth of word knowledge are part of the inevitable recycling process that results from extensive exposure to natural language (Krashen, e.g., 2004). Corpus assisted word profiles assist learners develop this depth systematically.

One of the course tasks required of the teacher trainers on the MU course, *Using Corpora in ELT*, is the creation of Word Profiles of particular words. DDL is generally more suited for depth of knowledge rather than for learning new items (Cobb 1999). Using the wiki tool in Moodle, course participants add data to their word profiles throughout the semester as they are introduced to new features. They choose from a list of General Nouns (Michaela Mahlberg, 2005) and for each one they add the results of their findings each time they learn new tools in the Sketch Engine such as Frequency, Word Sketch, Sketch Differences; and likewise for new linguistic concepts such as colligation, lexical support and semantic prosody.

Such an accrual of criteria has gone some way to ameliorating earlier Word Profile problems in which students seemed to regard frequency as the sum of all that could be gleaned from the data. In future, the end products need to be edited and published for the whole class, with an introduction and some commentary composed jointly.

Acquiring new metalinguistics concepts empowers students and teachers to make better choices when they deploy language. This may or may not require a corpus search for the answer. They develop a feeling for knowing what questions to pose and which ones are answerable given the resources at hand.

Creating a corpus

For the corpus building project with the post-graduate students at FIMU, Corpus Builder (CB), was modified so that each section of the academic papers would be tagged. In building the corpus, the students collected a series of research papers that represent their specialisation and save them as text files. After logging in to CB they entered metadata for each file; next they pasted each section of the text file into the appropriate tagging field, such as Abstract, Methodology, Future Work and Other. The corpus from an earlier attempt at this project was poured into the current one, and even though its sections are not tagged, such a volume of additional data proves invaluable for a wide range of explorations. On a specific date, uploading was closed so that searches would return the same results each time. This is particularly helpful for teachers preparing demonstrations and for setting students replicable discovery tasks.

Some of the language phenomena introduced in the academic writing course includes noun clauses, linking, hedging, emphasis, sexist language and the use of first person. In preparing these lessons, it was common to locate exponents of these phenomena in both the BNC and the students' Informatics Corpus. This was often useful as the main course textbook (Side and Wellman, 2002) illustrates everything with invented sentences. As Harwood (2002) writes: "the teacher will wish to consult the appropriate corpora to avoid the *ersatz* English of the textbooks which reflects little of the language's lexical variations and predominant patterns".

For example,.

It frightens me that there are so many criminals around. (Fronting, it + clause p.200)

I'd be very happy to be of any assistance. (Noun clause, after some adjectives p.163)

Because I'll be in tonight, I'll baby-sit. (Linking clauses, Time and Reason, p.102)

The first tier in the end of course assessment task required the students to locate three illustrative sentences for any five language phenomena. The task itself requires that they understand the concept to an extent where they are able to locate sentences and then choose the most illustrative of them according to a set of criteria discussed earlier in the course. Here are some examples of their findings representing the same concepts as illustrated above. Most come from their Informatics Corpus.

Fronting

That this will not do is what I have just argued. (BNC)

It is only relevant which of them is lowest .

Noun clause

For example, it is easy to see how to make one Svetlichny box using two XYZ boxes.

The problem is that the user has to trust the owner of the computing resource where her code will be executed.

Linking clauses, Time and Reason

> Users may store, update, and remove their credentials in the repository at will, after first authenticating to prove ownership of the credentials. (IC)
>
> Having four vertices, there are only 16 cases for which a given vertex is either larger or smaller than the isovalue.

It is motivating and strangely convincing for the students to see the language phenomena exemplified in the articles they personally contributed to the corpus. This is particularly the case where something would be salient in either a sub-field or a tagged section. The Sketch Engine demonstrates this clearly.

As the final task, the students are writing a 750 word text which demonstrates their use of as many of the target phenomena as possible. For the sake of this paper it is unfortunate these texts are not yet due. However, even if they had been submitted, marked and analysed, they would be incapable of serving as a research instrument evaluating the impact of corpus use: there is a plethora of other influences on their writing by this stage. Given this fact, it seems unlikely that anyone will ever be able to demonstrate the linguistic contribution corpus consultation makes to learner's linguistic development.

Referring to her survey of twelve papers researching DDL Chambers' (2007: 5) states, "it is worth asking why there are not more large-scale quantitative studies". And Boulton (2007), concludes his "meta-paper" with: It is at the least ironic that empirical evidence should be so lacking in a field relating to corpus linguistics, where the nature of evidence is crucial.

I can only conclude by comforting those despairing over the lack of empirical evidence with my decade-old tales of student delight and satisfaction with linguistic insights, and with the acquisition of skills and tools for life. No claim is made here that DDL is every language learner's panacea – their learning styles, attitudes towards computers and trusting themselves in data interpretation determine their comfort with DDL procedures. Every teacher has to juggle a wide range of variables in any class with more than one student.

I would like to conclude by suggesting that no amount of favourable empirical evidence is going to convince teachers or publishers to involve themselves in the use of corpora. Its proselytization depends on having the tools and techniques used by more teachers. This vicious circle can be broken with the involvement of teacher training institutions, which in turn requires corpus advocates to talk in their language of tasked-based learning, constructing knowledge, discovery and inductive learning, learner training, the role of the teacher and the like. Corpus activities are fun, challenging, hands-on and engaging. Empirical evidence is not needed to convince chocoholics to devour their first Godiva chocolate.

References

Chambers, Angela. 2007. "Popularising corpus consultation by language learners and teachers." In E. Hidalgo, L. Quereda & J. Santana. (eds) *Corpora in the Foreign Language Classroom*. Rodopi: Amsterdam, 3-16.

Boulton, A. 2007 "But where's the proof? The need for empirical evidence for data-driven learning." BAAL Proceedings

Cobb, T. 1999. "Breadth and depth of lexical acquisition with hands-on concordancing." *CALL* 12/4: 345-360.

Hanks, P. (forthcoming) "Analyzing the Lexicon: Norms and Exploitations"

Hanks, P. (forthcoming) "The impact of corpora on dictionaries", a chapter in a book on Corpus Linguistics edited by Paul Baker

Harwood, N. 2002. "Taking a lexical approach to teaching: principles and problems". International Journal of Applied LInguistcs, Vol 12 No 2, 2002 (139 – 155

Hyunsook Yoon 2008. "More than a linguistic reference: the influence of corpus technology on L2 academic writing. Language learning and Technology". June 2008 Vol 12 No.2 pp.31-48

Johns, T. 1991. "From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-driven Learning". In *CALL Austria*, 10, p. 14-34.

Kilgarriff, A., Rychly, P.., Smrz, P. and Tugwell, D. 2004. "The Sketch Engine" EURALEX 2004, Lorient, France

Kilgarriff, A. , Miloš Husák, Katy McAdam, Michael Rundell, Pavel Rychlý 2008. “Automatically finding good dictionary examples in a corpus”. Paper to appear in Proceedings Euralex 2008, Barcelona, July.

Krashen, S. 2004. “The power of reading : insights from the research” Heinemann

Lewis, M. "The lexical approach: The state of ELT and a way forward”. LTP Teacher Training Series Language Teaching Publications, Hove, UK

Mahlberg, M. 2005. “English General Nouns, A corpus theoretical approach”. John Benjamins

Nation, I. S. P. 2001. “Learning vocabulary in another language”. Cambridge: Cambridge University Press.

Norris, R. 2004 “Ready for first certificate: coursebook”. Macmillan Education

Nunan, D. 2004. “Task-based Language Teaching”. Cambridge University Press.

Side, R., and Wellman, G. 2002. "Grammar and Vocabulary for Cambridge Advanced and Proficiency"**.** Pearson Education Limited

Sinclair, J.M. 1991. “Corpus, Concordance, Collocation. Oxford University Press

Thornbury, S. 1998. “The lexical approach: a journey without maps?” Modern English Teacher 7.4: 7-13

Thornbury, S. 2004. "Natural Grammar" Oxford University Press

Willis, D. 1990. **“** The lexicall syllabus : a new approach to language teaching”. Collins ELT.

COMPARING TERMS OF TOURISM IN ESP TEXTBOOKS, LEARNER'S DICTIONARIES AND CORPORA IN ORDER TO BUILD A TOURISM ONTOLOGY

Patricia Tosqui-Lucks[346]

Abstract

In an attempt to fulfill the needs of a student of Tourism Graduation Courses to master the basic vocabulary of tourism and to contribute to the semantic-conceptual study of the lexicon, this poster presents an ontological structuring of the basic vocabulary of tourism which, on the one hand, constitutes a linguistic and pedagogical resource and, on the other hand, can be integrated to specific lexical data bases. To accomplish this goal, we have proceeded to the following tasks: (i) describing the tourism sector conceptually by carving its main concepts into a domain tourism ontology, (ii) delimitating the basic vocabulary of tourism, and (iii) anchoring the lexical items that constitute the vocabulary in English and Portuguese to the tourism ontology concepts. In order to do so, we have extracted the terms, restricted to nouns, which are more frequent in 5 textbooks of English for Tourism students. After that, we have compared all the information about them (definition, explanations, examples, use and usage notes, etc.) to the information about the same terms found in lexicographic works designed for learners and organized onomasiologically. In order to complement the research, we have consulted the on-line sample version of the Bank of English. The conclusion presents the main concepts for the Tourism ontology, which are: tourist, motivation, tourism business, destination, travel, transportation, accommodation, attractions and activities. For each concept all the terms in English and Portuguese were provided, consisting in a bilingual vocabulary. The semantic anchorage of the English and Portuguese vocabularies to the basic concept ontology built for the tourism domain resulted in a total of 267 terms.

Keywords: English for specific purposes; textbooks; learner's dictionaries; tourism; ontology

The purpose of building an ontology

According to Römer (2002:60), textbooks can prove a useful resource if one wants to find out more about the status of authenticity in ELT. The author states that it would be helpful to look carefully at textbooks to check out the input pupils really get in their English lessons. From this point of view, we believe that students of English for specific purposes need to learn in their classes the main terms in the area of study or work exactly the way they are actually used in authentic situations, with all their uses and particularities. In an attempt to fulfill the needs of a student of Tourism Graduation Courses to master the basic vocabulary of tourism and to contribute to the semantic-conceptual study of the lexicon, this poster presents an ontological structuring of the basic vocabulary of tourism which, on the one hand, constitutes a linguistic and pedagogical resource and, on the other hand, can be integrated to specific lexical data bases. To accomplish this goal, we have proceeded to the following tasks: (i) describing the tourism sector conceptually by carving its main concepts into a domain tourism ontology, (ii) delimitating the basic vocabulary of tourism, and (iii) anchoring the lexical items that constitute the vocabulary in English and Portuguese to the tourism ontology concepts. In order to do so, we have extracted the terms, restricted to nouns, which are more frequent in 5 textbooks of English for Tourism students: Stott & Buckingham (2000), Stott & Holt (1991), Wood (2003), Dubicka & O'keeffe (2003) and O'Hara (2004). After that, we have compared all the information about them (definition, explanations, examples, use and usage notes, etc.) to the information about the same terms found in lexicographic works designed for learners and organized onomasiologically. The 4 Learner's dictionaries are: *Cambridge Word Routes English-Portuguese* (2000), *Longman Lexicon of Contemporary English* (1981), *Longman Essential Activator* and *Longman Language Activator*. Word Routes was chosen to provide the information in both languages (English and Portuguese) and the *Longman* dictionaries were

[346] Patrícia Tosqui Lucks has a Bachelor Degree in Translation (English, Spanish and Portuguese) by UNESP – Sao Paulo State University. She holds a Master Degree in the area of Lexicography, in which she researched modal adverbs in bilingual learner´s dictionaries (English-Portuguese) in 2002, at UNESP – Sao Paulo State University. She has been an English teacher in Brazil for over 10 years. Since 2003, she has been teaching English for the under graduation course of Tourism at UNESP Sao Paulo State University. In 2007, she obtained her PhD Degree in Linguistics, at the same University, in which she proposed a Tourism ontology anchoring a bilingual vocabulary for pedagogical purposes. Parts of this thesis are presented in this article. Her main research interests are bilingual learner´s dictionaries and English for specific purposes.

chosen because the three of them are based on a British and American written and spoken corpus, the *British National Corpus*, and on a learner corpus, the *Longman Corpus Network.* In order to complement the research, we have consulted the on-line sample version of the British National Corpus.

The corpus selection

To build the textbook corpus, we have extracted the terms, restricted to nouns, which are more frequent in five textbooks of English for Tourism students: Stott & Buckingham (2000), Stott & Holt (1991), Wood (2003), Dubicka & O'keeffe (2003) and O'Hara (2004). The following areas were sellected for analysis:

1. Jobs
2. Making reservations / taking a booking – flight and hotel
3. Asking about time and timetables
4. Giving tourist information
5. Food and drink
6. Means of transportation
7. Describing a hotel/ Hotel facilities
8. Giving instructions
9. Dealing with money
10. Telephone: answering, giving information, taking messages
11. Correspondence
12. Email, Internet and IT
13. Car rental
14. Dealing with complaints / problems
15. Attractions and activities / Describing tourist attractions
16. Package holidays
17. Resort
18. Safari
19. Cruise ships
20. Winter holidays
21. Health and safety
22. Air travel
23. Conference and events
24. Cultural traditions / history and folklore / festivals
25. Explaining plans and itineraries
26. Ecotourism
27. Job interviews / CV
28. Marketing and promotion / brochures
29. Weather forecast
30. Create a tourism development plan

After this previous selection, we have consulted the most important English terms n dictionaries to see how they present their definitions. The four learner's dictionaries are: *Cambridge Word Routes English-Portuguese* (2000), *Longman Lexicon of Contemporary English* (1981), *Longman Essential Activator* and *Longman Language Activator*. These dictionaries were chosen because they are based on well reputed corpora, the British National Corpus and the Longman Corpus, a 10 million word computerized database made up entirely of language written by students of English. The Longman Learners' Corpus offers so much invaluable information about the mistakes students make and what they already know, that it is the perfect resource for lexicographers and material writers who want to produce dictionaries and course books that address students' specific needs.

Using corpus linguistics to check and complete information – some examples

The Bank of English (part of the Collins Word Web) is a collection of modern English language held on computer for analysis of words, meanings, grammar and usage. This huge collection is composed of many different types of writing and speech. It contains up-to-date English language from thousands of different sources: the written texts come from newspapers, magazines, fiction and non-fiction books, brochures, reports, and websites and the spoken material comes from television and radio broadcasts, meetings, interviews, discussions, and conversations. The Bank of English provides evidence about the English which people read, write, speak and hear every day of their lives. The corpus contains 524 million words and it continues to grow with the constant addition of new material. Research by Collins Cobuild over the last twenty years has shown that very large samples of text are necessary for proper linguistic study. In this research, we have used the free Concordance Demonstration and Collocation Demonstration, available in http://www.collins.co.uk/corpus/CorpusSearch.aspx. Here we present an example of the concordance of the term *accommodation* in the Bank of English, to illustrate the analysis we performed:

If you prefer, we can arrange hotel **accomodation** for you. The options

car hire from £ 200 a day. [p] **Accomodation**: Trailfinders offers

day. [p] Accomodation: Trailfinders offers **accomodation** in Paihia with

including travel, insurance, meals and **accomodation** is expected to be

Oxford or Cotswolds. ALA. Box 27209. [p] **Accomodation** offered. 40p per

essential. Phone David, 071-836 2314. [p] **Accomodation** offered 40p per

Sponsor effectively offers, and guarantees **accomodation**, board and all

memberss to provide their stock with such **accomodation**. Richard Jewson

comes here, " he continued `he should find **accomodation** a little easier."

which had been propped-up against an **accomodation** block housing

of the government say that good, cheap **accomodation** is being cut back,

Passengers are being offered alternative **accomodation** on fleets of buses

is otherwise going to shell out for secure **accomodation** on tour - would

good seal on the door between our sleeping **accomodation** and the toilet [p]
favourite sporting event, including travel, **accomodation** and entrance to The YMCA is already the largest
provider of **accommodation** for the young.

Approximately 7,027 sq ft of office **accommodation** is provided. The

the hotel. If you are interested in hotel **accommodation**, please consult

The Student Houses provide `self-catered" **accommodation**, with the Halls

two or three modules. [p] [p] PHOTOS [h] **Accommodation** [/h] [h] Special

and a tour of Cape **ACCOMMODATION** AND MEALS: we stay in comfortable hotels;

Hotels throughout though more modest **accommodation** in the Syrian interior.

beach hotel; half board **accommodation**. 21 First class sleeper train to

20 Arrive London **Accommodation** & Meals: Bed/breakfast at hotels in

include: return flight and Security Taxes, **accommodation** on a room and

July 1993. You may reduce the cost of your **accommodation** and car hire by

influences the success of a holiday. The **accommodation** is equally

adults and covers **accommodation** and breakfast for one person in a double

or elsewhere in Scotland, and can arrange **accommodation** for your free of farms comfortable, well-equipped
accommodation sleep up to 7 people share

Nearest tube: Green ParkHOTEL **ACCOMMODATION**: [/h] [p] The conference fee

Brochure giving you up to six nights free **accommodation**. We hope you will

at the coast. Free **accommodation**. American student welcome to apply. Box

main saloon, set to port. The well-finished **accommodation** is a strikin

category, but came second on `Sleeping **accommodation**' The ACM trailed

inclined to take indifferent **accommodation**, bad weather, poor snow quality,

or Manchester and two-weeks **accommodation** at the Southern Palms Hotel

operate a policy of only one offer of **accommodation**. Other people may be

and Hungary. Travel is by luxury coach, and **accommodation** is in guesthouses

and other seafarers, and offers fine **accommodation**. [p] And it's right

all this began. All passenger seating **accommodation** on overnight services

minimal duties exacted, and **accommodation** and other services for foreign

Depending upon your preferences, your **accommodation** will vary from luxury walk-in safari

Sample concordance of the word *accommodation*

Observing the concordances, it’s possible to identify some expressions composed with adjectives and the word *accommodation*, such as ‘self-catering accommodation’, ‘sleeping accommodation’, ‘two-week accommodation’, ‘seating accommodation’, and ‘hotel accommodation’. Some adjectives are often used to qualify it, such as: ‘well-equipped’, ‘fine’, ‘indifferent’, ‘cheap”, ‘secure’, ‘alternative’. It’s possible to see that some verbs are used with this word, such as ‘arrange’, ‘offer’, ‘find’ and ‘provide’. The expression ‘accommodation and meals’ was considered frequent too. Since our objective is to build a material for pedagogical purposes, in order to fulfill the needs of a student of Tourism Graduation Courses to master the basic vocabulary of tourism, this kind of information is very useful for our ontology.

Here we present another example of how the Bank of English was used in the research. We present the concordance of the term *hotel* in the Bank of English, in which we can see its collocate, corpus frequency, joint frequency and significance in the corpora. This kind of search helped us check if the related frequent words were present in our ontology. In this case, we can see that words such as ‘rooms’, ‘restaurant’, ‘accommodation’, ‘guest’, ‘stay’, ‘breakfast’, ‘nights’ and ‘luggage’ must be represented in the ontology for having some kind of relation to hotel. All the semantic relations were clearly described on the ontology (meronym, holonym, hypernym, hyponym).

Collocate	**Corpus Freq**	**Joint Freq**	**Significance**
the	2313407	3960	25.765107
at	226027	1022	24.821367
in	765730	1604	20.722066
a	973489	1679	16.958702
room	13713	269	15.556004
hotel	5845	192	13.429979
star	6198	179	12.910775
london	22502	184	11.887697

park	6603	126	10.630314
rooms	2578	110	10.239605
[p]	753638	1080	9.680752
[h]	106714	255	9.213140
where	42470	158	9.154221
[/h]	106294	248	8.924720
restaurant	2295	77	8.510572
grand	3169	78	8.469029
accommodation	1740	70	8.156362
beach	2364	66	7.829876
luxury	1214	61	7.653118
night	20466	94	7.561429
guests	1827	57	7.305203
stay	6749	65	7.216017
breakfast	2050	55	7.136762
house	22752	90	7.062403
bar	3715	55	6.909804
nights	1874	50	6.803154
hilton	743	45	6.596236
luggage	448	44	6.564974
outside	8826	59	6.519566
pound	58153	134	6.497389
savoy	249	42	6.441900
lane	2260	44	6.288826
lobby	615	40	6.226255
de	7035	52	6.224883
transported	232	39	6.207443
staying	1407	41	6.180991
royal	6909	51	6.163423
walk	5330	48	6.150493
village	4173	45	6.079345
back	47563	112	6.039704
stayed	1943	40	6.013989

comfortable	2504	41	6.007800
suite	639	37	5.976566
club	10579	55	5.974168
country	19542	69	5.928384
airport	2135	39	5.899396
luxurious	396	35	5.848413
independent	4060	42	5.847436
st	6844	47	5.846466
manager	6305	46	5.842569
palace	2513	37	5.665122
york	8918	48	5.626961
resort	1291	34	5.607132
chain	1564	34	5.559803
near	8259	46	5.551325
le	1851	34	5.510046
name	28566	77	5.484059
spa	571	31	5.464091
bed	5398	40	5.461747
forte	726	31	5.435949
sheraton	85	29	5.369209
board	6060	40	5.355934
grosvenor	177	29	5.351938
run	11538	49	5.333736
situated	479	29	5.295247
hyde	222	28	5.249091
from	183465	272	5.246890
la	2959	33	5.223849
ritz	180	27	5.161134
into	65158	123	5.151357
dinner	3398	33	5.146595
four	24648	67	5.141276
map	1839	30	5.137810
b	9314	42	5.027884

located	950	27	5.011331
pool	2194	29	4.973305
golf	3001	30	4.923344
offers	3794	31	4.878909
class	8070	38	4.841008
meals	1508	26	4.800051
catering	539	24	4.787757
[ZZ1]	25694	64	4.753219
reservations	710	24	4.752471
booked	813	24	4.731216
brighton	746	23	4.638583
lodge	751	23	4.637529
island	3616	28	4.600689
fine	6726	33	4.560946
victoria	1139	23	4.555743
inn	661	22	4.547953
w1	687	22	4.542349
restaurants	1251	23	4.532135
staff	7773	34	4.483354
dorchester	113	20	4.446593
hall	5588	30	4.445873
five	22552	56	4.436807
seasons	707	21	4.426613
conference	5334	29	4.383862
3	20185	52	4.381416
group	16895	47	4.364386

Table 1: Collocates and frequencies for the word *hotel*

Results

After consulting all of the selected sources, the ontology was built with the main concepts for the Tourism ontology, which are: TOURIST, MOTIVATION, TOURISM BUSINESS, DESTINATION, TRAVEL, TRANSPORTATION, ACCOMODATION, ATTRACTIONS and ACTIVITIES. For each concept all the terms in English and Portuguese were provided, consisting in a bilingual vocabulary. The semantic anchorage of the English and Portuguese vocabularies to the basic concept ontology built for the tourism domain resulted in a total of 267 terms.

Acknowledgments

The author is thankful to FUNDUNESP for the financial support.

References

CAMBRIDGE Word Routes – Inglês-Português**. 1999. São Paulo: Cambridge/Martins Fontes.**

Dubicka I. ; O'keeffe M. 2003. *English for International Tourism*. Longman / Pearson Education Limited.

LONGMAN Essential Activator. 1997. Essex: Longman Group Limited.

LONGMAN Language Activator. 2005. (2a. ed.) Essex: Longman Group Limited.

*McArthur, T**. 1981.*** **Longman Lexicon of Contemporary English.** ***Essex: Longman Group Limited.***

O'HARA, F. 2002. *Be my guest – English for the Hotel Industry*. Cambridge : CUP.

Römer, U. 2002. Comparing real and ideal language learner input: the use of an EFL textbook corpus in corpus linguistics and language teaching. *Proceedings of the Fifth Teaching and Language Corpora Conference.* Bertinoro.

Stott, T; Buckingham, A. *At your service: English for the Travel and Tourist Industry*. 4th Edition. Oxford: Oxford University Press, 2000.

Stott, T.; Holt, R. 1995. *First Class – English for Tourism*. Oxford: Oxford University Press.

Tosqui-Lucks, P. 2007. *Construção e ancoragem do vocabulário básico bilíngüe do Turismo com fins pedagógicos*. Doctoral Dissertartion, 248p, UNESP - Araraquara, São Paulo State, Brazil.

Wood, N. 2003. *Tourism and Catering Workshop*. Oxford: OUP.

SOFTWARE DEMOS

CORPUSLAB: BRIDGING THE GAP BETWEEN CORPORA AND LANGUAGE LEARNING

Michael Barlow[347]

Although there is a general interest in the use of corpora as a source of evidence of language usage in different genres and as a source of language learning materials, there are a number of impediments to the full exploitation of corpora due to the difficulty in obtaining suitable tools and resources appropriate for any particular teaching situation. This presentation briefly assesses the current situation and describes the development of a website for teachers that provides a variety of tools and resources.

The author will give an overview of a website that has been developed to promote the use of corpus-informed language teaching materials. At the present time, teachers tend to develop their own wordlists and collocation lists for use in specialised courses and one aim of the site is to provide a collaborative environment where teachers can upload/download corpus resources such as wordlists for a variety of genres: Business English, English for Tourism, etc. (Versions of the site in languages other than English are being developed.). In addition the site contains some tools such as a simple concordancer and text analysis utility that aid teachers or material developers in producing or evaluating materials with respect to the use of appropriate words and collocations. These resources feed into the third, main component, which is an exercise-generating utility allowing teachers/authors to produce a variety of corpus-informed exercises (matching, reordering, categorising, etc.) for Business English, Medical English, and so on.

The presentation will consist of a demonstration of the use of the different features of the website by students, teachers, and schools and an assessment of the current site design.

Keywords: e-learning, exercise-authoring, text analysis, concordancer, wordlists

[347] Michael Barlow is Associate Professor in the Department of Applied Language Studies and Linguistics at the University of Auckland in New Zealand where he teaches courses on Corpus Linguistics and CALL. He is the author of the text analysis programs: MonoConc, ParaConc and Collocate and is the co-author of two corpus-based ESL texts: Phrasal Verbs and Business Phrasal Verbs.

THINGS A PARALLEL CORPUS CAN DO

Ana Frankenberg-Garcia[348]
Pedro Sousa[349]
Rosário Silva[350]
Susana Inácio[351]

COMPARA, a bi-directional parallel corpus of English and Portuguese literary texts, was first made available to the public through its free, online search interface at the end of 2000. At the time, the corpus contained 65 thousand words and an embryonic search interface. After eight years of improvements, the corpus now contains 3 million words and an enhanced, user-friendly interface, designed not just for corpus and computational linguists, but also for language learners, language teachers, university lecturers, students and translators with little or no prior experience of using corpora. In addition to this, many new features have been added to the corpus, including grammar annotation for both Portuguese and English, semantic annotation for colour, and automatically-generated frequency lists for a number of specific, user-oriented types of distributions.

After a short introduction to the corpus, this software demonstration aims to present COMPARA's new web interface and show what some of the novel and less commonly used functionalities of the corpus are, how they work and how they can be used in language, literature and translation teaching and research. As many of the features we will present are unique, they should be of interest not just to new and regular users of COMPARA, but also to anyone interested in designing and using parallel corpora in general.

We begin with a very brief navigation of COMPARA's website[352]. Among other things, people wishing to embark on a similar, parallel-corpus adventure will be able to read about details on how we built the COMPARA; people who are new to corpora will be able to find out how they can teach themselves to use COMPARA; and people wishing to use the corpus in teaching and research will find useful background information about the corpus in addition to references to other studies that make use of it.

We will then use COMPARA's free, online search interface to demonstrate how to use the corpus, starting with parallel concordances for the classroom and then focussing on new and not so well-known facilities that can be particularly useful in a broader pedagogical framework.

Keywords: parallel corpora, language teaching, teaching translation, teaching literature, COMPARA

Acknowledgement

COMPARA is part of the Linguateca project, which is jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC. We also wish to thank Diana Santos for her pioneering work in the development of COMPARA.

[348] Ana Frankenberg-Garcia is senior researcher at the FCCN branch of Linguateca, where she is joint project leader of COMPARA. She is also Auxiliary Professor at the Instituto Superior de Línguas e Administração, in Lisbon, where she currently teaches ESP and Corpora for Applied Translation.
[349] Pedro Sousa graduated in Computer Science and Computer Engineering at ISEL, Lisbon. He is a full-time research assistant at the FCCN branch of Linguateca. He is responsible maintaning COMPARA and jointly responsible for developing the DISPARA interface.
[350] Rosário Silva is a translation graduate from ISLA, Lisbon. She works part-time for the FCCN branch of Linguateca, where she is responsible for digital text processing for COMPARA and revising the output of CLAWS. She is also a professional translator.
[351] Susana Inácio graduated in Modern Languages and Literatures at the University of Lisbon. She is a full-time research assistant at the FCCN branch of Linguateca. She is responsible for revising the output of the PALAVRAS tagger for Portuguese in COMPARA and jointly responsible for the development of the corpus.
[352] http://www.linguateca.pt/COMPARA/

THE NEW MICASE ONLINE INTERFACE AND ITS POTENTIAL FOR EAP TEACHING

Ute Römer[353]
Stefanie Wulff[354]

The *Michigan Corpus of Academic Spoken English* (MICASE) is a collection of more than 1.7 million words of transcribed speech (almost 200 hours of recordings) from the microcosm of the University of Michigan in Ann Arbor, MI, USA. MICASE contains data from a wide range of situations (152 different speech events altogether) and a variety of locations across the university. MICASE speech events include lectures, classroom discussions, lab sections, seminars, advising sessions, and dissertation defenses.

The entire corpus is made freely available through a custom-designed online search and browse interface (http://quod.lib.umich.edu/m/micase/) which has recently been updated. In this presentation, we provide a hands-on introduction to the new interface with a particular focus on the needs of language instructors and their students. We demonstrate how MICASE online can be profitably used to enrich the learning environment, from the initial stage of language data retrieval to the final product in the form of innovative teaching materials.

We start out with a basic introduction to the MICASE browse and search options. In browse mode, users can retrieve (and then read or download) transcript files that match certain pre-selected criteria (e.g. retrieve all Humanities & Arts seminars or all Biological Sciences events in which non-native speakers are involved). The search function allows users to search for words and phrases in the whole corpus or in a selected sub-set of files that match certain criteria, and to create concordances that include references to files, full utterances, and speakers. Moreover, the interface provides some descriptive statistics for each search term (e.g. number of hits by gender and academic division) and allows users to download query results in XML or tab-delimited format.

In a second step we illustrate how the interface can be used as a quick and easy tool to create EAP teaching materials. For one, we show how MICASE online output can serve as the basis of gap-fill exercises, e.g. on phrasal verbs or to distinguish between near-synonyms. Beyond that, we present an example of teaching criticism strategies using MICASE online, which accomodates the central role that discourse and pragmatic functions play in EAP teaching.

Keywords: MICASE, using online corpora, academic speech, EAP teaching, materials creation

[353] Ute Römer is currently Director of the Applied Corpus Linguistics Unit at the University of Michigan English Language Institute where she manages the MICASE (Michigan Corpus of Academic Spoken English) and MICUSP (Michigan Corpus of Upper-level Student Papers) projects. Ute’s primary research interests and areas in which she has published include corpus linguistics, phraseology, and the application of corpora in language learning and teaching. Her current research focus is on the creation of evaluative meaning in academic writing and on how corpus tools and methods can be used to identify meaningful units in specialized discourses.

[354] Stefanie Wulff is currently a postdoctoral research fellow at the English Language Institute at the University of Michigan. A quantitative corpus linguist at heart, Stefanie's primary research focus is on linguistic variation, be it lexico-syntactic, idiomatic, or dialectal variation; her publications include studies on the semantic similarity of serial verb constructions, the ordering of prenominal adjectives in English, the variety-specific usage of the so-called into-causative, and a book-long treatment of idiomaticity. Another major strand of her research is devoted to a constructionist account of foreign language learning and proficiency development.

THE SLOW PATH TOWARDS FAST LEARNING: SPEECH CORPUS INTEGRATION INTO THE TRAINING OF CHINESE INSTRUCTORS THROUGH SPEECHINDEXER

Jozsef Szakos[364]
Ulrike Glavitsch[365]

We have been trying to integrate corpus methods into Chinese language instruction over the past ten years, following the presentation at TaLC98 in Oxford. While this originally meant text-based instruction where concordancing was the main point, recently we are have been shifting our emphasis to incorporate authentic speech materials of Chinese into intermediate and advanced instruction and testing.

Our approach complements the great number of Chinese instructional materials and increases the efficiency of teaching through covering areas missing in the available teaching aids and through the return to living audio materials. The up-to-date recordings are broken down into usable units and aligned with character text, pinyin transliteration, translation, using the SpeechIndexer software. SpeechIndexer has a powerful semi-automatic indexing utility, making the alignment easier through pre-segmenting and correlating intonational units and suggesting an easier way to acquire complete phrasal units.

In our presentation, we intend to report (1) on the latest developments in the SpeechIndexer and SpeechFinder software, of which earlier versions have been introduced at other conferences, and (2) on the latest instructional DVD-s we have prepared in our courses together with our MA students of Chinese as a Foreign Language and which are already used in teaching. Topics include recent social events and changes, cultural aspects of life in China, regional specialities of China and scientific discoveries.

We are training instructors of Chinese to foreigners and in this way we are enhancing their future teaching methods, completing the change from the traditional to a corpus-based one.

Keywords: SpeechIndexer, Chinese learner corpora, speech/text alignment in teaching materials, speech-based testing and evaluation

[364] Jozsef Szakos is an Associate Professor in the Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University. He is the programme leader for MA in Chinese as a Foreign Language, the first one of this kind in Hong Kong. Previously he had been in Taiwan, director of an Indigenous Languages Department in Hualian, doing corpus based research on Formosan Austronesian languages. He had also taught at Providence University (Graduate School of Western Languages, English department) in Taichung, and other Taiwanese universities, mainly languages and linguistics. He founded the Chinese Language Education Center (CLEC) at Providence University, and advised dozens of corpus-based MA theses in Linguistics. He earned his Dr Phil in Bonn, Germany (1994), majoring in General Linguistics, minors Sinology and Comparative Religion and also has a Licenciate in Theology from Budapest (1982).

[365] Ulrike Glavitsch received her diploma in Computer Science from ETH Zurich in 1988 and her master’s degree from Stanford University, USA, in 1990. She was a software developer and group leader at Schmid Telecom AG, Switzerland, for several years. In 2000, she joined the speech processing Group at ETH and worked in the area of automatic speech recognition. Since 2005, she is a member of the native systems group at the Institute of Computer Systems at ETH Zurich. Her main research interest is the development of operating systems for reliability-critical applications. Ulrike Glavitsch designed and implemented large parts of the SpeechIndexer software.

WORKSHOPS

USING XAIRA TO EXPLORE YOUR XML CORPUS

Guy Aston[366]
Lou Burnard[367]

This workshop will introduce participants to the latest version of the XAIRA system developed originally for use with the British National Corpus, but now enhanced as a general purpose and open source cross-platform software architecture.

Participants will learn how this software can take advantage of all the XML markup in the new XML edition of the British National Corpus. They will also learn how to use Xaira with their own corpora. Xaira can operate on a simple collection of plain text files, with no markup at all. It can also operate on a collection of texts with very sophisticated embedded linguistic markup, provided this is expressed in some dialect of XML. Participants will learn how to customize the program for either kind of material.

We will provide a series of exploratory exercises, designed to show off the searching capabilities of the system when used with the BNC. These will include the production of concordances, word lists, collocation lists etc. in the usual way, but with an emphasis on the kind of application for such capabilities likely to be of most use in a language teaching environment. Particular attention will be paid to issues of integration and portability, in order to show how results obtained with Xaira can be integrated into other teaching material. We will also present and discuss strategies for encouraging students' own exploration of corpus resources using the program.

All material used at the workshop will be made available online, and we will therefore need network access. Participants will also be encouraged to bring their own materials for experimentation.

http://www.natcorp.ox.ac.uk/workshop/

http://www.xaira.org

Keywords: XAIRA, XML, retrieval software, indexing, computer-aided learning

[366] Guy Aston, MA (Oxon), MSc (Edinburgh), PhD (London), is Professor of English Linguistics at the Advanced School of Interpreters and Translators of the University of Bologna, where he teaches English for interpreters, Computer-aided translation and English-Italian liaison interpreting. His research interests include corpus linguistics, contrastive pragmatics, conversational analysis and autonomous language learning.

[367] Lou Burnard is Assistant Director of Oxford University Computing Services, and group manager for the Information and Support Group, one of its four major divisions. He set up the Oxford Text Archive in 1976, has been European editor of the Text Encoding Initiative since 1989, and is responsible for Oxford University's participation in the British National Corpus Project.

EXPLORING AND TEACHING THE PHRASEOLOGY OF ACADEMIC DISCOURSE

***Michael Barlow*[368]**
***Ute Römer*[369]**

Phraseology has had a rather marginal status in linguistic analysis and description in those linguistic theories that treat lexis and grammar separately and deal with words independent of their preferred grammatical structures and with grammatical structures independent of the words that typically occur in them (cf. Sinclair 2005; see also Ellis 2008). Consequently, the phrase has not always received the attention it deserves, given its central status as meaning-carrying unit (cf. Römer 2008; Sinclair 1996). Although corpus linguists have worked on different approaches to phraseology (see e.g. Biber, Conrad and Cortes 2003; Hunston and Francis 2000; Meunier and Granger 2008; Scott and Tribble 2006), there remains a gap between a general awareness of phraseology and knowledge of practical methods of identifying and analysing lexical units. The purpose of the present workshop is to provide hands-on experience in extracting and analysing phraseological units.

Phraseological items (or collocations, multi-word units, lexical bundles – to list a few commonly used alternative notions) can be broadly defined as repeatedly occurring contiguous or non-contiguous combinations of two or (usually) more words that carry meaning. We assume that it is essential for the language learner (and teacher) to know what word combinations are most commonly used in a particular type of discourse (e.g. in academic speech) so that learning and teaching activities can centre around those combinations and on the semantic and pragmatic meanings conveyed by them. But how can pedagogically relevant phraseological items be identified in corpora?

There is now becoming available a new generation of software tools that enable users to extract from a corpus lists of candidate phraseological items for inspection. One of these new phraseological search engines is *Collocate* (Barlow 2004). *Collocate* uses frequency information and statistical analyses (t-score, log likelihood, MI) in order to retrieve lists of

(a) collocations with a specified search word and within a set span (e.g. four words),

(b) n-grams (lexical bundles) of different lengths, and

(c) collocations extracted from the corpus as a whole.

In this 3-hour workshop, we will first demonstrate some of the *Collocate* facilities, focussing on the extraction of phrases and meaningful items from corpora such as MICASE (the Michigan Corpus of Academic Spoken English) and MICUSP (the Michigan Corpus of Upper-level Student Papers). The focus will be on corpora that capture English academic discourse but the analytic steps we will demonstrate are universal and can also be applied to other types of discourse and on data from languages other than English. We will discuss the ways in which *Collocate* can be used to provide insights into the phraseological profile of academic speech and writing, and how the program can highlight which word combinations the members of a particular disciplinary discourse community tend to use in order to convey particular meanings.

In the hands-on part of the workshop, the participants will be able to work with and evaluate different functions and statistics used in *Collocate* in order to produce and interpret lists of collocations from different disciplinary subsets of MICASE and MICUSP. We will then move on from analysis to pedagogy and show how *Collocate* output can be used in the creation of innovative teaching materials. After a discussion of learner and EAP instructor needs, workshop participants will be provided with exercise templates and get the opportunity to design their own exercises for use in the EAP classroom.

Keywords: phraseology, collocations, EAP teaching, academic speech and writing, disciplinary discourse

[368] Michael Barlow is Associate Professor in the Department of Applied Language Studies and Linguistics at the University of Auckland in New Zealand where he teaches courses on Corpus Linguistics and CALL. He is the author of the text analysis programs: MonoConc, ParaConc and Collocate and is the co-author of two corpus-based ESL texts: Phrasal Verbs and Business Phrasal Verbs.

[369] Ute Römer is currently Director of the Applied Corpus Linguistics Unit at the University of Michigan English Language Institute where she manages the MICASE (Michigan Corpus of Academic Spoken English) and MICUSP (Michigan Corpus of Upper-level Student Papers) projects. Ute received her PhD in English linguistics from the University of Hanover, Germany, in 2004. Her primary research interests and areas in which she has published include corpus linguistics, phraseology, and the application of corpora in language learning and teaching. Ute's current research focus is on the creation of evaluative meaning in academic writing and on how corpus tools and methods can be used to identify meaningful units in specialized discourses. More information about Ute's research interests and a list of her publications can be found at http://www.uteroemer.com

References

Barlow, M. 2004. *Collocate 1.0: Locating collocations and terminology*. Houston, TX: Athelstan.

Biber, D., Conrad, S. and **Cortes, V.** 2003. "Lexical bundles in speech and writing: an initial taxonomy." In A. Wilson, P. Rayson and T. McEnery (eds), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Frankfurt/Main: Peter Lang. 71-92.

Ellis, N. C. 2008. "Phraseology: the periphery and the heart of language." In *Phraseology in Foreign Language Learning and Teaching*, F. Meunier and S. Granger (eds.). Amsterdam: Benjamins, 1-13.

Hunston, S. and **Francis, G** 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Meunier, F. and **Granger, S.** (eds) 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: Benjamins.

Römer, U. 2008. "Identification impossible? A corpus approach to realisations of evaluative meaning in academic writing" *Functions of Language* 15/1: 115-130.

Scott, M. and **Tribble, C.** 2006. *Textual Patterns: keyword and corpus analysis in language education*. Amsterdam: Benjamins.

Sinclair, J. McH. 1996. "The search for units of meaning" *TEXTUS IX*: 75-106.

Sinclair, J. McH. 2005. "The phrase, the whole phrase, and nothing but the phrase" Keynote given at *Phraseology 2005*, Louvain-la-Neuve, October 13-15, 2005.

TALC AT TALC: TEACHING AND LINGUATECA'S (PORTUGUESE LANGUAGE) CORPORA

***Ana Frankenberg-Garcia*[370]**
***Belinda Maia*[371]**
***Cláudia Freitas*[372]**
***Diana Santos*[373]**

The use of corpora in language learning can be a valuable tool for both teachers and learners. In the past couple of decades, a steadily growing number of corpus-based dictionaries, grammars and textbooks have become available and been hugely successful in the teaching of English as a Foreign Language. Little progress has been made, however, with regard to the development of corpus-based pedagogic materials for languages other than English.

The aim of the present workshop is to introduce people working with the Portuguese language in educational settings to a number of corpus resources and tools created and made available to the general public by Linguateca over the past decade. These include the AC/DC project (free access to large quantities of Portuguese parsed text, including CETEMPúblico, a 180 million word corpus of newspaper text from the daily Portuguese newspaper *Público* and CETENFolha from the daily Brazilian newspaper *Folha de São Paulo*), COMPARA (a 3 million word bi-directional parallel corpus of Portuguese and English), the Floresta Sintá(c)tica (an expanding 1.5 million word syntactic treebank for Portuguese) and the Corpógrafo (a web-based platform for building and managing your own corpus).

Although most of the above were originally conceived for research in natural language processing, their usefulness in education cannot be overlooked. The scant availability of ready-made, corpus-based learner dictionaries, grammars and textbooks for the teaching of Portuguese makes it all the more important to disseminate resources such as these. Language teachers who learn how to use them can create their own corpus-based pedagogic materials to supplement areas that are particularly lacking, such as information on collocations.

This 6-hour, full-day workshop, which will be conducted in Portuguese, will begin with a brief introduction to Linguateca's corpus resources and practical demonstrations of their applications in language teaching. Later in the day, participants will have the opportunity to try out these resources hands-on and create a few of their own materials for teaching Portuguese. Participants who wish so will be able to share their corpus-based Portuguese language pedagogic materials on Linguateca's website.

Keywords: Portuguese corpora, Portuguese language teaching, Portuguese language resources, Linguateca

Acknowledgement

Linguateca is jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC.

[370] Ana Frankenberg-Garcia is Auxiliary Professor at ISLA, Lisbon, where she teaches English language and translation, and a senior researcher at Linguateca, FCCN, where she is joint leader of the COMPARA project. She holds a PhD in Applied Linguistics from Edinburgh University and her research interests include the use of corpora for language learning and translation studies, parallel corpora, corpus usability and user behaviour, learner autonomy, crosslinguistic influence and second language writing. See also http://adamastor.linguateca.pt/COMPARA/equipa/Ana/AnaHome.html

[371] Belinda Maia is an Associate Professor at the Faculdade de Letras da Universidade do Porto where she is responsible for teaching and research in the areas of contrastive linguistics, translation, information technology applied to translation, and terminology. She is the supervisor of the Polo CLUP/FLUP of the Linguateca project. See also http://web.letras.up.pt/bhsmaia/belinda/index.htm

[372] http://web.letras.up.pt/bhsmaia/belinda/index.htm

Cláudia Freitas obtained her PhD in Linguistics in 2007, with a thesis about the extraction of semantic relations from corpora. From 2002 to 2007, she taught grammar and writing skills to Brazilian students at Pontifícia Universidade Católica, Rio de Janeiro. In 2007, she joined Linguateca, working primarily at the Floresta Sintáctica project. See also http://eden.dei.uc.pt/~freitas/

[373] Diana Santos has worked with Portuguese corpora for the last 18 years. She led the development of the first corpus browser for Portuguese in 1992, did her PhD in corpus-based contrastive studies in 1996, and, under the scope of Linguateca, was involved in the creation of AC/DC, COMPARA, CETEMPúblico, CETENFolha and many other corpora for the processing of Portuguese. See also http:/www.linguateca.pt/Diana/diana.html

ANNOTATING PEDAGOGY: IMPLEMENTING LANGUAGE TEACHING AND LEARNING-ORIENTED ANNOTATION ON CORPORA

Pascual Pérez-Paredes[192]

José M. Alcaraz[193]

The main aim of this hands-on workshop is to show participants the practical ways in which corpora can be annotated from a pedagogical perspective and create a corpus that is pedagogically rich and usable in the language classroom. This workshop is an excellent opportunity to motivate language teachers and researchers to think pedagogy as an annotation target, along with the better-established morphology or syntax.

The TALC community has been debating for over a decade now on the pedagogic uses of corpora. In this debate, different contributions have addressed the exploitation of language corpora in the classroom, both as direct and indirect sources of information and activities. For the most part, these proposals have made an extensive use of L1 principled corpora. The BNC is a case in point. Proposals to integrate the BNC in the language curriculum abound and address a wide range of elements of the curriculum.

However, pedagogical annotation has not been in the agenda of language educators. The reasons for this are not easy to enumerate here. Suffice it to say that the fascination of the CL community for the usefulness of principled corpora such as the BNC have delayed the implementation of other initiatives that may meet the needs of non-linguists, or non-tertiary students of foreign languages.

The workshop is structured in five different stages which combine the lecture input with a more predominant hands-on approach. First, participants will be introduced into the notion of pedagogical annotation. Here, we will present a theoretical framework that will serve as the basis for the understanding of the following steps. Second, we will offer an overview of the annotation tool that will be used in this workshop: SACODEYL Annotator. This freeware and open source application has been developed within the frame of SACODEYL, an international EU-funded Minerva initiative that implements DDL online language learning opportunities for young people. Although participants will be given precise instructions on how to use the main functions of the tool, the emphasis of the workshop is not this particular software but the technology behind, that is, XML and, in particular, the extensibility quality of this markup language.

After this, participants in the workshop will be given the chance to annotate themselves part of a corpus so as to test the theoretical approach and the know-how discussed above. This is the most important activity of the workshop. The preparatory work for the annotation stage includes the division of a text/interview into learning units and the structuring of an annotation taxonomy tree. The level of granularity here is for the user or annotator to decide. The application of this user-driven annotation may include topics, grammar, lexis, target exploitation level and similar pedagogical units that may play a role in the teaching/ research context of those taking part in the workshop.

Subsequent to this hands-on practice, a debate will explore the different views on the text(s) annotated, laying emphasis on the flexibility of the tool and the structuring possibilities of XML annotation. A fifth stage of the workshop will offer more technical information on the possibilities for the exploitation and uses of annotated XML corpora.

[192] Pascual Pérez-Paredes started his collaboration with the English Department in the University of Murcia, Spain in 1996. After a research stay in the University of Texas at Austin, he completed his PhD in Applied Linguistics in 1999. He currently teaches CALL and Applied Linguistics. He is also a Sworn/Official Translator. His main interests are quantitative research of register variation, the compilation and use of language corpora and the implementation of Information and Communication Technologies in Foreign Language Teaching/Learning. He is a member of the Research Group Lingüística Aplicada Computacional, Enseñanza de Lenguas y Lexicografía (LACELL). Pascual Pérez-Paredes directs a research project funded by the SENECA agency on orality. He is also the coordinator of a MINERVA project funded by the European Commission: SACODEYL (225836-CP-1-2005-1-ES-MINERVA-M SACODEYL). At the moment, he is involved in corpus-based international projects such as LINDSEI (UCL) [http://cecl.fltr.ucl.ac.be/] and ICCI (TUFS) [http://www.tufs.ac.jp/insidetufs/doc/08012802.pdf].

[193] José M. Alcaraz-Calero is a Computer Science Engineer at the Computer Science School, University of Murcia. He has widely contributed to Corpus-based language research and thus has implemented solutions for different projects at UMU, including automatic semantic tagging of Spanish based of word sense disambiguation (WSD) and intelligent tagging algorithms for the CUMBRE corpus of Spanish. At the moment he is completing his PhD in autonomic computing, networks and ontologies.

We expect that by the end of the workshop participants will have not only mastered the basics of XML-driven pedagogical annotation, but also will have shared views and developed their own appreciation of how pedagogy can be actually annotated and further exploited in the classroom.

Participants will be provided with details on how to access and further use SACODEYL tools.

Keywords: Corpus annotation, pedagogical corpora, user-centered corpus exploitation, XML resources.