Center for Advanced Language Proficiency Education and Research (CALPER)

Using Corpora in Language Teaching

Michael McCarthy & Jane Evison

CALPER Digest 2004

There can be no doubt that corpora have revolutionized the way that we study and teach language (Biber et al., 1998, Hunston, 2002; McCarthy, 1998; Scott and Thompson, 2001; Sinclair, 2004, 1991). A corpus-based approach to language study involves storing, accessing and analyzing vast amounts of language data. Underlying this approach is the belief that the best way to understand and teach language is to draw inspiration from observing it being used to communicate in as many different ways and places as possible. Those who support a corpus-based approach argue that if we rely purely on introspection and anecdote to inform our teaching, we are likely to have an inaccurate view of how writers and speakers express themselves, not least because we tend to notice unusual things rather than 'typical' ones. This change in approach has been made possible by the technological advances which now allow vast amounts of language data from both spoken and written sources to be stored, accessed and analyzed electronically. However, despite the intrinsic part that technology plays in this corpus-based study of language, it is important to remember that those who use this approach do so to complement other more traditional ones, not replace them. The best corpus-based language studies, the most useful corpus-informed materials and the most accessible corpus-directed lessons are successful because they incorporate insights made possible by the use of computers with the existing knowledge base and shared experience of language professionals.

What are corpora?

Since the 1960s, when electronic corpora were first compiled, a growing number of teachers and researchers have come to believe that analyzing these databases of authentic language provides an alternative to intuition. But what is a corpus? A **corpus** (plural **'corpora'**) is simply a collection of texts (corpus is the Latin for 'body' and so a corpus is just a body of texts). A corpus is a large and principled collection of naturally occurring texts. The size of a corpus can range from tens of millions of words to a few thousand. These texts can be either transcripts of spoken language (increasingly with sound or visual files attached) or written language that has been scanned from books, newspapers etc. or downloaded electronically. The data that is stored in a computer can be in its 'raw' form i.e. just the words of the source texts. More often it is **marked up** to allow structural features (e.g. distribution of speaker turns; titles, authors and subheadings), and contextual features (where and when conversations occurred, or texts were written) to be included. It can also be **tagged** so that the grammatical categories of the individual words can be included. Some corpora are even **parsed** (i.e. the grammatical structure of sentences is automatically labelled in terms of subjects, verbs, objects, etc.). The size and make up of a corpus will depend on its purpose, but it is generally true that spoken corpora tend to be smaller than written ones because the data is more difficult to collect and much more expensive because of transcription costs (Biber et al.,1998; McCarthy, 1998).

What types of corpora are used?

Very large corpora are used to inform processes like the writing of dictionaries and grammars. It is unsurprising that these general corpora have to contain tens of millions of words because lexicographers and grammarians need as many examples as possible of all the words, expressions and structures in a particular language. On the other hand, a medium-sized corpus of a few million words, consisting of transcripts of lectures and seminars, could be used to design materials for learners who need academic language for their studies. Parallel corpora, which contain texts in more than one language, are compiled to assist translation and contrastive studies. Some corpora are owned by institutions, but many others can be bought on CD-ROM. There are also a growing number of websites which make corpora available online to anyone free of charge. Finally, there is one other particularly important type of corpus. This is the **learner corpus**, which will be discussed in the next section.

Learner corpora

The principled assembly and analysis of learner corpora can provide invaluable insight into second language acquisition (Granger et al., 2002). There are different types of learner corpora. Some contain millions of words from thousands of different learners. One source of language for corpora such as these is that which is collected from oral or written exams taken anywhere in the world. These corpora are of particular interest to writers of textbooks and compilers of exams, as well as to researchers into second language acquisition. Often such corpora are tagged for errors, a time consuming process, but one which adds another dimension to analysis. Other learner corpora are smaller and contain language samples from the same group of learners over a period of time. The advantage of these corpora is that it is possible to get a picture of how well learners are progressing, by analyzing, for example, how their vocabulary range is growing. The use of learner corpora is an exciting area of research, and more and more teachers are finding that, with the right guidance, it is possible to construct and analyze such corpora themselves.

What can corpora tell us?

Corpora themselves can tell us nothing; they are merely databases of texts in electronic form. In order to interpret the data that corpora contain, increasingly sophisticated computer programs have been developed. At their simplest level, these programs count and retrieve patterns of language use. There are two key advantages of analyzing language in this way. Firstly, the use of computers means that it is possible to analyze a far larger quantity of language than if this analysis were done 'by hand'. Secondly, such analysis is both reliable and consistent. However, it is vital to go beyond this 'quantitative' analysis, and use our experience of how we languages work and how they are learnt in conjunction with the mark up and tagging information mentioned above to suggest **why** these patterns

Developed with funding from the U.S. Dept. of Education (CFDA 84.229, P229A020010).

should exist. This 'qualitative' analysis can be particularly insightful. On exciting thing for both language teachers and language learners is that getting insights from corpora is not restricted to university-based researchers. There are a growing number of software programs available (either to buy, or as downloadable freeware), which allow corpora to be exploited easily on home computers. Two of the fundamental computer-managed processes are the generation of **frequency lists** (either in rank order, or sorted alphabetically), and **concordances** (examples of particular words or phrases in context).

Frequency lists

One of the first areas to be informed by the analysis of computer-generated frequency lists has been that of lexicography. For example, dictionary writers can now place more frequent meanings of a word or phrase first in dictionary entries, making it more likely that a user will find the explanation they need quickly and easily. However, the usefulness of frequency lists is not restricted to the compilation of dictionaries. Analysis of the frequency with which words or expressions occur in a particular corpus can also be used to inform teaching and learning. One key area that they help inform is that of syllabus and materials design. Many language courses aim to teach beginners to communicate orally in a target language as quickly as possible. One of the key issues facing those designing such courses is: what items **are** most frequently used in day-to-day spoken interaction? Analysis of frequency lists can help answer this question. Evidence from a corpus of spoken English (McCarthy, 1998) suggests that there is a core vocabulary of about 1,800 'heavy duty' words, which represent a realistic pedagogical target for beginners of a language. There is a sharp fall-off in frequency after the first 1,800, which shows us that these 1,800 words work harder than all the others. But as with any raw data which results from computer analysis of corpora, these figures need to be interpreted qualitatively if their significance is to be fully understood. As we said before, computer programs do not make judgements about why particular results occur. For example, the word 'know' occurs in fourteenth position in this frequency list for spoken English. However, this does not mean that speakers are constantly telling others 'I know this, that or the other'. If we investigate the words which tend to occur with 'know', we discover that a great many of the occurrences of 'know' are in fact part of the phrase 'you know'. Other items which occur near the top of our frequency list for spoken language are not 'words' at all in the traditional sense. For instance, 'mm', which expresses meanings such as acknowledgement, topic pausing, agreement and hesitation in spoken English is the 14th most frequent item, and 'oh', which commonly expresses a reaction of surprise or a desire to shift the conversation to something new, is the 32nd. This corpus evidence suggests that items like 'you know', 'mm' and 'oh' are too important to be ignored (Hunston, 2002; McCarthy, 1997).

Concordances

The fact that an item like 'know' tends to co-occur with 'you' can be explored by examining concordances (or concordance lines). A concordance (sometimes known as KWIC, or Key Word in Context) is a screen display or printout of a chosen word or phrase in its different contexts, along with the text that comes before and after it. For most people, generating concordances is their first experience of corpus analysis, which can easily be done online. Corpora that have free online access generally allow a sample number of concordances to be generated quickly and simply. These concordances can be invaluable for both teachers and learners who want to see a range of examples of a word or phrase in context (Hunston, 2002; Sinclair, 1991). Apart from giving a good idea of the grammatical forms associated with a word, concordances allow recurring phrases containing the word to be identified. Concordances are particularly useful for understanding **collocation**, an area of language learning that many learners find difficult. Collocation is simply the likelihood that two words will occur together. For example, the word 'fair' is likely to be used with 'hair', but not with 'car' or 'jacket'. We therefore say that 'fair' **collocates** with 'hair', but that 'beige' does not. By studying concordances of 'hair' or 'beige', we can see a range of examples of how these words collocate with others. Concordances can generally be 'sorted' alphabetically, either centrally, or to the left or right of the target word, which allows different patterns to be uncovered. Concordances can therefore provide an invaluable resource for teachers, allowing them to access to far more example uses of a word or phrase than can be found in even the best dictionaries.

Data driven learning (DDL)

So far we have focused particularly on how teachers and researchers can benefit from using corpora to inform their professional practice. However, the relative ease with which corpora can be accessed and the increasing availability of hardware and software, both in learning establishments and at home, means that learners are increasingly able to explore corpora for themselves. This kind of approach is frequently referred to as data driven or discovery learning, and can be extremely stimulating for more advanced level learners (Hunston, 2002; Johns, 1991; Sinclair, 2004, 1991).

Conclusion

In the past, many members of the teaching profession, and the research community too, were hostile towards the use of corpora to inform teaching and learning. They argued that nothing could substitute for the knowledge and intuition of experienced professionals. However, it is becoming more and more widely accepted that the principled collection and analysis of real language data has much to offer language teaching and learning. The combination of informed experience and the power of the computer is a formidable tool for all language educators.

References

Biber, D., Conran, S., & R. Reppen (1998) *Corpus linguistics: investigating language structure and use.* Cambridge: Cambridge University Press.

Granger, S., Hung. J., & S. Petch-Tyson (eds.) (2002) *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam: John Benjamins.

Hunston, S. (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal*, *4*, 1-16.

McCarthy, M. (1998) *Spoken language and applied linguistics.* Cambridge: Cambridge University Press.

Visit CALPER's Corpus Portal for more information: http://calper.la.psu.edu/corpus_portal.index.php

Please cite as: McCarthy, M., & Evison, J. (2004) Using Corpora in Language Teaching. (Digest 1104). University Park, PA: The Pennsylvania State University, CALPER

Additional copies can be downloaded at http://calper.la.psu.edu/publication.php?page=dig1

Developed with funding from the U.S. Dept. of Education (CFDA 84.229, P229A020010).