Telling ELT Tales out of School

Instructional uses of linguistic technologies

Thomas Cobb\textsuperscript{a}, *

\textsuperscript{a} Université du Québec à Montréal, Canada

Abstract

The earliest and most basic linguistic technologies are frequency lists and the concordance lines from a corpus. Although these technologies have been available for many years we are only now discovering their uses in language instruction. My talk will present the main instructional uses of lists and lines that are emerging, and show ways that language instructors are not only employing these linguistic technologies but also contributing to their further development through incorporation of the cognitive and affective dimensions of language learning. All ideas and theories will be presented in the context of real-life applications (mainly via the Compleat Lexical Tutor website (at www.lextutor.ca), and participants will leave with ideas and tools for immediate use in course design and execution.

Keywords: language and technology; computational linguistics; computer assisted language learning (CALL); data-driven learning; frequency; concordance; constructivism.

* E-mail address: cob.tom@uqam.ca

1877-0428 © 2010 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.sbspro.2010.07.007
1. Introduction

Computational linguists have developed some nice tools for looking at language, tools that have forced us to look at language in new ways and sometime very new ways. My presentation will argue and attempt to show that even intermediate language learners can benefit form using these tools when their teachers or they themselves have adapted them to learning purposes and not just purposes of language description or analysis. I will build my argument from two very basic linguistic technologies, concordancers and frequency lists.

The theme of this conference is “telling tales,” with the notion of a “tale” normally being a lengthy story with lots of twists that goes on for a considerable length of time. My tale is about how much adaptation is needed to join learner and technology in the case of some concordance learning tools that I have been working on for more than a decade.

2. Concordancers and language learning

Many of the discoveries about what language is and how it works in the last 50 years have been produced by analysing language samples as concordance lines. Concordance lines are little strings of language drawn out of a much larger text and arranged to show any similarities between then lines or other patterning that may exist. Figure 1 shows what one English concordance output looks like, in this case for all of the available meanings of any form of the word family *man* when in the presence of any form of *woman*. From this relatively minor output we can make some possibly unintuitive observations, which a native speaker may not have been fully aware of after a lifetime of using English, namely that in the vast majority of at least these instances *man* comes before *woman*. A native speaker may have vaguely suspected that this was the case, but the evidence in Figure 1 appears to provide some numbers: 16 out of 19 concordance lines place *man* first, and when *woman* does come first there is a sentence boundary between the *man* and *woman*, leaving only one true, same sentence, *woman*-first instance in the sample (1/19, or about 5%).

Of course this information is only interesting if we know what the larger text was that this information comes from. If this was a book catering to male interests like football or hunting, then we probably would not have learned anything about the language as a whole. In fact, it is from the Brown corpus (1967/1979), consisting of 1 million words of authentic American English, drawn from 2000 texts on a wide range of subjects, and thus the observation about man and woman may have some chance of being applicable to American users of English as a whole. But of course we might wish to see if this observation held to the same extent for more recent varieties of English, or for users of English other than Americans, and for this it would be necessary to test the same question in a more modern corpus, non-American corpus, such as the British National Corpus (Oxford Computing Services, 2006) – which incidentally at 100 million words could potentially strengthen our case considerably.
Figure 1: *Man* and *woman* as seen in a few concordance lines from the Brown corpus

Behind the development of a modern corpus is normally thousands of hours of work and a good deal of research finding. The Brown (Kucera & Francis, 1979) was seen as enormous in its time, with 500 authentic texts of 2000 words apiece on a systematic range of topics in three major areas (academic, press, and fiction). But then the British National Corpus of 100 million words in 100 major divisions dwarfed the Brown, and is being dwarfed in its turn by the Corpus of Contemporary American English (the COCA, Davis & Gardener, 2010). These and other corpora have produced “scientific” information about language, such as which language forms really get used most and least, and what native speakers really say or write in particular situations. Basically, concordance analysis of corpora has brought to an end the era of guesswork in language learning – which of course is particularly important when a language is being taught by non-native speakers, as is increasingly the case with English as a Second Language (ESL) worldwide.

Much of this information as interpreted by language specialists has made its way into second language (L2) learning materials in three main forms, - dictionaries (Scholfield, 1996), courses (McCarthy et al, 2005), and tests (Biber et al, 2004). My interest in this presentation, however, is to talk about language learners using concordance information themselves.

Why should L2 learners learn to use concordance information themselves? One reason is that no one can learn everything there is to know about a language from being told; at some point, the learning depends on observing how the language works and drawing one’s own conclusions from observations – this is called data-driven learning. If the habit of paying attention to how language is used can be inculcated in a learner, then the chances of a successful outcome to language learning are very good, and the conditions of paying attention are never likely to be better than when one has a corpus and a concordancer to work with.

There are many ways of asking advanced learners to work with concordance data, usually from a worksheet such as the one shown in Figure 2. In this task the learner is asked to use a concordancer to find the most frequent collocations (associated words, as *woman* was associated or collocated with *man* in Fig. 1) of the two meanings of the two homonymous words sharing the single word-form *bank* (money bank and river bank). Following this they are asked to draw a conclusion about these collocates, as to what degree they are overlapping. It is expected the collocates will be words like *finance, robbery, manager,* and *teller* on one side and *grassy, trees, boats* and *fishing* on the other, with little or no overlap. Learners are then asked to draw a conclusion from their evidence, namely that different words sharing a single word form can nonetheless be distinguished by their typical collocations.
For the advanced learner task shown in Fig. 2, a teacher can pretty much work straight from a corpus, and the only adaptation required is the worksheet itself. But for less advanced learners the adaptation process can be much more extensive. And “less advanced” learners are after all the majority of language learners who are likely to find themselves in a classroom. So if concordancing is worth doing, there should be ways to do it with less advanced learners too. There are indeed ways to do this, with equally long “tales” of adaptation in either case, to be recounted in the following – for both patterns of meaning and patterns of form.

2. Intermediate concordancing for patterns of meaning

For this discussion let us define an intermediate learner one who has meaning recognition knowledge for between 1,000 and 2,000 word families (as identified by Nation and Beglar’s Vocabulary Size Test, 2007, or at www.lextutor.ca/tests). This learner would be unaware, compared to the learner envisaged just above, that a word like bank has two independent meanings, and thus the exercise in Fig. 2 has no meaning for him or her. Further, the notion of a collocation is probably not of crucial importance to a learner who is building a basic word bank. Still further, if we assume such learners have some thousands of words to learn, then some sort of worksheet for every one of these is somewhat impractical. And yet in principle the concordance, as ultimate if implicit repository of all the information about how a language works, has value to offer this learner too, at his or her needs level.

What are the needs of such a learner. Arguably (Nation, 2001), one primary need is to build a basic word stock, both for its own sake and to leverage other kinds of learning like grammar and pragmatics. However, building a word stock can be done in a number of ways leading to rather different results. A learner can obtain a wordlist of “the basic words of English” and go through this writing in definitions from a dictionary and then take these away and memorize them; this is some effort but hardly an insurmountable task if the learner is mildly motivated to make some progress with the L2. A problem with this method is that research has consistently shown that it does not impact any of the uses to which this word knowledge would normally be put, such as reading with comprehension in the L2 or writing with some degree of lexical sophistication. What research has shown will help learners meet both of these lexical objectives is meeting new words in rich contexts. The problem of course is that meeting words in rich contexts is a rather haphazard not to say lengthy process. And yet a concordancer can present words in several rich contexts, although not in a way the learner is likely to get much from. What is the learner to “do” with the output in Fig. 1? What is needed is some way to turn the concordancer into a contextual learning tool for large number of words.
In a study of computer assisted vocabulary learning in the mid-1990s at Sultan Qaboos University in Oman, I developed several adaptations of the concordancing idea as a vocabulary learning tool for lower intermediate learners, and one of these tools in particular was quite popular with students and took on a life of its own. The learning tool (shown in Fig. 4) was adapted in two ways to appeal and be useful to such learners. First, a dictionary and a dictionary task was built into the tool (for a series of words, read the concordance lines until you can pick the best definition for all the lines), and second the corpus providing the concordance lines was guaranteed to be readable to these learners (since it consisted entirely of texts from the reading courses they had already taken).

The purpose of this experiment was to determine whether learners (a) would use, and (b) could benefit from such an approach to vocabulary expansion. The main test was whether they could apply meaning-knowledge thus gained to words in novel contexts. Users of this tool were tested weekly for ten weeks on a sampling of 200 words they had learned in that week (from the second thousand frequency level). The test was in two parts, first to choose a short meaning for 10 of their new words, and second to apply 10 other new words to gaps in a novel text which they had never seen before. A control group, who had not used the concordance program, also took these same tests. The result, described in further detail in Cobb (1997), was that both groups were equal at choosing definitions but concordance users were significantly better at applying their new words to novel contexts.

The benefits thus shown, I resolved to expand this system to a larger, less learner-specific tool, for which I reasoned I would need a corpus of at least 1 million words at both native-speaker level (the Brown would serve nicely) for learners working at the fourth 1,000 level and up, but more problematically 1 million words of simplified English for learners working at the second and third 1,000 levels. These latter are of course the vast majority in any formal learning context, and their reasons for needing a simplified corpus are obvious – they have to understand the examples before they can pick a definition that encompasses all of them (as the participants in the original experiment had done by dint of meeting their own materials in the concordance lines). Further, I would need several thousand definitions written in simple English.

Obtaining such a simplified corpus and lexicon was a project of roughly a decade. The definitions were slowly put together mainly owing to a kindly offer of access to the Wiki simple dictionary database (xxx). The corpus was put together by slowly collecting as many graded or simplified readers as I could find in machine-readable format (accessible from several corpus interfaces in Lextutor). While assembling these materials, I did not make this software available publicly since it was devoted to a specific context, but during this time the system was, interestingly, cited in a major work about applying corpora to language learning (O’Keefe, McCarthy, & Carter, 200x, p. 25) as a prime example of a data-driven learning tool from the 1990s. Figure 5 shows a sample of the simplified graded-readers based corpus, and Figure 6 shows a section of the dictionary thus being developed (in the figure, VP-1 refers to the Vocabprofile of the items being the first thousand - hence, 1 - and the URL is given for any reader wishing to follow the progress of this project).
To summarize then, from the basis of a sound empirical finding, a good learner response, and recognition in the research literature, it is a tale of roughly 13 years duration to expand an in-principle or “toy” learning tool up to a full, web-based learning system. The materials are now in place for this system to be launched, sometime in 2010, but already a prototype is up and running on Lextutor (at lextutor.ca/conc_infer, see Fig. 7).
3. Intermediate concordancing for patterns of form

A similar tale of lengthy adaptation can be told for making the concordancer a useful way of telling learners about patterns of both morphological and syntactic form. As every language teacher can attest, learners seem to find distinctions of form imperceptible during the early stages of learning. One reason for this is that they do not have the lexis to create a form-meaning mapping, and without meaning the perception of form is probably an impossibly abstract feat (except for computers, of course). Thus language instruction needs to emphasize lexis before it emphasizes form, although of course this is not always done.

Another reason learners cannot perceive formal distinctions is that these are often presented as one-off rules rather than as patterns. To have a pattern it is axiomatic that a form must be presented in several manifestations, and yet as we know many grammar books present one example of a form and then turn learners loose on the exercise, which as already mentioned may be in a lexis that is unfamiliar to them.

Yet another reason learners may not perceive a formal distinction is that it is imposed on them at a time they are not ready for it. For example, the conditional verb may be the topic of today’s lesson, but the learner has not so far attempted to express a conditional meaning and indeed may not be in the habit of doing so in his or her first language (L1).

To emphasize some element of language form to a language learner, then, we can postulate the following conditions for successful learning: the form should be couched in meaningful lexis that the learner understands; the form should be presented in several manifestations; and the form should be presented when the learner has shown he or she is ready to use it or has a need to use it. These conditions are to some extent met when learners’ attention is directed to form in the context of an error they have made, that is, when they have attempted to use the form but not entirely successfully, and they are asked to correct the error with the help of a concordancer. The concordancer will give them several examples of the correct form, if they are able to find it. However, they may not be able to find the correct form, since they do not already know it, and then there is the problem mentioned above that they may not find the lexis of the concordance lines comprehensible and no form-meaning mapping can occur, however many examples the concordancer produces.

One solution to this problem, of course, is to use the graded reader corpus already mentioned as the source of information about language form. Provided the learners have about 1500 word families in their lexicons, form-meaning mappings should be largely perceptible in this corpus. The other problem, that learners may not be able to generate the correct form from their error, is a bit more difficult. Ideally, a teacher would wish to generate the
concordance for the learner and put it in his or her text right beside the error. The learner’s task then would be to perceive the difference between his form and the concordance lines, not to generate it.

A long tale of corpus development has made something like this idea possible. In the 1980s, concordances were located on individual PCs (e.g., Johns, 1986), and the storage of a particular concordance meant copying the information onto a separate document. In principle, teachers could generate a few hundred pages of concordance output for every essay they assigned, but this was not practical. Then in the late 1990s and beyond, both concordances and most learner writing submissions went online. In this case, a whole complex concordance can be encoded as an URL and simply inserted in the learner’s text under a rubric like “Error 1” or something similar.

Such a plan for providing ESL learners with concordance writing feedback was developed and tested by Gaskell and Cobb (2004). Figure 8 shows a learner text bearing a standard error (“go *to home”) and the feedback link “Revision 1” which the learner will click to see what the problem could be with his or her sentence.

Figure 7: An error with its feedback link

Figure 9 shows the URL this link leads to on Lextutor, and Figure 10 shows the concordance lines encoded in this URL. These lines clearly provide the learner with some comprehensible pattern information to work with.

Figure 8: What’s behind the feedback link

Figure 9: What the feedback link shows the learner: no “to” with go home

The lines shown in Figure 10 clearly provide the learner with some comprehensible pattern information to work with, not only to correct the error but probably to learn something about this form that will be transferable to novel instances of the same form. This is a piece of important research that is waiting to be done. Although Cobb and Gaskell showed that learners were motivated to use this learning tool, the final word is not in on how useful it is. Following the outcome of Cobb (1997), there should be a case for optimism.

This would be particularly true in the case of subsequent developments since this software was developed and the study undertaken in 2004. A careful examination of the tightly coded URL in Fig. 9 will reveal that the corpus used at the time was the Brown corpus, which is a native speaker corpus. Since then the graded readers corpus has been completed, as mentioned above, and the potential increased for the concordancer to produce the form-meaning mappings on which grammars are built.

The teacher who wished to experiment with this approach to teaching form can inspect the pre-encoded forms that have been codified in the tutorial program Concordance Grammar on Lextutor (www.lextutor.ca/concord_grammar), a part of which features in Figure 11.
Concordance grammar on Lextutor provides numerous exercises on several grammatical topics (singular-plural, number, and many others). In Figure 11, the task is to use the concordance lines to rewrite “fed up of” (a common error for many learners) as “fed up with.” Teachers can use this page to see if their learners are ready for this approach, and if so then go and build their own links for insertion to their learners’ own writing submissions. They can do this at the main concordance input page (www.lextutor.ca/concordancers/concord_e.html, at the bottom).

4. Conclusion

In both the cases presented here, a tool developed by linguists for purpose of language description and analysis has been adapted to a language learning purpose – in one case to the perception of meaning patterns and in another to the perception of form patterns. In both cases a period of more than 10 years was required to achieve and test the adaptation, and probably more than that will be needed to make these generally useful and available. These of course are only two of many adaptations under way somewhere in the wide world of applied linguistics. Other researcher-developers are working on adapting corpus techniques to many other areas, particularly the exploitation of learner corpora (e.g., Granger & Pacquot, 2010, and other work by the Louvain group), and many others. In all cases, the adaptation process is long and slow but – so far – highly fruitful.

If this process of adapting corpus research is so slow and difficult, then someone approaching applied corpus work from the outside might legitimately ask, Why bother? The answer is simple, a corpus gives learners access to pattern information that they cannot get in any other way. In L1 acquisition, the patterns build up over countless exposures, following a “power law” of accumulation – that is, with each further exposure adding an additional minimal increment to the glacially accumulating competence (see N. Ellis, e.g., 2002). The power law clearly probably cannot function in L2 acquisition, certainly not initially, and so the patterns must be acquired in another manner. The tools of the linguist, in this case the concordance, provide such a manner to build up pattern information within the time frame available.

The kicker to this argument, however, is that in the end the linguists’ and the learners’ purposes are not all that different. Both require some way of making the repeated patterns of a language visible, as they are implicitly for native speakers, while both linguists and learners need to understand these patterns explicitly. The philosophy behind this approach was set out in theoretical terms in Cobb (1999) under the rubric “learner as linguist.” The argument presented there is only strengthened by the arrival of a comprehensible learner corpus and lexicon as described here.

At a time when computer and Internet are no longer novelties, and some of the potential of technology has been to some extent realized, or has failed to be realized, it is common to hear language practitioners say that they do not find their learners get much from using technology. Of course, these practitioners are largely referring to what their learners get from roaming over the Web and scan-reading various texts, or from doing one or two of the many simple multiple-choice quizzes that can be found on ESL or EFL Websites. The Web by itself will not teach learners
much, just as linguistic tools by themselves will not teach them much. In both cases, it is skilled adaptation of technologies invented for other purposes that will produce the powerful language learning tools.

References

[All Cobb references are available online at lextutor.ca/cv]


