# Training translators to use corpora hands-on: challenges and reactions by a group of thirteen students at a UK university

Ana Frankenberg-Garcia[1]

**Abstract**

The use of corpora is no longer restricted to a small community of researchers working on language description and natural language processing. Anyone with an Internet connection is now able to access corpora to help them with everyday questions about language, including questions for which dictionaries, grammars and other language resources do not always have clear answers. Translators are among those who have much to gain from using corpora, and this is widely acknowledged in the literature. Yet much of the research at the crossroads of translation and corpora seems to focus on the use of corpora in Translation Studies, and there does not seem to be enough information on the use of corpora in actual translation training and practice. In this paper, I discuss some of the challenges of training translators to use corpora and then describe how a group of thirteen students studying for an MA in Translation at the University of Surrey reacted to a hands-on module on learning to use corpora in everyday translation. The latter is based on the students' responses to a questionnaire and on a corpus of self-reports containing authentic examples of students using corpora in translation practice.

**Keywords**: Corpora, translation, translator education.

## 1. Introduction

Nowadays there are more and more ready-made corpora that are easily accessible to the public in general than there were a few years ago. It is no longer necessary to buy or apply for special licences or install any corpus

[1] Centre for Translation Studies, School of English and Languages, University of Surrey, Guildford, GU2 7HX, United Kingdom.
*Correspondence to*: Ana Frankenberg-Garcia, *e-mail*: a.frankenberg-garcia@surrey.ac.uk

*Corpora* 2015 Vol. 10 (3): 351–380
DOI: 10.3366/cor.2015.0081

www.euppublishing.com/journal/cor

software on one's computer in order to start using corpora. The BYU corpus website,[2] for example, provides free online access to a wide range of general English corpora, including the Corpus of Contemporary American English (Davies, 2008), the BYU-BNC interface to the British National Corpus (Davies, 2004) and the Corpus of Global Web-Based English (Davies, 2013). Likewise, the OPUS corpus collection[3] (Tiedemann, 2012), offers free online access to a wide range of diverse multi-lingual parallel corpora, including Europarl, with source texts and translations from the European Parliament Proceedings; EMEA, made with parallel documents from the European Medicines Agency; and OpenSubtitles2013, a multi-lingual collection of crowd-sourced movie subtitles. CorpusEye[4] (Bick, 2005) provides free online access to a number of general corpora in Danish, English, French, German, Norwegian, Portuguese, Spanish, Romanian, Swedish and even Esperanto. These are just a few examples, for there are many more free online corpora available.

The fact that access to corpora is no longer confined to a restricted community of researchers is not in itself sufficient, however. People will only begin to use a new tool or resource if they perceive it to be useful to them. For translators – and, indeed, any language service provider – to start using corpora, it is important that they realise that corpora can have the potential to help them find answers to questions for which there are often no clear answers in dictionaries, glossaries, Google searches and other tools and resources that they are accustomed to using. Translators in particular are constantly having to choose between different ways of presenting information in the target language; and, if used well, corpora can help translators with many of the decisions they are forced to make in the process, improving the overall quality of the translation product. This is true not only for inexperienced translators who are less accustomed to the terminology and phraseology required in a particular translation and less confident about distancing themselves from more literal translation strategies. It is also true for professional translators when they are working with subject domains less familiar to them, or when they need to accommodate a style of writing they are not accustomed to. Although expert translators might have fewer reference needs overall, corpora can eventually help both translation trainees and professionals cope better with unfamiliar terminology and phraseology, and with the styles and idiolects they may need to reproduce in a translation. Similarly, while translators working into their native language are likely to have fewer reference needs in terms of language production than translators working out of their native language, corpora can still be useful to both.

There are many examples in the literature of how translators can use corpora. Tognini-Bonelli and Manca (2004), for example, show how the

[2] Accessed 16 November 2014 at: http://corpus.byu.edu/
[3] Accessed 16 November 2014 at: http://opus.lingfil.uu.se/
[4] Accessed 16 November 2014 at: http://corp.hum.sdu.dk/

English word *welcome*, present 324 times in a corpus of English Farmhouse Holidays texts, is a lot more frequent than its literal translation *benvenutto*, which occurs only four times in a comparable corpus of Agriturismo texts in Italian. They demonstrate that, in this type of genre, *benvenutto* is probably not an appropriate equivalent for *welcome*, and they proceed to investigate Italian functional equivalents for *welcome*, obtaining some remarkable results which can be very enlightening to any translator working within this genre. Similarly, Philip (2009), uses comparable general reference corpora of Italian and English to show how even colours may not have one-to-one, literal translations. For example, she observes that red is the most common colour associated with rage and anger in English, whereas in this same context *nero* ('black') could in certain contexts be a suitable equivalent in Italian. Using a similar methodology for exploring comparable corpora, Kubler (2011) reports on how a translation student at a French university found the French adverb *agressivement* was not a good translation for its English cognate *aggressively* in the text he was working on, and on how corpora helped him to arrive at *avec virulence* – a much better choice in the given context. In another study, Frankenberg-Garcia (2014) shows how parallel corpora can raise translators' awareness of how discourse might need to be changed in a translation. By examining parallel concordances of English fiction translated into Portuguese, it became clear that there is a strong tendency for professional Portuguese translators to move time adverbs from their unmarked position at the end of the clause in English to the front of the clause in Portuguese. Findings such as these can boost translators' confidence, helping them to adopt bolder, less literal translation strategies. Bowker and Pearson (2002) and Varantola (2003) in turn demonstrate how corpora made of specialised texts from a specific subject domain can help translators come to grips with the specific terminology and phraseology that is often needed for specialised translation.

Yet despite many other examples reported in the literature of how corpora can help inexperienced and professional translators with equivalence and with specialised terminology and phraseology, and of how corpora should be part of translator education (for example, Aston, 1999; Rodríguez-Inés and Hurtado Albir, 2012; Varantola, 2003; and Zanettin, 2012), including four thematic CULT (Corpus Use and Learning to Translate) conferences,[5] translators do not seem to be using corpora much or at all. Bernardini (2006) reports on a survey of professional translators' use of corpora carried out in 2005, where 60.2 percent of the 623 respondents (mostly from the UK, but also from other European countries) replied that they did not use corpora in their translation practice, and 41.9 percent

[5] CULT 1997 (Bertinoro, Italy), CULT 2000 (Bertinoro, Italy), CULT 2004 (Barcelona, Spain) and CULT 2015 (Alicante, Spain). See also Zanettin *et al*. (2003) and Beeby *et al*. (2009) for published proceedings.

had never even heard of corpora. More recently, Gallego-Hernández (2015) analysed how 526 professional translators based in Spain deal with various translation resources. He found that nearly 50 percent never or almost never used corpora, 30 percent used corpora sometimes, and only 18 percent used corpora often or very often. These results are remarkably similar to those reported by Gough (2013) among a population consisting of 540 respondents of mainly EU-based professional translators: when researching terminology, 'Corpora are the least used resource, with over 50 percent of the respondents using them rarely or never' (Gough, 2013: 5). In contrast, the respondents in Gough's study preferred using translation–memory systems, terminology databases, glossaries and web searches. A quick examination of the international translator forums at Proz.com and TranslatorsCafe.com carried out in May 2014 pointed in the same direction: the forums did not contain any threads about corpora, compared to several daily queries about translation-memory systems and CAT tools in general.

One reason why translators are not using corpora could be that they simply do not know how to use them well enough to understand their potential benefits as a supplement to other resources and references. Indeed, corpora are not as intuitive as dictionaries, search engines and other resources that are more familiar to the public in general. Frankenberg-Garcia (2012: 476) refers to a number of studies that show that 'corpus skills that come as second nature to experts are not at all obvious to the untrained'. These studies suggest that for people to start using corpora, a certain amount of training is required. If translators are not trained to use corpora, they will not be able to decide for themselves whether using corpora will help them in their day-to-day practice.

However, as Kubler (2011) points out, there are not many translator training institutions teaching novices how to use corpora. A brief look at the current 2014 programme descriptions of MA in Translation programmes offered in fourteen different UK universities shows that less than a handful of these institutions offer specific modules on corpora for translation. It falls beyond the scope of this study to analyse translation degree programmes in other countries, but the situation is likely to be similar. The reason for not teaching trainee translators to use corpora cannot be because of technology constraints, for practically all MA in Translation programmes today are equipped with computer labs in order to teach students how to optimise their web searches and use translation-memory systems.[6] In fact, even though translation-memory systems can arguably be even less intuitive than corpora, there is a lot of pressure from the industry to train translators to use them. Translation-memory systems can result in a significant increase in translator productivity and there are important economic advantages to be gained here by translation agencies and clients

[6] Note that although some translation–memory programs include concordance searches, these searches are carried out within translation memories in use rather than within corpora.

that need large volumes of translations, especially if translators are required to use the memories owned by agencies or clients. It comes as no surprise that many translation jobs today *demand* the use of specific translation-memory systems. Moreover, software developers also stand to benefit from selling expensive proprietary programs to new translation graduates. In contrast, no such pressure exists with regard to the use of corpora. Corpus skills never seem to be mentioned in job advertisements on the translation market and, unlike translation memories, the use of corpora does not necessarily make translations cheaper. Furthermore, corpora are either completely free or comparatively inexpensive, so they are not aggressively marketed by translation industry stakeholders.

With little or no demand from the market requiring translators to use corpus tools in actual translation practice, corpora appear to be used more often for research purposes, and training translators to use corpora appears to be regarded as something optional or secondary, or only for those interested in doing research. Indeed, ever since Baker (1993) published her seminal paper on Corpus Linguistics and Translation Studies, there has been a steadily growing body of research in Translation Studies that is based on, or even driven by, corpora and corpus linguistics methods, as attested by the table of contents and abstracts of various translation journals around the world. There are even entire conferences devoted to using corpora in Translation Studies.[7] However, corpora are not just for those engaged in research. In addition to the previously mentioned studies on how corpora can help practising translators with equivalence and with specialised terminology and phraseology, anyone who knows how to use corpora to look up linguistic information that is not readily available elsewhere understands that practitioners have much to gain from corpora. It is not so much a question of improving productivity and making translations cheaper, as is the case with the use of translation-memory systems, but more a matter of boosting autonomy with regard to translators' decision-making processes and of improving the overall quality of the translation product.

Besides the fact that there is no particular pressure from the market to train translators to use corpora, yet another possible reason why translator training institutions have not so far paid sufficient attention to teaching learners how to use corpora in translation practice is that it is not something that is very easy to implement in translation training programmes, particularly in institutions where students are learning to translate into and out of a wide range of languages. Instructors qualified to teach how to use corpora are unlikely to have sufficient knowledge of all the language pairs their students are interested in, and instructors qualified to teach a particular language pair are often not familiar with corpora. To complicate

[7] At the time this paper was written, the fourth edition of the biennial conference on Using Corpora in Contrastive and Translation Studies took place at Lancaster University. In addition, many general translation conferences also have special corpus linguistics strands, and general corpus linguistics conferences usually have special translation strands.

things further, there is a general imbalance with regard to the availability of ready-made, off-the-shelf corpora in different languages. For example, while there are many free online general language corpora available for English, the same cannot be said for French. In addition, ready-made corpora of different languages are often integrated with different concordancers and make use of different corpus query languages, and this can be very confusing to learners. For example, the interface to the Spanish Corpus de la Real Academia Española (CREA)[8] is very different from the interface to the Deutsches Referenzkorpus (DeReKo) corpus in German.[9]

Finally, training translators to use corpora in translation practice takes time, and as Aston (2009: x) puts it, 'Not all translators, be they learners or professionals, appreciate that corpus use may have a medium- and long-term payoff which can override what they often perceive as short-term disadvantages'. So is it worth all the trouble? In this paper, I would like to give an overview of my experience of teaching a group of thirteen students studying for an MA in Translation at the University of Surrey on how to use corpora in translation practice, and then examine in detail how this group of students reacted to the training received. The latter is based on the students' responses to a questionnaire given at the end of the teaching period and on a corpus of self-reports compiled out of a graded piece of assessment on using corpora in translation practice that the students were required to submit.

## 2. Teaching to use corpora in translation practice

A group of thirteen students studying for an MA in Translation at the University of Surrey during the academic year 2013/14 took part in the present study. They were enrolled in an optional module focussing on the hands-on use of corpora for translation practice with a total of twenty-two hours of class contact time. The students in question constituted a multi-lingual group, with an interest in translating in the following language directions:

| | | |
|---|---|---|
| Spanish > English | French > English | German > English |
| Russian > English | Portuguese > English | English > Greek |
| English > Chinese | English > Italian | |

The diversity of translation language combinations within the group made the teaching particularly challenging, as it was not possible to base the content of the course on using corpora in the translation of a single, specific language pair. Moreover, different students would have to learn how to use corpora of different languages, most of which their instructor did not speak.

[8] Accessed 8 November 2014 at: http://corpus.rae.es/creanet.html
[9] Accessed 8 November 2014 at: http://www1.ids-mannheim.de/kl/projekte/korpora/

A decision was made, therefore, to teach the basics of corpora using English language corpora and parallel corpora containing English – the only language the entire group had in common. However, the students would also have to acquire enough autonomy to start using corpora in other languages if they wanted to learn how to use them in everyday translation practice. A decision was therefore made to subscribe to the Sketch Engine[10] (Kilgarriff *et al*., 2004), which, using the same search interface, provides access to corpora in several different languages, including but not limited to very large web-based corpora of all the languages relevant to the students enrolled in the class. Teaching the students how to use the Sketch Engine interface using the English corpora distributed by the Sketch Engine – particularly the British National Corpus and the enTenTen corpus (Jakubíček *et al*., 2013) – would then allow the students to explore by themselves corpora in other languages using the same interface. It was hoped this would then give them enough autonomy to be able to explore on their own non-English corpora outside the Sketch Engine, such as the previously mentioned CREA for Spanish and DeReKo for German, and so on.

All classes took place in a computer lab with individual workstations for each student. Lessons began with a general introduction, then progressed to providing the students with live demonstrations and as much hands-on practice as possible, with a focus on corpus consciousness-raising exercises such as those proposed in Frankenberg-Garcia (2012). Table 1 presents a summary of the course syllabus for 2013/14. The corpora that were used hands-on in class are listed in Appendix A.

Table 1 shows that the syllabus was not intended to be an introduction to the more theoretical aspects of corpus linguistics neither did it focus on research at the intersection of corpus linguistics and translation studies. Instead, it provided practical hands-on training, emphasising the basic knowledge and skills needed when using corpora to answer everyday questions about language, and it then introduced the students to building DIY corpora for practical translation purposes. For the latter, the students used the integrated corpus-building tools that come with Sketch Engine, which include the WebBootCat tool (Baroni *et al*., 2006), allowing the students to crawl the web in order to compile specialised language corpora. Besides the regular class contact hours, the students were encouraged to use corpora outside classes, especially during their practical translation assignments, but also to help the non-native speakers of English with their essays. From day one, the students were asked to keep a diary of the different ways in which they used corpora, which they would later need for their assignments.

As explained in the introduction, to assess the students' reactions to the training received, an anonymous questionnaire was completed by the

[10] Accessed 8 November 2014 at: https://the.sketchengine.co.uk/

| Session | Content |
|---|---|
| 1–2 | **Introduction**<br>Definition of corpora; empirical approaches to language description and evidence of how language is used by a community of users; looking up linguistic information in corpora (as opposed to dictionaries, Google and asking a native speaker); dealing with natural, unedited language in corpora (e.g., mistakes, non-standard language); differences between corpora and electronic libraries; corpus representativeness; corpus software: concordances, word lists and collocation; uses of corpora in translation practice. |
| 3 | **Different types of corpora, applications and implications**<br>Restricted access, public and DIY corpora; written, spoken and multi-media corpora; contemporary, non-contemporary and diachronic corpora; general and specialised language corpora; mono-lingual and multi-lingual corpora (parallel and comparable); full corpora and sub-corpora; lemmatised and annotated corpora. |
| 4–6 | **Single word queries**<br>Single-word queries in different corpora/corpus interfaces; case-sensitive and case-insensitive queries; using/not using diacritics; queries involving annotations; using wildcards; lemma queries; queries involving alternate forms; using POS tags. |
| | **Multiple-word queries**<br>Multiple word queries in different corpora/interfaces; conventionality, the idiom principle and realistic multiple-word queries; reformulating multiple-word queries: narrowing down and making queries more flexible. |
| | **Concordances**<br>Reading KWIC and full-sentence concordances; sorting, sampling and filtering concordances. |
| 7 | **Corpus frequencies**<br>Zipf's law; word lists; lemma lists; POS lists; *n*-grams; keywords; raw *versus* normalised frequencies; interpreting frequencies. |
| 8 | **Collocation**<br>Nodes and spans; left and right collocates; observing differences in MI scores, T scores and logDice statistics; using lemmas and POS tags in collocation queries; word sketches;[11] bilingual word sketches.[12] |
| 9–11 | **Building your own corpus**<br>Compiling corpora using pre-defined text files; compiling ad hoc specialised language corpora by crawling the Web; text alignment and compiling parallel corpora using tmx (translation memory eXchange) files. |

**Table 1**: Course syllabus for 2013/14

students at the end of the course.[13] One week later, the students were also required to hand in a graded assignment in which they had been asked to write a report on their use of corpora for translation; this was used to complement

[11] Word Sketches are automatic, corpus-based summaries of a word's grammatical and collocational behaviour (Kilgarriff *et al.*, 2004). The functionality is exclusive to the Sketch Engine corpora with POS tagging.
[12] See Kilgarriff *et al.* (2013).
[13] A similar survey, based on learning diaries, self-evaluation and student satisfaction questionnaires, has been carried out with translation students in Spain by Rodríguez-Inés and Hurtado Albir (2012).

the information yielded by the survey. The reports written by the students were used to create a small corpus which was, in turn, used to analyse the students' reactions in further detail.

## 3. End-of-semester questionnaire

The questionnaire given to the students at the end of their period of instruction was divided into four sections. The first part of the questionnaire was designed to assess the extent to which the students were familiar with corpora before starting their MA in Translation. As it was not possible to ask the students to answer this part of the questionnaire before they began their studies at Surrey, special care was taken to ensure some elicitation statements in this section were affirmative (e.g., 'I already knew what lemmatisation meant before I started this MA'), while others were purposefully formulated in the negative (e.g., 'Before my MA I didn't know what partofspeech tagging was'). This was done to prevent an acquiescence bias when requiring the students to think retrospectively about their responses. The first two questions were simply true or false questions that were designed to ascertain whether the students had heard of and/or had used corpora before their MA. As Table 2 shows, only one out of the thirteen students had heard of corpora and actually used a corpus before coming to Surrey. It was only this student (Respondent 2), therefore, that was required to answer the next six questions, which purported to capture through a five-point Likert scale the extent to which he/she was familiar with corpora before the MA. The responses by this student are summarised in Table 3.

Table 3 indicates that the only student in the group who had actually used a corpus before cannot be said to have been very familiar with corpora. He or she did not know there were different types of corpora available, did not know what a KWIC concordance was, did not understand what was meant by lemmatisation, part-of-speech (POS) tagging and normalised corpus frequencies, and did not know that it was possible to use corpora to extract information on collocation.

In the second part of the survey the students were asked to respond to a series of statements regarding how well they thought they could handle corpora after the eleven weeks of teaching. All thirteen students were required to answer this section, and their responses on a five-point Likert

| | |
|---|---|
| I had never heard of corpora before my MA. | 12 true<br>1 false |
| I had already used a corpus hands-on before I started my MA. | 1 true<br>12 false |

**Table 2**: Students' contact with corpora before they started their MA in Translation

| | |
|---|---|
| Before I began my MA at Surrey I didn't know there were different types of corpora (e.g., general mono-lingual corpora, specialised language corpora, parallel corpora, comparable corpora and so on). | agree |
| I already knew what KWIC concordances were before I came to study for this MA. | strongly disagree |
| I already knew what lemmatisation meant before I started this MA. | strongly disagree |
| Before my MA I didn't know what part-of-speech tagging was. | neither agree nor disagree |
| I didn't know collocation data could be obtained from corpora before I began my MA. | strongly agree |
| I already knew what normalised corpus frequencies were before I began my MA. | strongly disagree |

**Table 3**: Respondent 2's degree of familiarity with corpora before the MA in Translation

scale are summarised in Table 4. As shown, the central tendencies for all statements in this part of the questionnaire were quite favourable. The students generally felt that they understood the strengths and limitations of different types of corpora and agreed that they could carry out simple word queries as well as queries involving more than one word. They generally claimed to understand the difference between looking up lemmas and looking up plain words, and said they could use POS tags in their queries. In general they also felt they could use corpora to retrieve information about collocation. With regard to frequencies, they declared they were largely able to compare the frequencies of different words or combinations of words within a corpus and use normalised frequencies to compare words across different corpora or sub-corpora. Finally, they affirmed they could build a corpus of their own. Leaving central tendencies aside, the range of responses for each statement shows that, while all students felt fairly confident about looking at plain word queries and dealing with concordances, raw frequencies and collocations, some students were less happy about using POS tagging and normalised frequencies. One student also felt very strongly that he or she was not able to build a DIY corpus, but this may have been a student who did not attend the final sessions of the programme which focussed on corpus building.

In the third part of the survey, the students were asked to give their opinions on how helpful they found different types of corpus output. Their attitudes to concordances, word lists and collocation queries are summarised in Table 5, which shows that all corpus outputs were generally considered to be very helpful, with collocation coming out as the most helpful output according to the group as a whole.

| | |
|---|---|
| I understand the strengths and limitations of different types of corpora. | 3 strongly agree<br>**6 agree**<br>4 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I can carry out simple word queries to retrieve KWIC concordances. | 4 strongly agree<br>**9 agree**<br>0 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I can carry out queries involving more than one word. | 2 strongly agree<br>**11 agree**<br>0 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I understand the difference between looking up lemmas and looking up plain words. | 4 strongly agree<br>**8 agree**<br>1 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I can use part-of-speech tags in my queries. | 2 strongly agree<br>**8 agree**<br>1 neither agree nor disagree<br>2 disagree<br>0 strongly disagree |
| I can use corpora to retrieve information about collocation. | 5 strongly agree<br>**8 agree**<br>0 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I am able to compare the frequencies of different words or combinations of words within a corpus. | 4 strongly agree<br>**8 agree**<br>1 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I am able to use normalised frequencies to compare words across different corpora or sub-corpora. | 1 strongly agree<br>**6 agree**<br>3 neither agree nor disagree<br>3 disagree<br>0 strongly disagree |
| I am able to build a simple corpus on my own. | 5 strongly agree<br>**7 agree**<br>0 neither agree nor disagree<br>0 disagree<br>1 strongly disagree |

**Table 4**: Students' self-assessment after eleven weeks of instruction (central tendencies in bold)

| | |
|---|---|
| I find concordances helpful. | **8 strongly agree**<br>5 agree<br>0 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I find word lists (frequencies) helpful. | **7 strongly agree**<br>4 agree<br>2 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I find collocation queries helpful. | **10 strongly agree**<br>3 agree<br>0 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |

**Table 5**: Students' opinions about different types of corpus output (central tendencies in bold)

In the final part of the questionnaire the students were asked about their current uses of corpora and how they expected to use corpora in the future. First the students were asked to list the corpora they had so far used on their own outside classes. Their responses are summarised in Table 6. As shown, the students used a number of corpora on their own, especially those that they had also used in class. The only corpora not mentioned in class in the list were the Corpus of Translational Chinese (ZCTC), the Lancaster Corpus of Mandarin Chinese (LCMC) and the Babel English–Chinese Parallel Corpus, all developed at the University of Lancaster (Xiao, 2010). One of the responses given, the NYC corpus, does not seem to be a corpus at all, but could have been mistakenly used to refer to the BYU corpora developed at Brigham Young University (Davies, 2004, 2008, 2013). It is interesting to note that, when asked to list which corpora they used outside classes, some students noted they used the Sketch Engine – which is a not in itself a corpus but, rather, an interface that provides access to several different corpora. Likewise, TenTen in itelf is not a corpus, but the part of the name shared by a series of web-crawled corpora in different languages that are available through the Sketch Engine (Jakubíček *et al*., 2013). Note also that while some respondents listed EuroParl, EMEA, ECB and OpenSubtitles (Tiedemann, 2012) without stating which language sub-corpora of these multi-lingual parallel corpora they used, two students specifically referred to using EuroParl in English and Italian, and in English and French. They did not, however, specify a particular language direction.

Next, the students were asked to respond to a series of statements about their present and future uses of corpora against a five-point Likert scale. Tables 7 and 8 summarise their responses. Table 7 shows that, as a group, the students tend to use corpora more frequently when writing in a

| Corpus | No. of users |
|---|---|
| BNC [SE, BYU] | 5 |
| Sketch Engine (sic) [SE] | 4 |
| EuroParl [SE, OPUS] | 4 |
| enTenTen [SE] | 4 |
| GkWac [SE] | 3 |
| COCA [BYU] | 3 |
| EuroParl-en [SE, OPUS] | 2 |
| EMEA [SE, OPUS] | 2 |
| ZCTC | 1 |
| TenTen (sic) [SE] | 1 |
| ptTenTen [SE] | 1 |
| OpenSubtitles [SE, OPUS] | 1 |
| NYC corpus (sic?) | 1 |
| LCMC | 1 |
| itTenTen [SE] | 1 |
| EuroParl-it [SE, OPUS] | 1 |
| EuroParl-fr [SE, OPUS] | 1 |
| ECB [SE, OPUS] | 1 |
| BAWE [SE] | 1 |
| BASE [SE] | 1 |
| Babel | 1 |

**Table 6**: Corpora used by students outside classes ([SE] available through the Sketch Engine, www.sketchengine.co.uk; [BYU] available through the BYU interface http://corpus.byu.edu/; [OPUS] available through the OPUS interface, http://opus.lingfil.uu.se/)

language that is not their native language than when writing in their native language, and that they use corpora most frequently to help them with their translation assignments. A closer look at the responses specified by those students who claimed to use corpora for other purposes revealed two rather vague responses: 'To understand some collocations' and 'General'; and two very specific uses: 'to find out what are the most frequently used words in a certain area. For example, political speech' and 'Discussion on language interest groups on Facebook; for silly/informal discussions of frequencies'. Table 8, in turn, shows that the students generally plan to continue using corpora in the future, both for translation and for other purposes.

## 4. Student reports

As mentioned in the introduction, the students had been asked to hand in a report on their uses of corpora one week after the end of the teaching.

| | |
|---|---|
| I use corpora to help me when I am writing in my native language. | 0 very often<br>2 often<br>**6 sometimes**<br>3 rarely<br>2 never |
| I use corpora to help me when I am writing in a language that is not my native language. | 3 very often<br>**5 often**<br>3 sometimes<br>2 rarely<br>0 never |
| I use corpora to help me with my translation assignments. | 1 very often<br>**6 often**<br>6 sometimes<br>0 rarely<br>0 never |
| I use corpora for other purposes. | 0 very often<br>2 often<br>3 sometimes<br>**4 rarely**<br>4 never |

**Table 7**: Students' present uses of corpora (central tendencies in bold)

| | |
|---|---|
| I am likely to look things up in corpora during my translation exams or for writing my MA dissertation. | 3 strongly agree<br>**7 agree**<br>2 neither agree nor disagree<br>1 disagree<br>0 strongly disagree |
| I am likely to carry on using corpora in the future in my work as a translator. | 6 strongly agree<br>**6 agree**<br>1 neither agree nor disagree<br>0 disagree<br>0 strongly disagree |
| I am likely to carry on using corpora in the future for purposes other than translation. | 3 strongly agree<br>**4 agree**<br>5 neither agree nor disagree<br>1 disagree<br>0 strongly disagree |
| I am likely to build a corpus of my own to help me with my research or with my work in the future. | 2 strongly agree<br>**6 agree**<br>3 neither agree nor disagree<br>1 disagree<br>1 strongly disagree |

**Table 8**: Students' future uses of corpora (central tendencies in bold)

The report was a graded assessment and was divided into two parts. In the first part, the students were asked to describe how they had been using corpora in everyday translation, illustrating their account with examples from their own practice, which, as mentioned earlier, they had been asked to start collecting from the beginning of the semester. The students were given explicit instructions to describe the translation problems they encountered and the corpus queries carried out, and to explain how the latter had helped (or not) and how that had influenced their decisions as translators. In the second part, the students were asked to describe how they compiled a small *ad hoc* corpus in a specialist area of their choice and how they used it to research terminology and phraseology in the area. A 3,000-word limit for both parts of the assignment was imposed, excluding references.

All students completed the assignment. The reports were fed into a corpus totalling 47,123 running words in order to come to a better understanding of how the group as a whole had been using corpora, complementing the data obtained through the questionnaires. In addition to this, the reports were also read from beginning to end as they were marked and second-marked, during which it was possible to carry out a more fine-grained and detailed analysis of individual uses of corpora. A number of actual queries carried out by the students, as well as some of the opinions and insights they expressed, were selected in order to portray their individual uses of corpora in more detail.

The grades achieved by the students varied from 50 percent (pass) to 76 percent (distinction), showing that some reports were much better than others. However, in the analyses that follow no attempt will be made to focus on individual students. The corpus analysis in Section 4.1 is devoted to looking at the students' performance as a group, while the examples of queries carried out by the students in Section 4.2 as well as the students' opinions in Section 4.3 are intended to provide a balanced snapshot of what was revealed by the detailed reading of the student reports.

### 4.1 Corpus analysis of the student reports

The corpus analysis of the student reports was aimed at understanding how the group as a whole was using corpora and to verify whether some of the responses given in the introspective questionnaire described in Section 3 could be triangulated with what the students actually wrote in the reports.

The first exploration in this respect involved finding out whether there might be any other corpora that the students had used outside classes which the students had not mentioned in the questionnaire. A KWIC query for the lemma *corpus* was carried out and by examining the 966 concordance lines retrieved, it was possible to see that, in addition to the DIY corpora the students had been asked to compile, the students had actually used more corpora than those listed in Table 6. Separate KWIC queries were then carried

| Corpus | No. of users |
|---|---|
| **COCA** BYU | **9** |
| **BNC** SE, BYU | **7** |
| **enTenTen** SE | **7** |
| Sketch Engine (sic) SE | 4 |
| EuroParl SE, OPUS | 4 |
| **GkWac** SE | **4** |
| **EMEA** SE, OPUS | **3** |
| EuroParl-en SE, OPUS | 2 |
| **frTenTen** SE | **2** |
| ZCTC | 1 |
| TenTen (sic) SE | 1 |
| ptTenTen SE | 1 |
| OpenSubtitles SE, OPUS | 1 |
| NYC corpus (sic?) | 1 |
| LCMC | 1 |
| itTenTen SE | 1 |
| EuroParl-it SE, OPUS | 1 |
| EuroParl-fr SE, OPUS | 1 |
| ECB SE, OPUS | 1 |
| BAWE SE | 1 |
| BASE SE | 1 |
| Babel | 1 |
| **OPUS (entire fr>en)** SE, OPUS | **1** |
| **CREA (es)** | **1** |
| **COMPARA (pt<>en)** | **1** |
| **CCL (zh)** | **1** |
| **OpenOffice (en>zh)** SE | **1** |
| **ruTenTen** SE | **1** |

**Table 9**: Corpora used outside classes according to questionnaires and student reports with additions in bold (SE available through the Sketch Engine, www.sketchengine.co.uk; BYU available through the BYU interface http://corpus.byu.edu/; OPUS available through the OPUS interface, http://opus.lingfil.uu.se/)

out for each corpus cited in the reports to see how many different students had used them. An updated version of Table 6 is provided in Table 9, with the additions in bold. Note, however, that the students could have used other corpora as well, but may not have referred to them in either the questionnaire or the assignment.

A query for the lemma *concordance* then showed that in twelve reports the students had referred to concordance queries which they had carried out. A closer inspection of the only report that did not contain the word *concordance* nevertheless revealed that the student in question had

indeed looked up concordances but referred to them as 'searches' instead. All thirteen students carried out both single and multiple-word concordance queries.

Queries for the lemmas *collocation*, *collocate* and *word sketch* showed that all but one student included collocation queries in their reports. The assignment by the student who did not refer to any of those terms was quickly checked to see if she might have referred to the concept in a different way, but there was no mention of any collocation queries being carried out at all. Interestingly, the latter is at odds with the questionnaire responses in Table 5, where all students agreed or strongly agreed that collocation queries were helpful.

Next, queries for the lemmas *frequency*, *hit*, *occurrence* and *token* were carried out to inspect whether the students had used corpus frequencies in the look-ups described in their reports. The results showed that all students had indeed formulated queries that involved the analysis of frequencies. Most such queries involved checking the frequency of specific words or expressions in the same corpus. A search for the terms 'relative frequency' and 'normali(sz)ed frequency' showed that only two students in the group referred to comparing frequencies across different corpora or sub-corpora. This points in the same direction as the questionnaire responses in Table 4, which indicate that the students seemed less confident about examining normalised frequencies than they were about interpreting direct, raw frequencies.

Still on the topic of frequencies, a search for 'word list' and 'frequency list' revealed that only seven students looked up the overall distribution of words in a corpus (as opposed to looking up the frequency of specific words or expressions), which they did in relation to the DIY corpus they had built. Of these, only six students referred to the concepts of 'keyword list' and 'keyness', which enables one to extract the most salient words and expressions of a given corpus by comparing it with a general reference corpus. However, only five students actually carried out keyword analyses.

A search for *lemma* revealed that only five students explicitly referred to lemma queries (as opposed to word-form queries) in their reports. However, it should be noted that, in the Sketch Engine corpora, the default simple query automatically runs a lemma query rather than a word-form query, which means all the students carrying out simple queries in the Sketch Engine corpora were actually carrying out lemma queries, except of course for queries involving Chinese, which is a non-inflecting language.

Lemma searches for *part of speech*, *part-of-speech*, *pos* and *tag* then disclosed that while eight students had referred to the concept in their reports, only three actually used grammatical annotation in their queries. The questionnaire results in Table 4 also shows that the students were less confident about using POS tags in their searches. The above corpus analysis of the student reports is summarised in Table 10.

| | Student reports in which concept was cited | Student reports in which concept was used |
|---|---|---|
| Frequency, hit, occurrence, token | 13 | 13 |
| Concordance | 12 | 13 |
| Lemma | 5 | 13 |
| Collocation, collocate, word sketch | 12 | 12 |
| Word/frequency list | 7 | 7 |
| Keyword list, keyness | 6 | 5 |
| Part-of-speech, part of speech, pos, tag | 8 | 3 |
| Relative/normali(sz)ed frequency | 2 | 2 |

**Table 10**: Summary of corpus analysis of student reports

### 4.2 Examples of queries carried out by the students

Overall, the student reports revealed a mixture of successful and not-so-successful uses of corpora. The most common type of query referred to in the reports involved using concordances to check frequencies in mono-lingual corpora in order to find out which of two alternative forms was more conventional. For example, in an English to Chinese technical translation assignment, one student was not sure whether in her translation of *mRNA* she should keep the English form, *mRNA*, or use the Chinese form, *信使RNA* ('messenger RNA'). She looked up the frequencies of each of these terms in the Sketch Engine's zhTenTen corpus and found that the former seemed a lot more conventional than the latter, with a frequency of 1,674 against 106. This helped her decide to use the English form. In a similar type of query, a student translating from English into Greek wished to find out whether it was best to translate *ambivalent* into *αμφίθυμος* or *αναποφάσιστος*. Using the Sketch Engine's GkWaC corpus, she found there were only twenty-two hits for the word *αμφίθυμος* against 425 for *αναποφάσιστος* and said she chose to use the latter 'in order not to disturb the Greek readership with a word that is not widely used'. In both these cases, the frequency imbalance suggests that the students probably made the right choice. However, neither of the students commented on the appropriateness of the zhTenTen corpus and of the GkWaC corpus for these searches, and neither of them discussed whether they went on to analyse the concordance lines retrieved in order

to check whether the uses and contexts of the expressions in question were appropriate.

While the above two students at least had a measure of the sort of frequency imbalance that might be indicative of the more natural choice, some interpretations regarding frequencies seemed misguided. For example, a student translating from English into Greek was not sure whether to translate *tough competition* into σκληρός ανταγωνισμός ('tough competition') or έντονος ανταγωνισμός ('intense competition'). By looking at the frequencies of these terms in the GkWaC corpus, she concluded that it was better to use the latter, which had eighty-four occurrences, rather than the former, which had sixty-five hits in the corpus. However, the frequency difference between the two terms in a corpus of 150-million words does not seem to be sufficiently marked to justify this decision. More experienced corpus users would have perhaps concluded that both terms might be equally acceptable, and would have further explored the concordance lines to find out whether there could be subtle differences in usage associated with each of the terms.

Some students were not able to choose an appropriate corpus to look up answers to questions involving frequencies. For example, when translating an economic article from *Die Zeit* from German into English, a student wanted to check how typical the word *trendy* was of written English. To carry out the analysis, she chose to compare the frequency of the word in British Academic Spoken English (BASE) and British Academic Written English (BAWE). Although she was able to compare the two in terms of normalised frequencies (in view of the different sizes of the two corpora), she did not realise that for this query it would have been more appropriate to resort to a corpus of newspaper texts (e.g., the news sub-corpus of the BNC) instead of corpora of academic language.

Another problem noted with regard to checking frequencies in corpora was that some students seemed too attached to using concordance queries in cases where collocation queries would have been more appropriate. For example, a student translating an economic text from Spanish into English encountered some difficulty regarding the translation of the term *cuadro macroeconómico* ('macro-economic picture'), whose literal translation he found sounded rather unnatural. He used the enTenTen12 corpus to try out a series of concordance searches with nouns that could potentially combine with *macroeconomic* in the given context. He initially thought that *picture* would not generate many hits, so he tried *macroeconomic projection* and *macroeconomic prediction*, obtaining what he considered to be disappointingly few hits (thirty-five and nine, respectively). This made him revert to searching for the unnatural sounding *macroeconomic picture*, which to his surprise had the highest number of hits (sixty-seven), and made him decide to settle on it for his translation. The problem here was clearly an inability to see that a collocation query for *macroeconomic* would have solved his problem in a far more efficient way than the series of concordance queries based on the student's 'hunches'. A simple word

sketch for *macroeconomic* in the same corpus would have enabled him to immediately spot *macroeconomic environment*, with 714 hits, which could have been used as a translation for *cuadro macroeconómico*.

Despite the above, there were many examples of successful collocation queries. For example, a Chinese student was not sure how to best translate the adverb *inevitably* in the context of an English to Chinese business translation assignment. She explained that the word could be rendered as either *无可避免* ('unable to get rid of') or *势在必行* ('given the current situation, it must be done'). By looking at collocates of the two alternatives in zhTenTen, she found that the former collocated mostly with negative words such as *大战* ('war'), *灾难* ('disaster') and *问题* ('problem'), while the latter was generally used in a more neutral sense, with collocates such as *管理* ('regulation') and *改革* ('reformation').

In addition to collocations, one student found the automated thesaurus functionality of the Sketch Engine particularly useful to her work as a translator. She noticed that she tended to over-use the verb *allow* when translating the French verb *permettre*, and explained that when she had trouble looking for synonyms she had now got used to using corpora to arrive at alternative words and their collocates.

The students also used parallel corpora frequently. A student translating an article on stem cells from English into Greek was not sure how best to translate the word *treatment* in this context, and she found the translations supplied in dictionaries to be ambiguous. Some dictionaries translated *treatment* as αγωγή ('therapy') and others as θεραπεία ('cure'). The student therefore looked up the translation of this word in the European Medicines Agency (EMEA) corpus from the OPUS collection. She discovered that the majority of the results indicated that the established rendition of *treatment* in a medical/pharmaceutical context was θεραπεία.

Another student, this time translating a technical text about nuclear energy from French into English, explained she needed some inspiration to translate the word *pleine* in the context of 'en pleine guerre froide'. Initially, she was tempted to render it as 'in the middle of the cold war', but after looking up parallel concordances for 'en pleine guerre froide' in the French–English component of the OPUS collection she discovered 'at the height of the cold war', 'in the midst of the cold war' and 'in the throes of the cold war', which she felt were much better alternatives than her own initial option.

The reports also showed some evidence of students following up queries with further queries. The above student, for example, stated that she decided to explore further the word *midst* through a Word Sketch query in the enTenTen corpus. She noticed it collocated with many words associated with war, such as *turmoil*, *crisis*, *suffering*, *persecution*, *revolution* and *battle*, but was surprised to see that it did not collocate with *war*. The latter, however, was not really true. It simply showed that the student was not yet a proficient user of the Word Sketches, which only display a limited number of the most

significant collocates on its initial results screen. If the student had clicked on the 'more data' option, she would have been able to notice a very strong association between *war* and *midst*.

Another example of a follow-up query was given by the above mentioned Chinese student looking up suitable Chinese equivalents for *inevitably*. Her analysis of the collocates that went with the two translation candidates of the word led her to notice there was yet another option that could be used in the context: 迫在眉睫 ('very urgent and has to be done immediately').

One student in particular demonstrated a sophisticated awareness of different ways of reformulating queries so as to retrieve more useful results. For example, when translating an excerpt from a French novel into English, she used the frTenTen corpus to better understand how the French expression *histoire de* was used by native speakers of French in order to help her translate 'Histoire de présenter en position de force' (literally, 'Story of presenting in a position of strength'). Her initial query for *histoire de* returned mostly concordances with *histoire* in the sense of 'history', which was not very helpful. She therefore decided to insert a comma before *histoire*, in order to get results for the expression in the context of relative clauses. She was then able to retrieve exactly what she was looking for and supply the translation 'It was all about appearing in a strong position'.

On another occasion, this same student wanted to find the most suitable collocate to translate the French phrase 'Victoire total' (literally, 'total victory'). She used the BYU-BNC to look up collocates of *victory*, but the search brought up the adjectives and modifiers *Labour*, *great*, *Conservative* and *final*, which was clearly not what she needed. She therefore reformulated her query by looking for synonyms of *total* in the context of *victory* by typing in '[=total] victory'. This resulted in *complete*, *aggregate*, *unreserved* and *absolute*. She then decided to refine the search even further by looking for adjectives similar to *absolute*, typing in '[=absolute] victory'. This yielded *final*, *outright*, *total*, *complete* and *conclusive*, and *outright* immediately struck her as being the best option.

Having been asked explicitly to compile a DIY corpus to research the terminology and phraseology of a specialised domain of their choice, most students were able to describe the compilation process in detail, but only a few seemed to have understood the need for filtering provenance with regard to corpus files automatically retrieved by crawling the Web. On the other hand, a number of students reported on the useful information they were able to retrieve from their DIY corpora. For example, a student enrolled in German into English business translation built her own English corpus of different types of companies to be able to research specialised terminology in this domain. She then described how she used her corpus to research how the word *liability* was used and was able to arrive at terms such as 'joint liability', 'non-current liabilities', 'interest-bearing liabilities' and so on, which she then added to her glossary of business terminology.

A student who had built a corpus about the space industry noted that the words naturally occurring with *microwave* in her corpus were all scientifically based, more specifically in terms of radar and satellite communication. She explained this was very useful, because when looking up *microwave* in corpora of general English such as COCA and the enTenTen corpus, most of the collocates of the word had to do with cooking.

A Chinese student who back home was required to work out of her native language reported that she compiled a small specialised English corpus about cranes to assist her in the translation of user manuals about cranes. One of the examples she gave was about how she used the corpus to find suitable collocates of *load*, in order to translate the sentence '用履带式起重机将重物吊起' ('lift the load with the crawler crane'). According to her corpus, *load* could be preceded by both *lift* and *hoist*, but the former appeared to be more conventional than the latter.

A Greek student decided to build a microbiology corpus in English and a comparable corpus in Greek because she had been asked to translate a series of articles on the topic for her technical translation classes. She noted *infection* was very a frequent word in the English corpus and used the Greek corpus to look up its equivalent in Greek, after remarking that a bilingual dictionary had presented two options *μόλυνση* and *λοίμωξη*, and she was not sure which one to employ. With the corpus, she was able to find out that *λοίμωξη* was more appropriate in the domain of microbiology.

Few students had the initiative of consulting more than one corpus to address a single translation question. I have already given the example the student who compared the frequency of the word *trendy* in BAWE and BASE to find out whether it was appropriate to use it in the translation of news article. There was, however, a particularly perceptive analysis carried out by a student translating a short story from Russian into English. She explained how she used different mono-lingual corpora to help her decide whether she should add some extra information to an excerpt of the translation to make it more accessible to a target English audience. The problem in question was the sentence 'Ира – это не девушка – а мальчик' ('Ira – not a girl – but a boy'). She clarified that a Russian reader would expect *Ira* to be a girl's name, while a British reader might be a bit perplexed because there are not many associations with the name *Ira* in Britain, while Americans might think of *Ira* as a man's name, after names like Ira Gershwin. She thus proceeded to check across three mono-lingual corpora the frequency and context for the name Ira. In the Russian ruTenTen corpus, she established that *Ira* was very common and always used in the context of a female. To confirm this, she used *Ira* with a male form of the verb *be*, and found no results. Next, she looked up *Ira* in the BNC and, in the middle of a large number of references to the IRA (Irish Republican Army),[14] she found a small proportion of occurrences of

[14] Which, incidentally, could have been automatically excluded by carrying out a word-form query instead of a lemma query.

Ira as a man's name. Finally she looked up *Ira* in COCA, and was able to see that it was indeed a common but only male name. This convinced the student that she had to make the translation more explicit for English readers, and came up with 'Ira was not, as the name seemed to imply, a girl, but a boy'.

### 4.3 Students' opinions

The reports by the students contained not only examples of how they had used corpora, but also their views about them. A selection of verbatim quotes by the students are presented below.

In their opinions about different types of corpora, they seemed generally happy about large mono-lingual corpora like the BNC and COCA, and the TenTen family of corpora, but they had diverse opinions with regard to parallel and DIY corpora. While one student observed that 'Of all the types of corpora available, parallel are undoubtedly the easiest for translators to draw conclusions from because the necessary information can be accessed immediately and terms can be directly compared to their equivalents in another language', another student complained that 'The parallel corpus often produced few results'. Of course, these two views are not contradictory, for while parallel corpora can provide immediate and easy to interpret answers to translation queries, they tend to be much smaller and specialised than large mono-lingual corpora. As Frankenberg-Garcia (2009: 60) notes, 'Only a very small part of what people in general say or write ever gets to be translated, which seriously limits the number and types of texts available for the compilation of parallel corpora. Indeed, this is one of the main reasons why parallel corpora are usually much smaller in scale than mono-lingual corpora.'

With regard to the DIY corpora they had been asked to compile, one student said that, 'Although my corpus was put together in only a matter of minutes, it still allowed me to study terminology and phraseology related to astronomy in a reasonable amount of depth', while another one remarked that 'I find that compiling corpora is more suitable for researchers, linguists and teachers, rather than translators and interpreters'.

There were some students who commented on the difficulty of becoming acquainted with corpora: 'The translator spend [sic] a huge amount of time familiarise [sic] him or her with the tool and then spend extra effort on mastering the code and tag language these things [sic], but he or she may never use some of the functionalities in a corpus [sic]'; 'Overall, it has been a useful resource but has been limited by my relative inexperience of applying the available functions and occasional searches taking too long'; 'the use of corpora [. . . ] takes some time to get used to but has proved to be a good resource for translation practice'.

One student found using corpora could be distracting: 'One thing that can make using corpora time-consuming is that once concordances are

begun, in my experience, I can find myself looking further and often find interesting things out that I wasn't looking for in the first place, which isn't necessarily a negative observation.'

There were several comments about coping with raw corpus data as opposed to the polished language of dictionaries. One student observed that corpora contained language mistakes: 'errors crop up from time to time as I discovered when trying to make a concordance for the English noun "attention". I accidentally made a typo in the spelling, typing out "attenton", missing the "i" [… and ] retrieved sixty-two results of the misspelling of attention, nonetheless, reinforcing the fact that corpora really do represent real language use, mistakes included'. Another student felt corpora should be used in conjunction with dictionaries: 'Although corpus is [sic] highly informative, it is no substitute for other authoritative resources like dictionaries. A better solution would be to combine them both and utilise the advantages of both.' And one student emphasised that corpora complemented dictionaries, but needed to be used with care: 'Compared to dictionaries, they [corpora] offer translators with extensive genuine examples in various contexts, thus can be a powerful complementary tool for understanding the usage of language. However, it is also noted that translators should be careful with their own interpretations for [sic] data presented by corpora and examine the reliability of some examples in corpora before making further analysis.'

Finally, there were several positive comments about the overall usefulness of corpora:

> 'using my comparable [DIY] corpora saved me time and effort.'
> 'it has given unexpected insights on the native language and a showed to be a [sic] precious resource especially in regards with working into a non-native language, in this case English, during the writing of essays.'
> 'Producing an authentic-sounding TT is, however, especially difficult when you are working out of your native language and I therefore found corpora to be especially useful when translating a text about an Aztec artefact from English into my non-native language, German.'
> 'corpora certainly possess the potential to be excellent resources for my translation projects and I will continue learning how to use them effectively.'
> 'I have found that corpora have been most useful to me when dealing with issues of collocation.'
> 'Corpora can be useful, not only for translating, but also for the writing of essays and reports.'
> 'corpora either monolingual, multilingual, general, specialized, comparable, parallel or not, have always been my ally in tackling translation challenges. Despite the fact that they might have failed to help me in some cases, I still consider them really helpful when used in the correct way and I recommend them to any translator or to anyone who just wants to explore how languages function.'

## 5. Discussion and conclusion

Before I had the opportunity to analyse the students' responses to the questionnaire and their assignments, my overall impression of the module was that it had been successful in teaching the students the basics of the applied uses of corpora in translation practice. However, it was very frustrating that not enough attention could be given to the use of corpora to address concrete translation problems. Not only were different students working with different language pairs, but also there was not enough time to address specific translation problems in class. In a nutshell, the students themselves commented that 'It's very useful to learn about how to use corpora', but 'we should spend more time working with corpora in actual translations'. This same feeling was also observed by Rodríguez-Inés and Hurtado Albir (2012) in a survey with translation students at the University of Barcelona.

There were also problems teaching the students to use POS tags in their queries, which is not something that is easy and intuitive. The fact that different corpora are coded with different POS tags means that getting used to the tags that go with one corpus will not help much when trying to use tags with another corpus. Even when using the Sketch Engine corpora, where regardless of the corpus used the search routines remain the same, teaching the students to use POS tags for the English corpora did not automatically help them to use POS tags in corpora of other languages. And, indeed, despite their responses in the questionnaire stating that, on the whole, the students agreed that they could use POS tags in their queries, only three students exemplified their use of such tags in actual queries in their reports.

Teaching the Chinese students to transpose what they had learnt with the English corpora in the Sketch Engine to the Chinese zhTenTen corpus proved to be particularly challenging when we discovered that, unlike for other languages, multiple-word queries for Chinese did not work in the Sketch Engine's simple query option. This meant these students had to be taught to use the more complex CQL query language separately. To complicate things further, the CQL query language requires the students to surround separate words with double quotation marks of the English keyboard, which meant the students had to keep switching back and forth between the English and the Chinese keyboards all the time. Still, despite these glitches, the Sketch Engine proved to be an excellent means of providing a multi-lingual group of students with access to large general corpora in different languages using the same interface. This also enabled the students to transpose most of what they had learnt through the Sketch Engine English corpora to corpora of other languages distributed by the Sketch Engine.

The benefits and challenges noted during the actual teaching of the module seem to be reflected in both the students' responses to the end-of-course questionnaire and in the students' assignments. According to the questionnaire, all but one student had never heard of corpora before coming

to study at Surrey, and the only student who had heard of corpora did not seem to have had much practice in using them.[15] In contrast, by the end of the course they generally agreed that they could perform all basic corpus operations they were taught about in class; they found using concordances, comparing frequencies and using collocations helpful; they were able to use a variety of corpora on their own outside classes, including corpora that had not been seen in class; they were often using corpora in their translation assignments and sometimes to help them write – particularly when writing in a language that was not their native language; and they intended to carry on using corpora in the future. Despite the fact that the responses to the questionnaire were generally very positive, there was some variation with regard to how confident the students felt about using POS tags and normalised frequencies, and about building DIY corpora.

The analysis of the student assignments then showed that some students had grasped the basics of corpora better than others and that some students seemed to be under-using corpora while others were using them rather well. It also became apparent that some aspects of using corpora would have benefitted from more support from the teacher, especially with regard to using POS tags, interpreting frequencies, comparing frequencies across different corpora and sub-corpora, following up initial queries with further analyses, and extracting word lists and keyword lists from DIY corpora.

It is interesting to note that the most common type of query carried out by the students involved checking the frequencies of different translation options against corpora in the target language in order to determine which one seemed more conventional, which is in a sense similar to checking the frequency of search results using a search engine.[16] However, the examples in Section 4.2 show that corpora were also used for a variety of other purposes, too, especially analysing collocations – and this is not something that can be done easily or systematically using a search engine. The reports also showed several other examples of queries for which dictionaries, glossaries and web searches and other more conventional resources would not have provided satisfactory answers. Having said this, it is important to note that it was not possible, of course, to analyse what the students left out of their reports. With a limit of 3,000 words for their assignments, there may have been many other queries and details about queries which they simply did not have room to describe.

---

[15] This could be interpreted as a sobering reminder that people who claim to know about corpora (see Bernardini, 2006; Gallego-Hernández, 2015; and Gough, 2013) may in fact know very little about them.

[16] The advantages of using corpora, of course, are that provenance is easily traceable, frequency counts are stable and exclude repeated texts (and are, thus, more reliable), KWIC output is more informative than web snippets, and, for many corpora, it is possible to look up word inflections and resort to POS tags to refine queries.

The students' opinions of corpora were generally very favourable, although they did comment on the difficulty of mastering the use of corpora. This seems to corroborate Aston's (2009) previously mentioned assertion that the medium- and long-term advantages of using corpora can over-ride the steep learning curve that is required in the beginning. Of course, it will only be possible to actually test whether translators can benefit from corpora when translators are able to use corpora effectively. The students' intention of continuing to use corpora in the future is nevertheless very positive, and, as with any other new technology, it is likely that the more they use corpora the better they will be able to use them. An interesting follow-up would be to contact these students in a few years' time and ask them if they have continued to use corpora.

To conclude, this study pointed out aspects of the module that can be improved the next time it is taught, and I hope it can also raise awareness of the challenges and possible advantages of teaching translation students to use corpora, despite the lack of incentive from the translation industry.

## Acknowledgments

I would like to thank my two anonymous reviewers for their comments and valuable suggestions on a previous version of this paper.

## References

Aston, G. 1999. 'Corpus use and learning to translate', Textus 12, pp. 289–314.

Aston, G. 2009. 'Foreword' in A.P. Beeby, P. Rodríguez-Inés and P. Sánchez-Gijón. (eds) Corpus Use and Translating, pp. ix–x. Amsterdam and Philadelphia: John Benjamins.

Baker, M. 1993. 'Corpus linguistics and translation studies: implications and applications' in M. Baker, G. Francis and E. Tognini-Bonelli (eds) Text and Technology: In Honour of John Sinclair, pp. 233–50. Amsterdam and Philadelphia: John Benjamins.

Baroni, M., A. Kilgarriff, J. Pomikalek and P. Rychly. 2006. 'WebBootCaT: instant domain-specific corpora to support human translators', Proceedings of EAMT-2006, pp. 247–52.

Beeby, A., P. Rodríguez-Inés and P. Sánchez-Gijón (eds). 2009. Corpus Use and Translating. Amsterdam and Philadelphia: John Benjamins.

Bernardini, S. 2006. 'Corpora for translation education and translation practice: achievements and challenges', Proceedings of the Third International Workshop on Language Resources for Translation

Work, Research and Training (LR4Trans-III). Available online at: http://mellange.eila.jussieu.fr/bernardini_lrec06.pdf

Bick, E. 2005. 'CorpusEye: Et brugervenligt web-interface for grammatisk opmærkede korpora' in P. Widell and M. Kunøe (eds) 10. Møde om Udforskningen af Dansk Sprog 7.–8.okt.2004, Proceedings, Århus University, pp. 46–57.

Bowker, L. and J. Pearson. 2002. Working with Specialized Language: A Practical Guide to Using Corpora. London: Routledge.

Davies, M. 2004. BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at: http://corpus.byu.edu/bnc/

Davies, M. 2008. The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at: http://corpus.byu.edu/coca/

Davies, M. 2013. Corpus of Global Web-Based English: 1.9 Billion Words from Speakers in 20 Countries. Available online at: http://corpus2.byu.edu/glowbe/

Frankenberg-Garcia, A. 2009. 'Compiling and using a parallel corpus for research in translation', International Journal of Translation XXI (1), pp. 57–71.

Frankenberg-Garcia, A. 2012. 'Raising teachers' awareness of corpora'. Language Teaching 45 (4), pp. 475–89.

Frankenberg-Garcia, A. 2014. 'Understanding Portuguese translations with the help of corpora' in T. Sardinha and T. Ferreira (eds) Working with Portuguese Corpora, pp. 161–76. London: Bloomsbury.

Gallego-Hernández, D. 2015. 'The use of corpora as translation resources: a study based on a survey of Spanish professional translators', Perspectives: Studies in Translatology 23 (3), pp. 375–91. DOI: 10.1080/0907676X.2014.964269.

Gough, J. 2013. Survey of professional translators' use of on-line resources for terminology research. Unpublished interim PhD report. September 2013. University of Surrey.

Jakubíček, M., A. Kilgarriff, V. Kovář, P. Rychlý and V. Suchomel. 2013. 'The TenTen Corpus Family', paper presented at 7th International Corpus Linguistics Conference. July 2013. Lancaster.

Kilgarriff, A., V. Kovar and A. Frankenberg-Garcia. 2013. 'Bilingual word sketches: three flavours', paper presented at Electronic Lexicography in the 21st Century: Thinking outside The Paper (eLex 2013). 17–19 October 2013. Tallinn, Estonia.

Kilgarriff, A., P. Rychlý, P. Smrz and D. Tugwell. 2004. 'The Sketch Engine', Proceedings of Euralex, pp. 105–16. Lorient, France.

Kubler, N. 2011. 'Working with corpora for translation teaching in a French-speaking setting' in A. Frankenberg-Garcia, L. Flowerdew and G. Aston (eds) New Trends in Corpora and Language Learning, pp. 62–80. London: Bloomsbury.

Philip, G. 2009. 'Arriving at equivalence: making the case for comparable general reference corpora in translation studies' in A. Beeby, P. Rodríguez-Inés and P. Sánchez-Gijón (eds) Corpus Use and Translating, pp. 59–73. Amsterdam: John Benjamins.

Rodríguez-Inés, P. and A. Hurtado Albir. 2012. 'Assessing competence in using electronic corpora in translator training' in M. Borodo and S. Hubscher-Davidson (eds) Global Trends in Translator and Interpreter Training: Mediation and Culture, pp. 96–126. London: Continuum.

Tiedemann, J. 2012. 'Parallel data, tools and iInterfaces in OPUS', Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012), pp. 2214–18. 21–7 May. Istanbul, Turkey.

Tonnini-Bonelli, E. and E. Manca. 2004. 'Welcoming children, pets and guests: towards functional equivalence in the languages of "Agriturismo" and "Farmhouse holidays"', TradTerm 10, pp. 295–312.

Varantola, K. 2003. 'Translators and disposable corpora' in F. Zanettin, S. Bernardini and D. Stewart (eds) Corpora in Translator Education, pp. 55–70. Manchester: St Jerome.

Xiao, R. 2010. 'How different is translated Chinese from native Chinese?: a corpus-based study of translation universals', International Journal of Corpus Linguistics 15 (1), pp. 5–35.

Zanettin, F. 2012. Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies. Manchester: St. Jerome.

Zanettin, F., S. Bernardini and D. Stewart (eds). 2003. Corpora in Translator Education. Manchester: St. Jerome.

**Appendix A:** Corpora used in class

**Mono-lingual**

BASE (British Academic Spoken English Corpus)[SE]

BAWE (British Aacademic Written English Corpus)[SE]

BNC (British National Corpus)[SE, BYU]

BLC (Business Letter Corpus), see: http://www.someya-net.com/concordancer/

COCA (Corpus of Contemporary American English)[BYU]

Sketch Engine web-crawled corpora:[SE]

enTenTen (English)
zhTenTen (Chinese)
itTenTen (Italian)
ptTenTen (Portuguese)
frTenTen (French)
deTenTen (German)
esTenTen (Spanish)
ruTenTen (Russian)
GkWaC (Greek)

**Parallel**

COMPARA (bidirectional corpus of Portuguese and English fiction), see: www.linguateca.pt/COMPARA

OPUS collection of parallel corpora (especially EuroParl, EMEA, OpenSubtitles and European Central Bank)[SE, OPUS]

SE Available through the Sketch Engine, see: www.sketchengine.co.uk
BYU Available through the BYU interface, see: http://corpus.byu.edu/
OPUS Available through the OPUS interface, see: http://opus.lingfil.uu.se/

**Your short guide to the EUP Journals Blog http://euppublishingblog.com/**

***A forum for discussions relating to Edinburgh University Press Journals***

## 1. The primary goal of the EUP Journals Blog

To aid discovery of authors, articles, research, multimedia and reviews published in Journals, and as a consequence contribute to increasing traffic, usage and citations of journal content.

## 2. Audience

Blog posts are written for an educated, popular and academic audience within EUP Journals' publishing fields.

## 3. Content criteria - your ideas for posts

We prioritize posts that will feature highly in search rankings, that are shareable and that will drive readers to your article on the EUP site.

## 4. Word count, style, and formatting

- Flexible length, however typical posts range 70-600 words.
- Related images and media files are encouraged.
- No heavy restrictions to the style or format of the post, but it should best reflect the content and topic discussed.

## 5. Linking policy

- Links to external blogs and websites that are related to the author, subject matter and to EUP publishing fields are encouraged, e.g.to related blog posts

## 6. Submit your post

Submit to ruth.allison@eup.ed.ac.uk

If you'd like to be a regular contributor, then we can set you up as an author so you can create, edit, publish, and delete your *own* posts, as well as upload files and images.

## 7. Republishing/repurposing

Posts may be re-used and re-purposed on other websites and blogs, but a minimum 2 week waiting period is suggested, and an acknowledgement and link to the original post on the EUP blog is requested.

## 8. Items to accompany post

- A short biography (ideally 25 words or less, but up to 40 words)
- A photo/headshot image of the author(s) if possible.
- Any relevant, thematic images or accompanying media (podcasts, video, graphics and photographs), provided copyright and permission to republish has been obtained.
- Files should be high resolution and a maximum of 1GB
- Permitted file types: *jpg, jpeg, png, gif, pdf, doc, ppt, odt, pptx, docx, pps, ppsx, xls, xlsx, key, mp3, m4a, wav, ogg, zip, ogv, mp4, m4v, mov, wmv, avi, mpg, 3gp, 3g2*.