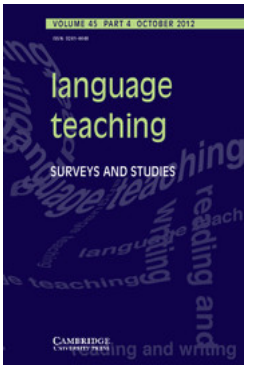

Raising teachers' awareness of corpora

Ana Frankenberg-Garcia

Language Teaching / Volume 45 / Issue 04 / October 2012, pp 475 - 489
DOI: 10.1017/S0261444810000480, Published online:

Lang. Teach. (2012), **45.4**, 475–489 © Cambridge University Press 2010
doi:10.1017/S0261444810000480 First published online 6 December 2010

Raising teachers' awareness of corpora

Ana Frankenberg-Garcia Instituto Superior de Línguas e Administração – Lisboa
ana.frankenberg@gmail.com

The last couple of decades have seen a dramatic increase in corpus availability and a steady growth in the number of supporters of the use of corpora in language teaching. Yet there still seems to be a long way to go before corpora can be understood and used by language teachers in general. Novice corpus users often fail to grasp that corpora do not work in the same way as the more familiar language learning resources – such as dictionaries, grammar books and textbooks – that they are accustomed to using. I therefore propose a series of task-based, consciousness-raising exercises to help teachers (who are not corpus linguists) understand the basics of corpora. The tasks proposed are not about learning how to use a specific corpus or software, but about learning how to use corpora in general.

1. Language teachers and corpora

Many language teachers in the EFL context have been using corpus-based dictionaries, grammars and textbooks for some time now without actually knowing what a corpus is. Indeed, understanding corpora and how to use them is not strictly necessary when the language teaching community resorts to off-the-peg corpus-based materials designed by experts. The picture changes when teachers are encouraged to handle corpora themselves, to promote data-driven learning in classroom activities such as those put forward by Johns & King (1991), Tribble & Jones (1997), Aston (2001), Sinclair (2004), Bennet (2010), Reppen (2010), and many others.

In spite of the enthusiastic acceptance of the few language teachers who have managed to break the barrier and have put their hands on corpora, only a very small number of teachers have actually tried using corpora directly. In the survey that Tribble circulated on the Linguist List (www.linguistlist.org) and on Corpora (http://torvald.aksis.uib.no/corpora/) ten years ago, only 52.8% of the 89 respondents declared they had used corpora in their teaching (Tribble 2001). This percentage is extremely low if we bear in mind that the readers of these lists are far more likely to know about corpora than the average language teacher. In what seems to be the only available well-documented survey on the use of corpora in actual schools, Mukherjee interviewed 248 secondary school teachers in Germany and found that nearly 80% had never even heard of corpora (Mukherjee 2004). And although the number of free,

Revised and updated version of a plenary address given at the 7th Teaching and Language Corpora conference, Bibliotheque National de France, Paris, 1–4 July 2006.

online corpora that can be used by anyone with access to computers and the Internet has taken a giant leap over the past few years (Anderson & Corbett 2009), the fact that several scholars continue to show concern over the still very limited use of corpora beyond a small community of experts reinforces the notion that there is a long way to go before corpora can be understood and used by language teachers in general (Mukherjee 2004, 2006; Braun 2005; Römer 2006, 2009; Boulton 2009; Breyer 2009; Frankenberg-Garcia 2010; Gilquin & Granger 2010).

Yet corpora are far from being a secret closely guarded by a restricted community of corpus linguists. On the contrary, there is a wealth of information on the specific uses of corpora in language teaching and a growing body of literature that strives to make corpus resources in general accessible to non-experts. This includes: (1) books about using corpora in language teaching (for example, Johns & King 1991; Tribble & Jones 1997; Burnard & McEnery 2000; Aston 2001; Kettemann & Marko 2002; Aston, Bernardini & Stewart 2004; Sinclair 2004; Hidalgo, Quereda & Santana 2007; O'Keeffe, McCarthy & Carter 2007; Aijmer 2009; Anderson & Corbett 2009; Bennet 2010; Reppen 2010; Frankenberg-Garcia, Flowerdew & Aston 2011; Kübler forthcoming); (2) online introductions to corpora in language teaching (see examples in appendix); (3) corpus-specific tutorials (see examples in appendix); and (4) countless conference papers and journal articles.

One of the generally acknowledged reasons why the direct use of corpora in language teaching has not caught on is that the majority of corpus resources are neither pedagogically oriented nor user friendly. In spite of the recent development of a few remarkable projects like the SACODEYL corpus (Widmann, Khon & Ziai 2011) and IFAConc (Kaszubski 2011), most corpora that can be used in the classroom were originally conceived for research rather than for teaching purposes. In this presentation I would like to argue that another major drawback is that so far little attention has been paid to training teachers in basic corpus skills.

2. Novice corpus users

With a limited number of teachers using corpora, it comes as no surprise that there do not seem to be any studies of this particular kind of user behaviour. Some of the difficulties encountered by novice corpus users in general are, however, described by Bernardini (2000), Kennedy & Miceli (2001), Frankenberg-Garcia (2005), Bianchi & Manca (2006) and Santos & Frankenberg-Garcia (2007). Although these studies differ quite substantially from one to another, they all converge to suggest that corpus skills that come as second nature to experts are not at all obvious to the untrained. Apart from corpus-specific difficulties in handling different search interfaces and query languages – and the human–computer interaction issue should not be overlooked – these studies bring to light a number of very basic problems that novice users encounter, no matter which corpus or concordancer they use.

To begin with, novice users do not always know how to choose between different types of corpora. They tend to overlook factors as seemingly obvious as provenance and publication dates, what text types are represented and how big different corpora are. In classroom observations of undergraduates using corpora in applied translation, for example, I have come across students choosing the British National Corpus (BNC), a general language corpus, to

search its concordances for highly technical terms such as *electrostatic precipitator*, or for looking up neologisms such as *Bluetooth technology*, which were not yet in use when the BNC was compiled in the early nineties. In her description of language students exploring corpora for the first time, Bernardini (2000) noted that the students were content to use the full BNC instead of narrowing it down to the specific sub-corpora that were relevant to their particular queries. Paradoxically, in our analysis of the log files of COMPARA – a much smaller, three-million-word parallel corpus of English and Portuguese (Frankenberg-Garcia & Santos 2003) – Diana Santos and I were able to observe that certain users tended to limit their searches to minute sub-corpora that did not return any hits for their queries, although they would have found relevant and useful hits if they had chosen to use less restrictive sub-corpora (Santos & Frankenberg-Garcia 2007).

Novice users also appear to experience considerable difficulties in formulating corpus queries. For example, in the same study of log files from the COMPARA corpus (Santos & Frankenberg-Garcia 2007), we found records of queries reflecting serious misconceptions about the kind of information that can be retrieved from a corpus, including queries as absurd as the string *this still did not give me the happiness I thought it would or for which I sought*. Logs with queries such as *water shining*, *bill quantities* and *like a manor* also suggest that people who are new to corpora have very little idea of the way chunks of words behave. Indeed, an in-depth analysis of 1000 queries returning null results in this corpus shows that only around 20% had actually been plausible queries. Some of the problems identified could be traced back to users applying dictionary and web browsing strategies to corpora, but most of them seemed to reflect a complete failure to grasp the very basics of corpus exploration (Santos & Frankenberg-Garcia 2007).

Existing studies on corpus-user behaviour also report that many of the queries by newcomers to corpora are either too general or too restrictive, and that apprentice users will often give up searching instead of trying to reformulate their queries. For example, in their classroom observations of undergraduates using corpora, Kennedy & Miceli (2001) noted that the students rarely carried out follow-up queries in order to refine their searches. Likewise, when analysing log files from COMPARA to check what happened after a user obtained a null result, we found that in almost half the user sessions we examined, people simply ended the session instead of trying out a modified or an altogether different query (Santos & Frankenberg-Garcia 2007).

Another important point to bear in mind is that newcomers to corpora may find it difficult to interpret corpus data. My classroom observations of students using corpora directly show that they usually have no qualms about accepting as positive evidence results returning just one or two hits. If they see something in the British National Corpus, they tend to assume it corroborates typical English usage, even when there is not enough evidence to attest this. It also seems very easy for novice users not to take corpus size and composition into account when interpreting their results. In an end of term exam, I asked a group of Portuguese undergraduate translation students to compare the distribution of the misspelled Portuguese verb ∗*caiem* 'fall' in CETEMPúblico, a Portuguese newspaper corpus of 180 million words and in DiaCLAV, a smaller comparable corpus of 6.7 million words (both available online at www.linguateca.pt). The larger corpus had 44 instances of this spelling mistake and the smaller one had six. However, only one out of sixteen students managed to

conclude that CETEMPúblico contained comparatively fewer mistakes. All other students made the simplistic assumption that the larger corpus was less accurate.

In a study where I examined how students used dictionaries, corpora, terminology databases and other language resources during a translation task, I noted that the students found it difficult to deal with unedited language and felt more comfortable using resources mediated by lexicographers and terminologists than corpora, which require autonomous user interpretation (Frankenberg-Garcia 2005). In this same study, I compared the students' perceptions of the usefulness of corpora with the effects of corpus lookups on their translations. The few students who had used corpora to assist them with their translations mistakenly felt that corpora had helped them more than they had actually done. Similarly, Kennedy & Miceli (2001) wrote that their students were frequently lured by misleading near-matches in their results and tended to make a summary analysis of the concordance lines retrieved, forgetting to pay attention to both co-texts and contexts. Bianchi & Manca (2006) also observed this behaviour and reported that their students often failed to enlarge a concordance word span in order to arrive at a more significant analysis, or even to select a relevant concordance line for analysis in the first place.

3. Developing basic corpus skills

The findings and observations I have described suggest that language teachers who are new to corpora may find it difficult to grasp that corpora do not work in the same way as the familiar language learning resources – such as dictionaries, grammar books and textbooks – that they are accustomed to using. Although most corpora provide detailed information about their composition, non-experts cannot be expected to understand the full significance of this without being explicitly taught how to do so.

To begin with, there is the question of selecting an appropriate corpus. While most people are aware of the differences between large and small dictionaries – indeed, they come in different sizes and weights – the distinction between a corpus of 100 thousand words and a corpus of 100 million words is not readily noticeable, nor are the implications of such a distinction obvious. Awareness that different corpora use different criteria for text selection is also important. Teachers are familiar with what distinguishes monolingual from bilingual from specialized language dictionaries, but the differences between corpora of newspapers, fiction, business letters, parliamentary debates and general language may not be immediately visible to them. Likewise, the distinctions between corpora of spoken and written texts or between monolingual and multilingual corpora can be just as elusive.

The ability to devise reasonable corpus queries is also something that cannot be taken for granted. Language teachers who feel comfortable looking up headwords in dictionaries and are familiar with using search engines on the Web – where it is actually very difficult not to come up with some sort of search result – may not be aware that the principles underlying corpus exploration are different. They may need help in developing the basic skills required to formulate meaningful and plausible queries.

And, of course, there is the question of interpreting corpus results. Teachers who are used to dealing with language learning materials that have been carefully selected and edited by

experts cannot be assumed to know how to handle raw data samples containing language mistakes and idiosyncrasies, too many or not enough hits, and unforeseen findings that can go against their intuitions. We cannot presume that teachers will be automatically able to derive reasonable conclusions from all the information provided by corpora without any training. As corpus output is very different from the polished materials normally used in the classroom, teachers may need some help to decode the results of their initial corpus searches.

4. Task-based, consciousness-raising corpus exercises

In the remaining part of this presentation I would like to propose a few consciousness-raising exercises that can help language teachers assess different corpora, develop basic corpus-searching strategies, and begin to learn how to interpret corpus data. The exercises proposed are task-based and, unlike most corpus tutorials available, they are not corpus-specific. Rather, they were conceived for different corpora that are available online and do not require users to pay subscription fees, so they are easily accessible to anyone with a computer and an Internet connection. Although similar exercises can be devised for practically any corpus, those presented here make use of the following English language corpora and multilingual corpora containing English: the BNC via the Simple Search service provided by the British Library and Mark Davies' BYU interface (Davies 2004), the Corpus of Contemporary American English (COCA), also developed by Mark Davies (Davies 2008), Collins Wordbanks Online (CWO), also known as the COBUILD Concordance and Collocation Sampler, the Business Letter Corpus (Someya 2000), COMPARA (Frankenberg-Garcia & Santos 2003) and the EuroParl Corpus via the OPUS interface conceived by Jörg Tiedemann (Tiedemann 2009). The exercises proposed focus on some of the fundamental principles of corpus querying in general, rather than on the capacities and the query languages of the search interfaces of each corpus in particular.

4.1 Understanding different corpora

Perhaps the most basic way of helping novice corpus users to become aware of the implications of using different types of corpora is to ask them to take a few different types of corpora, read the details about the texts they contain, and compare them. Explicit questions on the size and composition of each corpus may suffice to help people who are not familiar with corpora to understand that, just as there are different types of dictionaries, corpora too may – and often do – differ. I provide an example of this very simple consciousness-raising exercise in Table 1.

Another way of helping novice users to assess different types of corpora is to ask them to repeat the same query in a number of different corpora and compare the results obtained. The task-based exercise I present in Table 2 is an example of how this can be achieved. The different types of keywords selected for comparison were deliberately chosen so that they (a) would highlight important differences between the different corpora in analysis and (b) were single words so that novice users would not, at this stage, be confused by different query languages.

Table 1 *Comparing corpora I*

The table below contains information on six different English language corpora or sub-corpora.
a. Which is the largest corpus?
b. Which is the smallest?
c. Which corpora contain spoken language?
d. Which ones are made up of writing alone?
e. Which corpora include translated English?
f. Which corpus is most likely to include English by non-native speakers?
g. Which corpus includes non-contemporary English?

(Sub-)Corpus	**Size in words** (m = million)	**Type of English**	**Time**
Corpus of Contemporary American English (COCA)	450m (in 2010)	General American	1990–present (regularly updated)
British National Corpus (BNC)	100m	General British	Early to mid-1990s
Collins Wordbanks (CWO) (spoken British English sub-corpus)	10m	Spoken British	1990s
EuroParl (English sub-corpus)	26m	British, Irish, international and translated (EP debates)	1998–2003
COMPARA corpus (English sub-corpus)	1.5m	Original and translated fiction	1837–2002
Business Letter Corpus (BLC)	1m	US and UK business letters	1990s

Just as dictionary users might want to try out a number of different look-ups to get a feel for a given dictionary, new corpus users can be encouraged to try out different queries in a specific corpus in order to obtain a measure of its capacities and limitations. In Table 3 I give an example of a task-based exercise aimed at helping new corpus users become familiar with the contents of a specific corpus.

4.2 Formulating corpus queries

Apart from the need to become aware of what different corpora are composed of, novice users can also benefit from being explicitly taught how to formulate basic corpus queries.

4.2.1 Single words

People who are learning to use corpora should perhaps realize first of all that some corpus-browsing software, unlike many search engines and search facilities in word processors, is by default sensitive to upper and lower case distinctions, while others may require users to

Table 2 *Comparing corpora II*

Carry out a search for the words in italics in the corpora described in Table 1 and complete the table below to keep a record of how many times they appear in each corpus.

Type of word	Search word	COCA	BNC	CWO (speech)	EuroParl (English)	COMPARA (English)	BLC
old-fashioned	*counterpane*						
new	*MP3*						
common	*with*						
rare	*epicure*						
typically oral	*yeah*						
literary	*amiable*						
technical	*pelagic*						
regional	*lass*						
sentimental	*darling*						
religious	*rosary*						
political	*coalition*						
foreign	*rapporteur*						

a. Which are the only corpora that have *counterpane*? Think of some more old-fashioned words and check in which corpora they appear.
b. Which corpora have *MP3*? Do the dates of the texts included in the corpus give you any explanation of why this might be so?
c. Why do you think *with* appears in all six corpora? Why do you think some corpora have more hits for *with* than others?
d. Which corpora have the word *epicure*? Think of some more rare words and check in which corpora they appear. Does corpus size affect the chances of finding rare words in them?
e. Which two corpora do not have *yeah* in them? Why do you think *yeah* does not appear in these corpora?
f. In which corpora is *amiable* reasonably frequent? Can you think of an explanation for this?
g. Which corpora have the word *pelagic*? Why is a technical word like this unlikely to be found in the other corpora?
h. Which corpora have the word *lass*? In which corpora are regionally marked words like this least likely to be found?
i. Which corpus does not have *darling*, and which corpus has only one occurrence of this word? Why do you think this is so?
j. Which corpora have the word *rosary*? In which corpora is this word reasonably frequent? Why?
k. *Coalition* appears in all six corpora. In which corpora is it comparatively very infrequent? Why?
l. The foreign word *rapporteur* does not appear in some English language corpora, but in one of them it is exceptionally frequent? Why?

Table 3 *Assessing a specific corpus*

Read the description of the Business Letter Corpus in Table 1. Based on this information, decide which of the words and expressions below are likely to be very frequent in the corpus, and which ones are unlikely to be found in it. When you finish, use the corpus to test your predictions.

Cheerio	I am pleased to	I love you	looking forward to	soup
Thank you for	very funny	We regret	footprint	Yours sincerely

a. Which of the above words and expressions is the most frequent in the corpus?
b. Which four search terms cannot be found in the corpus?
c. Were all your predictions right? If not, which one(s) surprised you? Why?

ignore diacritics and capital letters altogether. For example, in his tutorial on the use of the CWO corpus, Thomas (2002) prompts novice users to try out a search for *BBC* (in capital letters) and then explains that it will only work if we counterintuitively search for *bbc* (in lower case). In contrast, in the OPUS interface to EuroParl and in the COMPARA corpus, both of which use the CPQ query language (Evert 2010), users will find hits for *BBC* only if they use capital letters.

Diacritics can also be important. Users will get no results if they look up *café* in CWO, but if they type in *cafe* (without an accent) they will get concordances for both *café* and *cafe*. In contrast, in COMPARA and the OPUS interface to EuroParl, *cafe* and *café* are processed separately.

When users are trying out a corpus for the first time, they should be encouraged to experiment with capital letters and accents in order to understand how the software treats diacritics and upper and lower case distinctions. They should come to realize that different corpus tools tend to deal with these non-trivial details differently.

Another point that must be made explicit is that corpora are sensitive to word inflections. People who are used to employing dictionary strategies should learn that looking up uninflected forms in corpora will only retrieve uninflected forms. A very simple, task-based exercise to make new corpus users aware of this is to ask them to carry out queries for uninflected forms (e.g. *look*) and different inflections (e.g. *looks, looked, looking*), and then discuss the results obtained. Having done this, they can then be encouraged to read the information about the query language associated with a specific corpus in order to find out if there is a shortcut to capturing the different inflections of a word in a single query.

4.2.2 Strings of words

Once novice users are familiar with the basics of single-word queries, they can begin to get acquainted with queries involving strings of words. Perhaps the first point that apprentice users need to bear in mind is that, unlike search engines, corpora require people to retain words such as *the*, *of* and *it*, which when entered into search engines are treated as 'stopwords' and ignored, to speed up searches. In the OPUS interface to EuroParl, for example, there are eighty-five hits for *European Year of Languages*. However, if we leave out the word *of* and enter

Table 4 *Expansion and reduction exercise*

Look up the following strings of words in the BNC and write down their frequencies. What can you conclude from your results?
It
It was
It was fine
It was fine as
It was fine as far
It was fine as far as I
It was fine as far as I could
It was fine as far as I could see
was fine as far as I could see
fine as far as I could see
as far as I could see
far as I could see
as I could see
I could see
could see
see

the expression *European Year Languages* instead, we get absolutely no hits at all. Novice users who are encouraged to carry out these two queries will be able to conclude for themselves that, when using corpora, words such as articles and prepositions should not be ignored.

Once apprentice users realize that corpus searches need not be limited to single-word queries, they also need to understand that while it is usually possible to retrieve frequent combinations of words from corpora, strings of words put together in ways that have not been used before or that have only been used in very limited contexts are very unlikely to be found. In Tables 4 and 5 I propose two different task-based exercises to help novice users understand this.

The exercise in Table 4 shows learners that there are no instances of the string *It was fine as far as I could see* in the whole of the BNC. From that, novice users can be led to conclude that, even though it is a perfectly well-formed English sentence, it is a unique utterance which is not represented in any of the texts that make up the BNC. In contrast, they will also be able to see that English speakers often use clusters of words like *it was fine* and *as far as I could see* in that precise combination and order. They should also be able to notice that as the word strings get shorter, the frequencies get higher. The idea is for them to deduce that the shorter the sequence of words, the more likely it is that other people have used it before.

When novice corpus users do the exercise in Table 5, they should realize that although all clusters are made up of exactly three words each, there is a dramatic difference between them with regard to the frequency with which they are represented. The tri-grams *as a rule*, *a rule of*, *rule of thumb, you need a*, *a litre of* and *square metres of* are clusters of words often used by native speakers, while sequences like *of thumb you*, *of paint to*, *every 12 square*, and so on are not. This should help people who are new to corpora understand that it is not just the length

Table 5 *The trigram exercise*

Below are sequential three-word clusters taken from the sentence *As a rule of thumb you need a litre of paint to every 12 square metres of wall*. Which clusters are likely to turn up in the BNC? Which ones are unlikely to be found? Can you guess which will be the most frequent? Test your predictions and then discuss your results.

As a rule	thumb you need	litre of paint	every 12 square
a rule of	you need a	of paint to	12 square metres
rule of thumb	need a litre	paint to every	square metres of
of thumb you	a litre of	to every 12	metres of wall

of a string of words that matters: it is also the actual words chosen, together with the exact order in which they occur. Apprentice users should realize that when using corpora to look up strings of more than one word, we can only realistically expect to find sequences that occur in clusters that represent conventional uses of the language appearing in the corpus.

4.2.3 Reformulating queries

Another important point to introduce to novice corpus users is that queries often have to be reformulated. They should find it helpful to be explicitly reminded that they might have to broaden word queries so as to capture words with alternative spellings such as *generalise* and *generalize*. More commonly, however, single-word queries may have to be narrowed down. For example, newcomers to corpora should realize that homographs – such as *look (V)* and *look (N)* – need to be filtered if they want to capture just one of the meanings of the word. Using COCA or the BYU-BNC, both of which have user-friendly, drop-down part-of-speech (POS) annotation options, it is quite simple to ask new users to run queries with and without POS annotation in a series of short, hands-on exercises to draw attention to these intricacies.

Multiple-word queries, in contrast, are often too restrictive. Apprentice users can be shown that a multiple-word query that produces no results at first can be made more flexible by replacing certain elements of the string with wildcards or, if the corpus is grammatically tagged, with POS categories. For example, they can be asked to look up a search string like *give Barbara a hand* in COCA or the BNC, which will return no results. They can then be asked to replace *Barbara* with a wildcard and run the query again, which will give them plenty of analogous hits which sanction the use of the string *give Barbara a hand*. As a follow-up, they can then be shown how to replace *Barbara* with a noun or pronoun.

Many of the intricacies of single and multiple-word queries are difficult to anticipate, and even experienced corpus users may fail to foresee them. Novice users must learn that corpus querying is often a recursive process, and that the unsatisfactory results we get from initial queries can teach us how to improve our subsequent attempts to extract the information we are after.

4.3 Interpreting corpus output

Perhaps the most difficult aspect of learning to use corpora, according to Sinclair (2004), is interpreting our results. As already mentioned, unlike teaching materials mediated by experts – such as published grammars, textbooks and dictionaries – corpus users must get used to deciphering corpus output autonomously. People who are not yet very familiar with corpora can easily misinterpret their results.

A very simple consciousness-raising exercise novice users can try out to begin with involves looking at frequently misspelled words and comparing their frequency with that of their correct equivalents. For example, newcomers to corpora can be asked to look up the following pairs in the BNC: ∗*beleive:believe*, ∗*defendent:defendant*, ∗*payed:paid* and ∗*accomodation:accommodation*. Their results will contain hits for both the correct and the incorrect forms. However, as the figures below indicate, the correct forms will be far more frequent:

∗ *beleive (9)*	*believe (20,431)*
∗ *defendent (8)*	*defendant (3,300)*
∗ *payed (45)*	*paid (1,542)*
∗ *accomodation (46)*	*accommodation (4,361)*

These straightforward findings can be used to draw attention to the fact that (a) the texts that make up corpora are authentic texts that have not been linguistically edited, so they probably contain mistakes, errors and non-standard uses of language and (b) forms that are considered to be correct and standard are nevertheless likely to be much more frequent. Considerations such as these can help novice users to become aware that corpora are different from the carefully edited published materials language teachers are accustomed to using. If they find one or two hits for a misspelled word or a common grammar mistake in a general language corpus, they must not jump to the conclusion that this is sufficient proof of conventional usage.

Another consciousness-raising exercise that can help novice users become aware of the intricacies of evaluating corpus data involves introducing them to the concept of relative frequencies. For example, using the BYU interface, which provides user-friendly, drop-down options for selecting specific sub-corpora of the BNC, apprentice users can be asked to look up a word like *honestly* in the sub-corpus of speech and then in the sub-corpus of fiction. They should encounter 439 hits in the former and 459 hits in the latter, which could lead them to mistakenly conclude that the word *honestly* is a little more frequent in fiction than in speech. Next, they can be asked to pay attention to the sizes of the spoken and the fiction components of the BNC, and they will see that the fiction sub-corpus has over 50% more words. They should then be prompted to work out that the fact that *honestly* appears 439 times in 10.33 million words in the sub-corpus of speech, and 459 in times in 16.19 million words in the sub-corpus of fiction means that, according to the BNC, the word is proportionally more frequent in speech than in written fiction.

Apprentice users may also need to be explicitly taught not to make a summary analysis of the data. A simple exercise to train them to take a closer look at co-texts entails asking them to look up a word like *congratulations* followed by a preposition in the CWO corpus, which will render mainly concordances with *congratulations on*, *congratulations to* and *congratulations from*. After identifying these prepositions, the important question is whether these three prepositions

can be used interchangeably. To help novice users to become aware of the importance of paying close attention to co-texts, they can be asked to take a more careful look at what comes after each preposition and discuss whether they can see any patterns emerging. They should be able to conclude that although *congratulations on*, *congratulations to* and *congratulations from* are all acceptable, they are not synonymous.

Another very straightforward exercise to help novice users learn about the basics of interpreting corpus output involves drawing their attention to the significance of context and medium. This can be easily exemplified with a search for a word like *whatsit* in CWO using the entire corpus first, and then each of the three available sub-corpora separately. They will see that there are 27 hits for *whatsit* in the entire 56 million-word corpus, but 22 of them are from 10 million words of spoken British English; only 5 from 36 million words of British books, ephemera, radio, newspapers and magazines, and none of them are from the 10 million words of American English. With this, they should be led to conclude that *whatsit* is an expression that is more typical of spoken British English, and that it is probably not very common in America.

5. Concluding remarks

Several of the studies I have referred to here indicate that there is a long way to go before corpora can be understood and used by language teachers in general. A closer look at novice-user behaviour suggests that a major hurdle on the way to popularizing corpora is that so far little attention has been paid to training newcomers in basic corpus skills. Experts who use corpora routinely usually take these skills for granted, but language teachers who are not yet very familiar with corpora may need help in order to understand how to choose between different types of corpora, how to retrieve information from a corpus and how to evaluate that information.

I have presented a series of task-based, consciousness-raising exercises aimed at helping language teachers understand the basics of corpora. Exercises such as those proposed may be too easy for experts, but the ideas underlying them are not necessarily self-evident for people who have never used corpora before.

I have only had time to describe a limited number of exercises addressing just a handful of issues that are important to bear in mind when using corpora. Many more straightforward, hands-on, consciousness-raising exercises can be created. Of course, if you have ever tried doing so, you will know that it is actually quite hard to produce pedagogic materials that will make complex corpus concepts and procedures look simple to newcomers.

Yet exercises such as these can be very important if we are to make the direct use of corpora in the classroom more popular. After all, corpora do not work in the same way as the familiar language-learning resources – such as dictionaries, grammar books and textbooks – that language teachers are accustomed to using. Helping teachers who are new to corpora to take their first steps in using corpora autonomously will hopefully encourage them to want to find out more about using corpora in the classroom. And there is no lack of follow-up material available: as I mentioned at the beginning of this paper, there is a growing body of literature that strives to make corpus resources in general accessible to non-experts, including

general introductions to corpora, corpus-specific tutorials, and books and articles on how to use corpora in language teaching.

Appendix

Six free online corpora used in the task-based exercises exemplified

1. British National Corpus (BNC)
   General British English from the early 1990s; 100 million words; available from various sources; recommended interfaces for apprentice users: Simple Search: www.natcorp.ox.ac.uk. BYU interface – http://corpus.byu.edu/bnc.
2. Business Letter Corpus (BLC)
   American and British business letters from the 1990s; 1 million words; available at www.someya-net.com/concordancer.
3. Collins Wordbanks Online English (CWO)
   Also known as the COBUILD corpus; a demo version of the Bank of English containing general British and American English from the 1990s; 56 million words; available at www.collins.co.uk/corpus/CorpusSearch.aspx.
4. COMPARA
   Parallel corpus with original and translated English and Portuguese fiction; 3 million words (1.5 million English); available at www.linguateca.pt/COMPARA.
5. Corpus of Contemporary American English (COCA)
   General American English from 1990–2010 (regularly incremented with new texts); 450 million words in mid-2010; available at http://corpus.byu.edu/coca.
6. EuroParl
   Parallel corpus with European Parliament debates from 1998 to 2003 in 11 European languages; the English component contains 26 million words of British, Irish, International and translated English; available online via the OPUS interface at http://urd.let.rug.nl/tiedeman/OPUS/bin/opuscqp.pl?corpus=Europarl3;lang=en.

Five corpus-specific tutorials

COCA: www.americancorpus.org/help/learners_e.asp
COMPARA: www.linguateca.pt/COMPARA/docum/Tutorial.pdf
Collins Wordbanks Online: http://web.quick.cz/jaedth/Introduction%20to%20CCS.htm
MICASE: http://micase.elicorpora.info/using-micase-tips-tutorials
SACODEYL: http://www.um.es/sacodeyl/data/publications/SACODEYL_guidelines_for_teachers.pdf

Two general tutorials on corpora and language teaching

CALPER: http://calper.la.psu.edu/corpus_portal/tutorial_overview.php
ICT4LT: www.ict4lt.org/en/en_mod2–4.htm

References

Aijmer, K. (ed.) (2009). *Corpora and language teaching*. Amsterdam and Philadelphia: John Benjamins.

Anderson, W. & J. Corbett (2009). *Exploring English with online corpora*. Basingstoke: Palgrave Macmillan.

Aston, G. (ed.) (2001). *Learning with corpora*. Houston: Athelstan.

Aston G., S. Bernardini & D. Stewart (eds.) (2004). *Corpora and language learners*. Amsterdam and Philadelphia: John Benjamins.

Bennet, G. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Michigan: University of Michigan Press.

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (eds.) *Rethinking language pedagogy from a corpus perspective*. Frankfurt am Main: Peter Lang, 225–234.

Bianchi, F. & E. Manca (2006). Discovering language through corpora: needed abilities and student difficulties in corpus analysis. Paper presented at the *7th Conference on Teaching and Language Corpora*, Paris, 1–4 July 2006. www.openstarts.units.it/dspace/bitstream/10077/3193/1/04Bianchi_Manca.pdf

Boulton, A. (2009). Data-driven learning: reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics* 35.1, 81–106.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL* 17.1, 47–64.

Breyer, Y. (2009). Learning and teaching with corpora: reflections by student teachers. *Computer Assisted Language Learning* 22.2, 153–172.

Burnard, L. & T. McEnery (eds.) (2000). *Rethinking language pedagogy from a corpus perspective*. Frankfurt am Main: Peter Lang.

Davies, M. (2004). *BYU-BNC: The British National Corpus*. http://corpus.byu.edu/bnc.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 410+ million words, 1990–present*. www.americancorpus.org.

Evert, S. (2010). *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial*. http://cogsci.uni-osnabrueck.de/~korpora/ws/CWBdoc/CQP_Tutorial.

Frankenberg-Garcia, A. (2005). A peek into what today's language learners as researchers actually do. *International Journal of Lexicography* 18.3, 335–355.

Frankenberg-Garcia, A. (2010). Encouraging EFL teachers to use corpora in the classroom. Invited presentation at the BAAL and Cambridge University Press Seminar *Using corpus evidence in the classroom: Working with teachers and learners*. University of Birmingham, 24–25 June 2010.

Frankenberg-Garcia, A. & D. Santos (2003). Introducing COMPARA: the Portuguese–English parallel corpus. In F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in translator education*. Manchester: St. Jerome, 71–87.

Frankenberg-Garcia, A., L. Flowerdew & G. Aston (eds.) (2011). *New trends in corpora and language learning*. London: Continuum.

Gilquin, G. & S. Granger (2010). How can DDL be used in language teaching? In A. O'Keeffe & M. McCarthy (eds.) *The Routledge handbook of corpus linguistics*. London: Routledge, 359–370.

Hidalgo, E., L. Quereda & J. Santana (eds.) (2007). *Corpora in the foreign language classroom*. Amsterdam and New York: Rodopi.

Johns, T. & P. King (eds.) (1991). *Classroom concordancing*. Birmingham: The University of Birmingham Centre for English Language Studies.

Kaszubski, P. (2011). IFAConc – a pedagogic tool for online concordancing with EFL/EAP learners. In A. Frankenberg-Garcia et al. (eds.), 81–104.

Kennedy, C. & T. Miceli (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology* 5.3, 77–90.

Kettemann, B. & G. Marko (eds.) (2002). *Teaching and learning by doing corpus analysis*. Amsterdam and New York: Rodopi.

Kübler, N. (ed.) (forthcoming). *Corpora, language, teaching, and resources: From theory to practice*. Bern: Peter Lang.

Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In U. Connor & T. Upton (eds.) *Applied corpus linguistics: A multidimensional perspective*. Amsterdam and New York: Rodopi, 239–250.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: the state of the art and beyond. In S. Braun, K. Khon & J. Mukherjee (eds.) *Corpus technology and language pedagogy: New resources, new tools, new methods*. Frankfurt am Main: Peter Lang, 5–24.
O'Keeffe, A., M. McCarthy & R. Carter (2007). *From corpus to classroom*. Cambridge: Cambridge University Press.
Reppen, R. (2010). *Using corpora in the language classroom*. New York: Cambridge University Press.
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* 54.2, 121–134.
Römer, U. (2009). Corpus research and practice: What help do teachers need and what can we offer? In K. Aijmer (ed.) (2009), 83–98.
Santos, D. & A. Frankenberg-Garcia (2007). The corpus, its users and their needs: A user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics* 12.3, 335–374.
Sinclair, J. (ed.) (2004). *How to use corpora in language teaching*. Amsterdam and Philadelphia: John Benjamins.
Someya, Y. (2000). Online business letter corpus: KWIC concordancer and an experiment in data-driven learning/writing. Paper presented at the *3rd Association for Business Communication International Conference*, Kyoto, 9 August 2000.
Thomas, J. (2002). *A ten-step introduction to concordancing through the Collins Cobuild Corpus Concordance Sampler.* http://web.quick.cz/jaedth/Introduction%20to%20CCS.htm
Tiedemann, J. (2009). News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova & R. Mitkov (eds.) *Recent advances in natural language processing*. Amsterdam & Philadelphia: John Benjamins, Volume V, 237–248.
Tribble, C. (2001). Corpora and teaching: adjusting the gaze. Paper presented at the *22nd ICAME* conference, Louvain, 16–20 May 2001.
Tribble, C. & G. Jones (1997). *Concordancing in the classroom: a resource guide for teachers*. Houston: Athelstan.
Widmann, J., K. Kohn & R. Ziai (2011). The SACODEYL search tool – exploiting corpora for language learning purposes. In Frankenberg-Garcia et al (eds.), 167–178.

ANA FRANKENBERG-GARCIA is Auxiliary Professor at Instituto Superior de Línguas e Administração and invited Auxiliary Professor at Universidade Nova de Lisboa. She was joint project leader of COMPARA, a three-million word parallel corpus of English and Portuguese fiction, available online at www.linguateca.pt/COMPARA. Her work on using corpora in language and translation teaching and research has been published in various books and journals, including *International Journal of Lexicography*, *International Journal of Corpus Linguistics*, *Corpora* and *ELT Journal*. Most recently, she has co-edited *New trends in language learning and corpora* with Lynne Flowerdew and Guy Aston.