Corpus linguistics and the ESL classroom

Jan Hendrik Beck, University of Freiburg

Although corpus linguistic research has steadily grown in popularity since the appearance of the first computerized text corpora in the 1960s, many ELT professionals are still unaware of the immense potential corpus linguistics has for the improvement of language instruction.

The aim of this article is twofold: first, it provides a basic introduction to the theory of corpus linguistics and shows why corpus linguistics is relevant and interesting not only for language theorists but also for practitioners. To illustrate how corpus linguistics can directly contribute to classroom teaching, the article introduces the reader to *data*-

driven-learning (better known as classroom concordancing), which empowers language learners to become their own corpus researchers. Secondly, the article

provides some resources for instructors who might consider using corpus linguistics in their own class-rooms. To that end, the last section of the paper points out some introductory readings about corpus linguistics and classroom concordancing, and refers the reader to software, corpora and tutorials. As many of our colleagues work in teaching contexts where money is tight, all corpora and software referred to in this paper are available on the Internet free of charge.

What is corpus linguistics? What is a corpus?

Corpus linguistics deals with the analysis of large collections of natural language data, the *corpora*, for underlying patterns of language use. Put quite simply, a corpus linguist looks at large amounts of written text, including text transcribed from oral

discourse, either to discover new patterns in how language is used or to test and exemplify theories.¹ To give a few examples, a researcher might be interested in how a word changes in meaning in different linguistic contexts, how frequently it appears in different sorts of texts (informal vs. formal, written vs. spoken), whether it is associated with particular syntactic structures, and so forth. Counting and the quantitative presentation of the findings are integral parts of most corpus linguistic studies. Most instructors have strong intuitions about language, but because the corpora consist of actual language uttered or written by language users, corpus linguistics is always strictly empirical.

classroom concordancing ...
empowers language learners to
become their own corpus researchers

Because of the objective nature of corpus linguistics, a corpus should represent a language or a variety of a language as accurately as

possible. Therefore, the designer has to make choices in the selection of the texts. It is important to note that corpora are not conglomerates of random texts, but that any well-designed corpus follows certain principles. A comprehensive corpus such as the British National Corpus (BNC), for instance, which attempts to represent the British English dialect as a whole, must not only contain face-to-face conversations but also phone conversations, newspaper articles, TV speech, novels, etc. For a more detailed discussion on the principles of corpus design, see Biber, Conrad and Reppen, 1998.

A second important characteristic of corpora is their size. Since corpus studies are always empirical, in order to have valid results, the researcher must make sure that the corpus is large enough to be representative of the language or the variety it reflects. According to Sinclair (2001), the first

Volume 25, 2007

electronically-stored text corpora were assembled in the United States and Europe concurrently, but the project teams were unaware of each others' projects until 1965. Brown University designed a one-million-word corpus called the Brown Corpus of Standard American English (more commonly referred to simply as the Brown Corpus), which contained written language from various registers. Today, COBUILD (Collins Birmingham University International Language Database), one of the largest English corpora, contains over 600 million words.

By today's standards, a corpus of one million words is fairly small, but it is evident that without the help of computers any kind of analysis is incredibly time-consuming, if not impossible. For this reason, researchers use computer programs that can do much of the quantitative work such as the counting of frequencies and the search for particular linguistic items or collocation patterns.

Concordancing

It is a common misconception that corpus linguistics is only for those who have strong computer skills and who are able to write their own soft-

ware. Although some linguists need to develop their own software for very advanced corpus analyses, highquality software capable of performing a variety of investigative tasks is available and is more than sufficient for a multitude of interesting research projects.

The standard program for analyzing corpora is called a *concordancer* and operates somewhat like a search engine on the Internet: the researcher loads a corpus into the concordancer and can then "tell" the concordancer to search for a word or phrase. Laurence Anthony is the author of AntConc, an excellent free concordancer (available for Windows, Mac, and Linux). Most corpora offer a wide variety of analysis functions. The following are available in AntConc.

KWIC (Key Word in Context) is the most basic function of all concordancers and at the same time their most versatile feature. The researcher can look up a search term and be presented with all hits (instances of the search term) in their linguistic contexts. These so-called concordance lines give the researcher valuable information about linguistic association patterns that go beyond quantitative descriptions provided by the software.

For example, questions that can be answered by looking at key words in concordance lines include: Is the search term associated with a particular syntactic structure? Is it associated with a particular speech style (such as vulgar language)? Concordance lines are also an extremely powerful tool to help language learners interpret meanings or changes in meaning of polysemic words according to their contexts without being presented with direct word-for-word transla-

tions. Table 1 below shows some concordance lines of the word *issue* from a corpus of native speaker undergraduate essay writing.

Even from only these six concordance lines, one can distinguish three quite distinc-

tive meanings of the word. In the first two lines, issue has the meaning of matter, whereas in the second two lines, it is more likely *problem*. The last two lines refer to a volume in a sequence. While native speakers generally have few problems inferring the appropriate meaning intuitively, this can be a challenging task for learners of English. By looking at keywords in context, however, the learners more likely will be able to "guess" the correct meaning by considering the immediate linguistic context (for example line 3: "this issue will be difficult to resolve..." or line 6: in the April 4, 1994 issue of the Wall Street Journal..."). Also by being presented with so many examples of the word, the learners can gain a systematic and broader knowledge of a word's multiple meanings and uses.

without the help of computers any kind of analysis is incredibly time-consuming, if not impossible

```
... their own claims. However, with an ... a prison warden. Finally, there is the ... care workers are underpaid, and the ... sensitive, horrifying, and controversial ... community. In the December 7, 1994 issue of the New York Times, Anna... issue of the Wall Street Journal states ...
```

Figure 1. Sample Key Word in Context concordance

KWIC can also be effectively used to study grammar patterns. For instance, a learner might be interested in American English usage patterns of the future tenses. Textbooks generally tell their readers that the *going to*-future is restricted to intentions, assumptions, and other non-definite statements about the future. However, performing a concordance search, the learner can see that the distribution of the tenses primarily depends on the register: *going to* is much more common in many spoken registers, and there, it is not restricted to the textbook rules of usage but quite universal.

Evidently, there is no clear-cut line between lexis and grammar (Hunston & Francis 2000); for instance, words are often associated with particular grammatical structures. As Biber, et al. (1998) show, *little* and *small*, often thought to be synonymous, have very distinct preferences for grammatical

structures. In both conversation and academic prose, small is found much more frequently in predicative position than little. In concordance analysis, it is therefore always worthwhile to consider association patterns between lexis and grammar.

Most commercially available ESL/EFL materials do a poor job of reflecting language as it is actually spoken and written

Another basic function of every concordancer is the *frequency count*, which tells the researcher how often the search word or phrase occurs in the corpus. Sometimes, this feature alone can yield interesting insights into a language. For instance, the researcher might be interested in the distribution of a particular word over different registers or dialects, in which case he or she would search for the term in question in different corpora (for instance, one containing American English and one containing British English). Another example of the use of frequency counts is to find out more about synonymous or near-synonymous words such as *sofa* vs. *couch*.

Most concordancers can list common *collo-cates* and *word clusters* of the selected search term in ranking orders. These practical tools are especially interesting for investigating formulaic expressions, which play an important role in native speaker discourse (for a corpus linguistic study on formulaic expressions in second language acquisition, see De Cock, et al., 1998), but also about such learner-

relevant aspects like phrasal verbs and distribution of prepositions.

AntConc is capable of producing word lists from the texts in a corpus, presenting the most frequent words in the corpus in ranked order. This is a great way to identify the most common words in a language or a language variety, such as academic prose. Examples of word lists produced from frequency lists are the General Service List (GSL) (West, 1953), presenting the 2,000 most common word families in English, and the Academic Word List (Coxhead, 1998), containing the 570 most common word families in academic prose (excluding the ones already contained in the GSL). Both are based on corpus research and are important tools for ESL/EFL learners.

On a smaller scale, word lists can also be created for single texts before having students read the texts in the English classroom. Doing so will help

the teacher determine which words are central in the text and should therefore be given special attention. For this purpose, the teacher creates a mini-corpus by simply loading the one text (as a regular .txt file) into the concordancer.

Authenticity: Relevance of corpus-based research

It should be evident that corpus linguistics can contribute to valid linguistic research. Processing vast amounts of authentic language data allows researchers to base their work on reliable empirical evidence, rather than on abstract theories or native speaker intuition. For example, every traditional grammar textbook tells the reader that English requires subjectverb agreement. However, a simple investigation of any given corpus of spoken American English presents many instances where this is not the case, as in *There is* + *plural* constructions ("There's four cats on the patio."). In addition, corpus linguistics encompasses crucial aspects of descriptive research. It describes linguistic phenomena quantitatively and qualitatively and formulates theories based on data that are authentic, rather than stating prescriptive rules of what language "should" be like.

The notion of authenticity has important implications for language teaching. Most commercially-

Volume 25, 2007

available ESL/EFL materials do a poor job of reflecting language as it is actually spoken and written by native speakers, focusing instead on what is considered correct grammar or vocabulary usage. On the same note, what is considered correct language is also almost always written language, which presents the standard for assessing any learner's language. McCarthy and Carter (McCarthy & Carter, 1995; Carter, et al., 1998) address this correct-incorrect dichotomy and argue for a stronger awareness of spoken grammar.

From a textbook writer's point of view, a clear-cut definition of what is acceptable English makes sense, and in fact there are good arguments for using a language of instruction that is grammatically correct and consistent with written English. Carter, et al. (1998) concludes that "on the one hand, we have real English which, as far as classroom treatment is concerned, can be unrealistic; and on the other hand, we have unreal textbook English which, as far as classroom treatment is concerned, is

frequently handled in peda-

gogically viable and realistic

ways" (p. 50).

students can perform their own research projects and share their findings with the class

Carter also acknowledges the often-quoted argument that learners do not have to sound like native speakers and need not become experts at navigating the vast cultural contexts authentic language brings about, as most non-native speakers learn English to communicate with other non-native speakers (Cook, 1998; McKay, 2003). Yet, from my own experience learning and teaching English, I know that virtually any serious learner of English desires to sound as much like a native speaker as possible and also wishes to connect the target language to a target culture. This is especially true for ESL students, who can profit directly from sounding more natural, as this makes their language use less marked for native speakers and therefore raises the learners' acceptance in the target culture.

Regardless of one's position of which variety of English to teach, it is important to acknowledge that there are numerous varieties of English one could learn and teach. Depending on the corpus used, corpus linguistics can answer questions about many varieties of English and raise awareness of differences in dialects and styles.

Data-driven learning (DDL)

If the class has access to computers and the learners are at least at an intermediate level of English proficiency, the teacher can have the students do their own corpus research. Tim Johns (1991, 1997), the originator of *Data Driven Learning* (DDL) believes that the language learner is at the same time a language researcher, and that to learn the language effectively, the learner needs to have access to authentic linguistic data (Cheng, Warren, & Xu-feng, 2003).

In small groups or on their own, students can perform their own research projects and share their findings with the class. In most cases, learners of English will be interested in meanings and real usage of words, phrases, and constructions. Using a concordancer and a corpus, they are presented with a

large number of items in various linguistic contexts. To stay with the earlier example, a student might be confronted with an instance the word *issue*, which has a number of very different senses as a verb and as a noun. The student could, of course, simply look

up the word in a dictionary, but still might not be able to find the correct meaning, as dictionaries do not usually provide much context. Furthermore, the student is likely to have to look up the word again the next time it is encountered because little cognitive effort is involved using dictionaries, especially bilingual ones.

An alternative is to run a concordance search for issue (or in this case, a search for the term issu* with the wildcard * that allows all forms of the word as a noun and as a verb to be retrieved), presenting the learners with a huge number of examples of how native speakers use the word. Teachers generally agree that presenting learners with ready-made translations and explanations is not a very effective teaching method. DDL, however, engages the learners in an inductive learning process, in which they have to make their own inferences and generalizations about meaning and use until they find what they need. This requires some more time than using a dictionary, but the students will gain a deeper understanding of the meaning and use of the items in question and improve their general problem-solving

abilities (such as hypothesis forming and testing, an integral part of language acquisition). Another benefit of this approach is that student motivation is generally high, as the students feel that they are doing real research and actually accomplishing something.²

Besides such useful ad hoc studies, which can be done by individual students in a short time, teachers should consider larger concordancing projects for small groups. Depending on the needs and the proficiency level of the learners, the scope of these projects can range from basic grammatical or lexical patterns to very complex cross-linguistic or cross-dialectal comparisons. To give some examples of the various possible research topics, Table 1 presents a study by Cheng, Warren, and Xu-feng (2003), who introduced concordancing to English majors at a Hong Kong university.

After the students in the study finished their projects, they were asked for their comments. It speaks well for the method that 81% of the students in the class found it "interesting" or "very interesting" and that 87% found it "useful" or "very useful."

Concordancers in the classroom

Classroom concordancing is most useful in areas of ELT that are not sufficiently covered in course curricula and standard teaching materials. As

classroom concordancing projects take some time, both for the introduction to the method and for the actual projects, the teacher has to consider carefully what topics are worth including into a concordancing session to make it maximally effective. In ESL contexts, students often might have questions about words and constructions they hear or read outside the classroom, but for which they cannot find an answer in their textbooks. The teacher should also take notes of problematic areas (certain words, phrases, or structures) in the learner language whenever they arise and include chronically troublesome features.

A feature of English that seems to be problematic for many learners is the appropriate use of the words *can*, *to be able*, and *to be capable*. Investigating meaning and use of the three items makes a great concordancing project, as all three words describe roughly the same concept. However, there is plenty to find out about the exact meanings and morpho-syntactic structures associated with the items.

The first step of a concordancing project should always be activating the students' existing knowledge. In the case of *can*, *to be able*, and *to be capable*, it is sufficient to write the items on the board and ask the students for the meanings. This provides them with a working hypothesis such as "All three words mean the same." Anything works at this point; it is important to keep in mind that the

Table 1. Students' research areas	
Research areas	Examples
Single lexical items	Nouns and pronouns (god, I, you, that, they); adjectives (little, great); modal verb (shall); adverbs (yet, anyway, actually)
Pairs of words	because and since, between and among, after and afterwards, could and must, able and capable, lovely and sweet, almost and nearly, in and into, however and but, enough and sufficient, guy and man, for ever and forever, anyway and but anyway, at the beginning and in the beginning, at the end and in the end, and each other and one another
Taboo words	damn, shit, hell, fuck, fucking
Discourse markers	you know, excuse me
Comparing spoken and written discourses	but, however, shall, I, you, that god, at the beginning and in the beginning, guy and man
Comparing female and male speakers	great, you know
Comparing native & non-native speakers of English	they, because
From Cheng, Warren, & Xu-feng, 2003	

Volume 25, 2007 5

learners are the researchers, so the teacher should be careful not to steer the students in a particular direction.

The first task for the learners is to run a concordance search in a first corpus to determine the frequencies of the items.³ For this example, I am using a 4.2 million-word corpus of U.S. conversational English and get the following result:

can: 3,347 / 1 million words – *able*: 283 / 1 mil. – *capable*: 7 / 1 mil.

In a second task, the learners run the same search in a different corpus (in my example, academic writing) to compare the frequencies across the registers:

> can: 3,091 / 1 million words - able: 309 / 1 mil. - capable: 86 / 1 mil.

It appears that *can* is by far the most common of the three items in both corpora, but it seems to decrease in frequency in favor of *able* and *capable* in academic writing. The data would indicate that *capable* seems to be much more typical for academic writing than for conversations.

In a next step, the students look at the key words in context to investigate individual meanings. In most cases, it is sufficient to have the students look at the first 20 hits of each word. One reason for *can* being so common immediately becomes clear: it has a much broader range of senses than its two counterparts. While *able* and *capable* seem always to refer to actual ability, in only the first 20 of my concordance lines, *can*, besides ability, expresses possibility, permission, request, and suggestion. The teacher should encourage the students to define the meanings of the items as accurately as possible and come up with their own examples to share with their classmates.

Finally, an investigation of collocation patterns and grammatical context is worthwhile. Can serves as a modal auxiliary, following the noun and preceding the main verb. Able almost always occurs in the form to be able to [V], although in a few cases, it serves as an adjective attribute ("He's an able person") or as a predicative attribute ("He's so able"). Capable, being the most restricted item, almost always occurs in the construction to be

capable of, except in the rare cases where it serves as a adjective attribute modifying a noun ("He's a highly capable guy") or as a predicative attribute ("He's so capable"). Able and capable have very similar grammatical associations, except that able to calls for an infinitive verb, whereas capable of calls for a gerund form. Having the students sort their concordance lines according to the words neighboring the search terms (most concordancers can do this) makes such facts jump to the learners' attention.

These steps are just some of many basic concordance operations that learners can use to discover language. Other tasks can entail further frequency and context comparisons across registers or dialects (for example, British vs. American English), looking at grammatical and semantic/functional context in more detail, comparing concordance findings to textbook rules, and investigating collocation patterns.

Conclusion

This paper has presented a brief introduction to the concepts of corpus linguistics and concordancing and has discussed some aspects of how the methodology can be relevant to language practitioners and learners. It was shown that the use of authentic language in the ESL classroom can contribute to learning success, and how classroom concordancing can empower learners to be language researchers. The following section describes some free resources that provide a start for those interested in pursuing corpus linguistics and concordancing.

Resources

Websites

Devoted to Corpora: http://devoted.to/corpora
David Lee's website has it all: links and descriptions of corpora, software, tutorials, conferences, and other corpus-related topics make this website the ultimate link collection for everyone interested in corpus linguistics. The hyperlinks are well-maintained; only the descriptions of some references seem a little subjective occasionally.

The Gateway to Corpus Linguistics on the Internet: http://www.corpus-linguistics.de/
The Gateway offers a variety of links to corpora and other corpus-related topics and resources on the web. This clearly-arranged website is in

excellent shape and probably somewhat easier to navigate than Devoted to Corpora.

Tim Johns Data-driven Learning Page: http://www.eisu.bham.ac.uk/johnstf/timconc.htm Tim Johns is the originator of DDL, and his website provides a useful bibliography of literature on the topic, as well as some links to materials and software.

Data Driven Vocabulary Learning:
http://darkwing.uoregon.edu/~billwalk/ddvl.htm
This website presents the script of Bill Walker's
(University of Oregon) presentation on the use of
concordancing for vocabulary learning at the
TESOL 2003 Convention and contains some useful
and clear examples of DDL exercises.

AntConc 3.2 Help
http://www.antlab.sci.waseda.ac.jp/software/
AntConc_Help/AntConc_Help.htm
The AntConc website offers an excellent tutorial
on the functions in AntConc.

Concordancers

AntConc: http://www.antlab.sci.waseda.ac.jp/
software.html
This powerful and user-friendly program by
Laurence Anthony is capable of performing all of
the concordancing tasks that were discussed in the
concordancing section. The website also has a
useful tutorial on using the concordancer.

TextSTAT 2.7: http://www.niederlandistik.fu-berlin.de/textstat/software-en.html
This excellent and simple concordance tool by
Matthias Hüning (Freie Universität Berlin, Germany) allows the construction and analysis of simple corpora by incorporating text either from data stored on the computer or the Internet.
Simply adding web addresses, the user can feed the software with whole websites and search them with the concordancing tools included.

Corpora

Unfortunately, most authors do not give their corpora away for free, which is not surprising considering how much time and effort go into a well-designed corpus. However, there are a few high-quality free corpora available. A new corpus of written English is being assembled at Portland State University. The *Viking Corpus of Academic Student Writing* will contain papers and dissertations by

university students and will be in the public domain once it is released.

International Corpus of English: http://www.ucl.ac.uk/english-usage/ice/index.htm

The ICE Project gathers corpora in fifteen different research centers in the English speaking world, each corpus containing one million words of spoken and written text. The corpora are free for academic non-commercial use and can be downloaded or ordered on CD-ROM. Unfortunately, the U.S. corpus is not released yet, but there are a

MICASE: http://www.lsa.umich.edu/eli/micase/index.htm

number of other corpora available already.

The University of Michigan brings us MICASE (Michigan Corpus of Spoken Academic English), an excellent on-line corpus, which comes with its own concordancer and can be accessed for free. The corpus' scope is limited to spoken English in the university context but offers a huge variety of social, demographic, and situational search filters. This is an exciting page to explore for learners and teachers alike.

The Compleat Lexical Tutor: http://www.lextutor.ca/
This website by Tom Cobb (University of Quebec at Montreal) focuses on vocabulary and offers a rich selection of self-learning tools for learners and resources for teachers. Its online concordancer has access to the Brown Corpus, the spoken and written BNC, and a variety of specialized corpora. Most features of the website are available in French and English.

References

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

Carter, R., Hughes, R. & McCarthy, M. (1998).
Telling tails: Grammar, the spoken language and materials development. In B. Tomlinson (Ed.),
Materials development in language teaching.
Cambridge: Cambridge University Press.

Cheng, W., Warren, M., & Xun-feng, X. (2003). The language learner as language researcher: Putting corpus linguistics on the timetable. *System*, *31*, 173-186.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

Volume 25, 2007 7

- Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal*, *52* (1), 57-63.
- De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer*. London: Longman.
- Hunston, S. & Francis, G. (2000). *Pattern grammar:* A corpus-driven approach to the lexical grammar of English. Amsterdam: John Benjamins.
- Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom concordancing (English Language Research Journal 4)*. ELR, Birmingham, 1-16.
- Johns, T. (1997) Contexts: The background,development and trailing of a concordance-basedCALL program. In A. Wichmann, S. Fligelstone,T. McEnery, and D. Knowles (Eds.), *Teaching*and language corpora. London: Longman.
- McCarthy, M. & Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal*, 49 (3), 207-218.
- McKay, S. (2003). EIL curriculum development. *RELC Journal*, *34*, 31-47.
- Sinclair, J. (2001). Preface. In M. Ghadessy, A. Henry, & R.L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins.

West, M. (1953). A general service list of English words. London: Longman, Green.

Notes

- 1 Some researchers (e.g., Cheng et al. 2003) distinguish between *corpus-driven research*, which develops theories based on corpus research, and *corpus-based research*, which uses corpora to check theories that are not necessarily based on corpus research.
- 2 However, not everyone is comfortable working with computers, and some people even consider themselves "computer-blind" or feel anxious about working with them. Teachers should be aware of this and make sure affected students get extra assistance while working on their tasks.
- 3 For a tutorial of how to use the different functions of a concordancer, please refer to the resource section of this paper.

Jan Hendrik Beck came to the U.S. as an exchange student in 2005. He fell in love with Oregon, stayed, and received his M.A. in TESOL from Portland State University in the summer of 2007. Jan is currently pursuing graduate studies with the University of Freiburg, Germany, and is trying to get as much teaching experience in Oregon as possible.

Calling all authors!

Want to write for *The ORTESOL Journal*? Here are some suggestions:

- 1. Look at the Guidelines for Contributors on the inside back cover.
- 2. Decide if your topic will work best as a feature article, research note, or teaching note.
- 3. Remember that feature articles should be readable, draw on research, and have practical class-room application. Feature articles will include references to the literature as appropriate.
- 4. Watch word count, especially in Research and Teaching Notes.
- 5. Include images in attached files in resolution that is as high as possible (300-600dpi).
- 6. Include your bio data in a separate file.

Feel free to contact the Editors with ideas and questions. The peer review process generally takes two to three months to complete. We know Oregonians have a lot to share, so let us hear from you.