Please cite as: Watson Todd, R. (2001) Building and using your own corpus and concordance. ThaiTESOL Bulletin vol. 14 no. 2.

Recently, there has been an increasing emphasis on corpora and concordancing in language teaching. The use of corpora and concordances has, however, been largely restricted to institutions with large resources, and the majority of teachers and students have been excluded from gaining the potential benefits. In this paper, I will show how specific genre and general corpora can be built and analysed without any special programs. This allows some innovative applications of corpora to become available in most classrooms, including getting students to create their own concordances for self-correction and creating databases to guide course design.

"The use of corpora and concordancing is now an area of considerable interest."

(Coniam, 1997: 199)

This quotation is the starting point of one of many recent articles highlighting the growth of the use of corpora and concordancing in language teaching. Originally designed as instruments aiding descriptive research into the use of language (Sinclair, 1991), corpora and concordancing have within the last five years jumped the barrier between research and teaching, and have become an important tool for many language teachers (Woolard, 2000).

There is, however, a danger that the value of corpora and concordancing will only be realised in institutions with large resources. Many of the authors of the articles on corpora and concordancing assume that readers have access to an existing large corpus of text and concordancing programs (e.g. Fox, 1998; Kettemann, 1995; Stevens, 1995; Tribble, 1997; Wichmann, 1995). Most institutions in Thailand, however, do not have these resources, resulting in Thai teachers and students running the risk of being left behind and missing out on the potential benefits of using corpora and concordances as aids for language learning.

In this paper, I hope to show that this need not be the case. It is possible to gain many of the benefits of corpora and concordances using standard readily available programs which many teachers and students have access to and are familiar with, namely, an Internet browser and a word-processing program.

Building Your Own Corpus

Before you start to build your own corpus, you need to decide what kind of corpus you want. There are two main kinds of corpus (Lewis, 2000). The first is a corpus of a specific genre of text, for example, academic articles, business letters, and newspaper feature articles. The second is a general corpus which includes texts from a wide variety of different genres.

Building a Specific Corpus

To build a specific corpus, texts within the chosen genre can be found on the Internet and downloaded. A specific corpus does not have to be very large. The only requirement for size is that a reasonable number of examples of each target word should be present in the corpus. If the target words are relatively high frequency words, a corpus of around 50,000 words should produce at least 10 examples of each word (and probably several hundred of *the*). If the words are relatively low frequency, then the corpus could be expanded to around 200,000 words. Beyond this point, the law of diminishing returns and the size of the file containing the corpus make further expansion not worth the effort.

One problem with making a corpus from the Internet is the small amount of text on most Internet pages. To overcome this, the search for suitable texts can be started from sites which link to pages giving full-text articles and the like. Four particularly useful starting points are:

- http://digital.library.upenn.edu/books/
- http://www.ipl.org/
- http://www.findarticles.com/
- http://www.lib.utexas.edu/Libs/PCL/Etext.html

Using sites like these, a specific corpus of around 100,000 words can be built in under an hour.

Building a General Corpus

There are far more problems in building a general corpus. If the corpus is to be representative of most written genres (building a corpus of spoken English should be avoided for practical reasons), then the corpus will have to be very large. Building your own general corpus is therefore a daunting task.

The alternative to building your own general corpus is to use the world's largest corpus - the Internet (for comparison, the COBUILD corpus which is frequently publicised on the basis of its size has 544 occurrences of the word *reticent* whereas there are 31,842 occurrences of *reticent* on the Internet. To use the Internet as a corpus, search engines with a wide coverage which search within pages as well as in meta-tags are needed. Two examples of such search engines are FAST Search (http://www.alltheweb.com) and Google (http://www.google.com). Using these search engines, a search will produce a list of pages in which the target item occurs. Entering the pages, using the 'Find' command, and copying the context of the target item enables the production of a concordance from the corpus of the Internet.

Making Your Own Concordance

Having built a corpus or decided to use the Internet as a corpus, the next stage is to make a concordance. The word *concordance* has two different meanings, both of which have applications in language teaching.

Word-count concordances

The original meaning of *concordance* is a word count of all the different words in a text. For example, for the invented sentence 'The boss was the same old boss.', a word-count concordance is:

boss	2
the	2
old	1
same	1
was	1

To produce a word-count concordance from a corpus, a concordancing program is helpful, although some other programs such as the database program *SortItOut* and the Web Frequency Indexer (http://www.georgetown.edu/cball/webtools/web_freqs.html) can also create this kind of concordance.

Examples-of-use concordances

The more recent meaning of *concordance* is a collection of examples of the use of a word, such as the concordance for *despite* below.

the influence of foreign capital. Villagers,	despite	being conceived as a "floating mass"
thus continued in the world spotlight,	despite	Jakarta's effort to remove it. Foreign
some new species have been introduced.	Despite	the lack of data, there is
tolerant to smoke from any source.	Despite	these concerns, people should recognize that
the central problem, successful implementation occurred	despite	disagreements about how best to solve
to continue participating as mentors and tutors,	despite	having been laid off from work

new major sugar schemes since 1978.	Despite	the great expansion of the area
and economic needs of the country.	Despite	this stated policy, the Department of
of which are experiencing neotectonic movements.	Despite	their relative low position the intermontane
need for flexibility in solving problems	despite	uncertainty) also suggest the importance of

A concordance like this can be created using ‘Find’ commands with a specific corpus or by searching the Internet to obtain examples of use. These examples can then be placed in a table using a word-processing program. Although more laborious than using a concordancing program, this procedure is far easier than constructing concordances by hand as suggested by Willis (1998). More importantly, it allows examples-of-use concordances to become available to many teachers and learners, enabling innovative uses of concordances in language teaching to be implemented.

Using Your Corpus and Concordance

Using a word-count concordance

A word-count concordance can only be used with a specific corpus (and even then, most concordancing programs limit the maximum number of words in the corpus to be concordanced). It is, therefore, most likely to be of use in the teaching of English for Specific Purposes. The main pedagogical use of a word-count concordance is in course and materials preparation. Knowing the frequency of occurrence of words in a specific corpus can indicate which words need to be taught in a course (this is especially important for procedural vocabulary, see Marco, 1999) and can help in finding representative texts to be used as materials in teaching.

Using an examples-of-use concordance

The most common use of concordances in language teaching is for the teacher to create a concordance of a word (such as the one for *despite* above) and ask students to induce rules of use from the concordance. For *despite*, induced rules of use may include the facts that *despite* is frequently used at the start of sentences and that *despite* is always followed by a noun phrase or gerund. This approach highlights the collocations and colligations (or grammar) of a word (Woolard, 2000), and also encourages students to realise the benefits of inducing their own rules from language data.

While the use of a teacher-chosen examples-of-use concordance is an effective teaching technique, it reinforces the erroneous viewpoint that the language points to be learnt are best chosen and presented by the teacher, and may even encourage some students to believe that the only language points learnable are those presented to them by the teacher. Proponents of this technique, on the other hand, argue that teacher control over the examples presented in a concordance is important since it gives the concordance a clear and learnable focus rather than being a hotchpotch of mixed meanings and uses of a word (e.g. Lewis, 2000).

Nevertheless, most of the language data that students come across, including all of the English they meet outside the classroom, has not been vetted by the teacher. If students are to learn from all this potentially valuable language, they need to be able to make valid inductions from collections of examples which have not been specifically selected as clear and unambiguous illustrations of a language point. Whether students can make valid inductions in such situations is a point needing investigation.

Recently, I conducted such an investigation (see Watson Todd, 2001) based on the premise that students could use the Internet as their corpus and construct their own concordances from it. The situation of the investigation was self-correction of written work. Students had written drafts of a report as an assignment for a course, and as part of the feedback two words which had been used incorrectly were indicated. Students were required to conduct an Internet search for examples of use of the indicated words, create a concordance from these examples, induce rules of use from the concordance, and apply the induced rules in self-correction of their writing. The findings showed that students' induced rules correctly described 78% of the examples in their concordances, and that students were able to self-correct their mistakes in writing in 78% of instances as well. An example of a student's self-selected concordance, the rules he generated and the correction of his written work is given in Appendix 1.

This investigation illustrates a valuable use of concordances built by the students themselves. Without relying on a teacher-generated corpus and concordance, and only using programs they were already familiar with, students were able to build their own concordance and use it to help their own language learning.

The methods of building corpora and concordances that I have suggested earlier in this paper and which are readily available to students as well as teachers and researchers facilitate innovative applications of concordancing in language learning. Students' use of a self-selected concordance in self-correction of written work is one possible application. Another is for the teacher to construct a set of awareness-raising questions concerning a word and to ask the students to build their own concordance to answer the questions. Two possible sets of awareness-raising questions are given in Appendix 2. This approach is situated somewhere between the classroom use of a teacher-selected concordance and the use of student-selected concordances in self-correction, and has the benefits of providing a clear focus for learning through the awareness-raising questions while also highlighting the students' ability to learn from any language data they meet.

Corpora and concordances, then, have a wide range of uses in language learning extending from the standard use of teacher-created concordances in the classroom through awareness-raising to the use of student-selected concordances in self-correction. The more learner-centred of these approaches can help students to realise that value of learning from any language input they come across, as well as serving the more usual purpose of learning important aspects of vocabulary.

Conclusion

In this paper, I hope I have shown how both teachers and students can build and use their own corpora and concordances. Although the suggestions in this paper involve a little more work than is required when using a dedicated concordancing program, experience with students at King Mongkut's University of Technology Thonburi has shown that the approach can be easily implemented. Using the approaches suggested in this paper allows teachers and students outside a few elite institutes to gain the benefits of using corpora and concordancing in learning and provides bountiful opportunities for students to discover the patterns of English.

References

Coniam, D. (1997) A practical introduction to corpora in a teacher training language awareness programme. *Language Awareness* vol. 6 no. 4 pp. 199-207.

Fox, G. (1998) Using corpus data in the classroom. In Tomlinson, B. (ed.) (1998) *Materials Development in Language Teaching*. Cambridge: Cambridge University Press. pp. 25-43.

Kettemann, B. (1995) On the use of concordancing in ELT. Available at: http://gewi.kfunigraz.ac.at/~ketteman/conco.html.

Lewis, M. (2000) Materials and resources for teaching collocation. In Lewis, M. (ed.) (2000) *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications. pp. 186-204.

Marco, M. J. L. (1999) Procedural vocabulary: Lexical signalling of conceptual relations in discourse. *Applied Linguistics* vol. 20 no. 1 pp. 1-21.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stevens, V. (1995) Concordancing with language learners: Why? When? What? *CAELL Journal* vol. 6 no. 2 pp. 2-10.

Tribble, C. (1997) Improvising corpora for ELT: Quick and dirty ways of developing corpora for language teaching. Available at: http://web.bham.ac.uk/johnstf/palc.htm.

Watson Todd, R. (2001) Induction from self-selected concordances and self-correction. *System* vol. 29 no. 1 pp. 91-102.

Wichmann, A. (1995) Using concordances for the teaching of modern languages in higher education. *Language Learning Journal* no. 11 pp. 61-63.

Willis, J. (1998) Concordances in the classroom without a computer: assembling and exploiting concordances of common words. In Tomlinson, B. (ed.) (1998) *Materials Development in Language Teaching*. Cambridge: Cambridge University Press. pp. 44-66.

Woolard, G. (2000) Collocation - encouraging learner independence. In Lewis, M. (ed.) (2000) *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications. pp. 28-46.

Appendix 1 A sample of a student’s work with a self-selected concordance

Educational Assistant is an educational tool	capable	of increasing a student’s attention, comprehension
Actions MooWP robots (and puppets) are	capable	of giving multi-line responses, and these
concept is one of a vehicle	capable	of traversing an antipersonnel minefield carrying
Vehicle) project involves building a robot	capable	of finding and extinguishing a fire
autonomous mobile robot navigation prototype system	capable	of performing office delivery tasks in
and built an RC servo “pup”	capable	of sitting, standing, walking and barking.
created a robot capable of … well,	capable	of navigating a maze.
A robot	capable	of juggling 3 balls was built
an autonomous mobile robot that is	capable	of competent, safe behavior.
Somehow, the Shadow is	capable	of generating quasi-real projections of itself.

Rules of capable:

Capable is used between verb to be and of.

Capable is always followed by verb ing.

Work to be corrected:

“It is *capable* taps all kinds of parts stamped and bar headed and die cast nuts, flange nuts, wing nuts 12 pt.”

Correction:

“It is *capable* of tapping all kinds parts stamped and bar headed and die cast nuts, flange nuts, wing nuts 12 pt.”

Appendix 2 Sample awareness-raising questions to be used with concordances

Example 1

Try to find 10 examples of *affect* and 10 examples of *effect*.

1. What part of speech is *affect*?
2. What part of speech is *effect*?
3. What are the grammatical patterns for using *affect*?
4. Are there any phrases which include *effect*?
5. What do you think *affect* and *effect* mean?
6. How can you prevent yourself from confusing *affect* and *effect*?

Example 2

Try to find 20 examples of *in case*.

1. Do any of your examples contain the pattern *just in case*?
2. Do any of your examples contain the pattern *in case of*?
3. What structure follows *in case* and *just in case*?
4. Are there any other uses of *just in case*?
5. What structure follows *in case of*?
6. What do you think *in case*, *just in case* and *in case of* mean? Are the meanings the same or different?