SUMMARY. Through small-scale sampling of relevant specialized texts to craft hands-on inferential vocabulary tasks, both students and teachers can benefit from corpus linguistic information. By discovering ways to collect and access real data, second-language teachers can create topic-specific corpora and use software to sort and highlight the data to create more rich and revealing classroom materials for improving vocabulary learning. This student-centered, data-driven learning can be easily adapted for different levels of reading students. Free or purchased software, as well as features of existing programs, can be put to use to access online texts. [Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <http://www.HaworthPress.com> © 2005 by The Haworth Press, Inc. All rights reserved.]

KEYWORDS. Vocabulary, second-language acquisition, online texts, corpus linguistics, reading, context, concordancing

LAUREL SMITH STVAN is Assistant Professor, Department of Linguistics and TESOL, The University of Texas at Arlington, Arlington, TX 76019-0559 (E-mail: stvan@uta.edu).
One area that has greatly benefited from the use of new technology is language learning. Better explanations of language use have come about as scholars have been able to access and examine larger and larger samples of language, a task made easier by the increased speed and storage abilities of computers. Many dictionaries, for example, have improved their entries over the years by citing from collections of naturally occurring data rather than made-up examples. While reference works based on such a body or “corpus” of attested data date back as far as the first Oxford English Dictionary publication in 1884, recent works that rely on collections of computerized texts for the data, such as the Collins COBUILD English Language Dictionary (Sinclair, 1987) and the Longman Grammar of Spoken and Written English (Biber, Johansson, Leech, Conrad, & Finegan, 1999), have now set the standard.

Beyond verifying actual usage, much work in linguistic analysis has been clarified by data from real (often genre-focused) data. But without being trained in software that will concordance a text—that is, sort and highlight instances of the item in question—and without purchasing copies of large corpus collections, how can language teachers find a way to access the insights derived from corpora to enhance the tasks that language learners use in the classroom? By doing small-scale sampling of relevant specialized texts to craft hands-on inferential vocabulary tasks, both students and teachers can benefit from corpus linguistic information. In discovering ways to collect and access real data, they can create topic-specific corpora and then sample their data to create more rich and revealing classroom materials.

Maddux, Johnson, and Willis (2001) present definitions of contrasting types of technology uses in which they separate those uses that merely speed up or put a clean face on tasks that are already going on in teaching, from those that allow a new exploration of ideas. These second uses are the Type II technologies, those that allow students and teachers to use inductive reasoning, data manipulation, and individual encounters with material in new ways. Similarly, Warschauer (1996, 1998) details levels of technology use in language classrooms, sorting out those that are purely instrumental from those more critical approaches that can “bring about new power relations in a classroom or community” (1998, p. 759). One type of language learning task that is both student centered and data driven is determining vocabulary usage based on multiple, real-life examples. Allowing students to infer meaning and usage patterns, from collections of real language use is a prime Type II use of technology. And today, concordancing software and im-
proved collections of text corpora make it easier to access relevant materials for each set of learners.

**COMPUTERIZED TEXT CORPORAS**

Early work in compiling a language corpus was aimed at gathering large samples that represented a whole language. The one million word Brown Corpus gathered by Francis and Kucera (1967), for example, was a general language corpus made of American English texts excerpted from documents, journalism, and fiction published in 1961. As each corpus compiled since then has gotten bigger, researchers’ ability to use it to more accurately describe the general language has improved as well. However, many teachers find that, in addition to general language vocabulary, their students need to become familiar with the language of a specialized subject or domain. For work, for travel, or for graduate studies, they need to read and master representative materials from a new genre. Thus, recent works like Bowker and Pearson (2002) and Ghadessy, Henry, and Roseberry (2001) focus on the insights to be gleaned from small, specialized text corpora.

Incorporating authentic language is not a new concept in communicative second-language teaching. Bringing in examples of real-life texts such as menus, cookbooks, appliance manuals, and job applications has been found to be a way to expose learners to the variety of text types that they may encounter in using a new language. The idea of collecting language samples of the type most useful to their students, therefore, is not out of line with many language teachers’ training.

In discussions of ways to gather language examples, the World Wide Web is often put forward as a source of naturally occurring data. In many ways it is the ultimate free and ever-growing corpus. But the Web is not without its downsides. It contains a mix of native and non-native speaker utterances, a mix of edited and unedited sources, a mix of spoken, written, and multi-genre material, and is a repository of language from many different eras. In other words, the material to be found on the Web is often too uncontrolled to rely on for usage samples. Thus, especially for beginning students, the Web can offer poor data from which to draw conclusions.

For first-language learners, Miller and Gildea (1987) discuss the difficulties of acquiring word meanings through reading traditional printed texts. In an argument pre-dating the Web, they suggest a need for a method with more contextual options available at precisely the spot
where the words in question appear, rather than forcing a reader’s eyes to leave his current sentence while he goes to look up the word in another book. Furthermore, when readers do move to a dictionary, they are forced to guess from multiple senses in a dictionary entry which one best matches their word’s original context, by which time that context may now be gone from short-term memory. Second-language readers face the same problems, of course, with the added difficulty of having had even less exposure to the new term to help them to puzzle out the correct sense.

Hyperlinking abilities across computer networks, as well as the ability to simultaneously access multiple applications, as is now found in most operating systems (OS), have eliminated some of the problems of a reader moving from book to book. That is, the modern reader does much of his or her reading and reference work in front of a single screen, with the ability to immediately click back and forth to seek further information.

**LANGUAGE TASKS THAT CAN EXPLOIT THESE TECHNOLOGIES**

Thus, OS software is rapidly improving to make it easier for readers to deal with new words. How can application software supplement this? Two kinds of linguistic software come into play: (a) collections of text corpora, and (b) concordancing and indexing tools to analyze them. Electronic corpus material is now being used in many ways to enhance language teaching, as can be seen, for example, in the special 2001 issue of *Language Learning and Technology* (volume 5, number 3) focusing on using corpora in language instruction. The improvement of grammars and dictionaries, mentioned earlier, has been one such outcome. Another focus is for second-language teachers to collect examples of writing produced by non-native speakers of a language in order to find error patterns in the use and acquisition of second-language grammar (e.g., Granger, 1998). These usages are then contrasted with a corpus of native speaker data that serves as the target usage. Most corpora, however, including those proposed in this article, focus only on collecting the language produced by native speakers.

Another application of text corpora is in translation; here work often relies on “parallel corpora”—two translations of the same material. With these, students examine aligned segments of the texts to observe contrasting constructions. (Special software—such as Multiconcord, created
by David Woolls at University of Birmingham, and the WinAlign features of Trados Freelance—is available to align translated texts.)

Finally, in writing instruction, corpora are especially useful as representative genres or styles that can be accessed as models for writers, in both discourse layout (such as contrastive rhetoric) and academic or other specialized vocabulary. For details on using corpora in writing instruction, see work by Chris Tribble (1997, 2001a, 2001b among others).

As can be seen from the examples above, much language teaching could benefit from Type II technology uses such as accessing computerized corpora. In this paper, however, the focus will be on a skill that plays a crucial role in reading classes, but also in many other areas of language study: vocabulary acquisition. Concordances have been produced for centuries, as every word of a particular body of work—say the Bible, the Koran, or the complete works of Shakespeare—was meticulously itemized by hand to show the distribution of each word in the oeuvre. For language students, using a concordance allows a chance to learn vocabulary nuances from context. That is, students act as their own lexicographers, discerning senses from a set of real uses in context. This is what Johns calls “data-driven learning” (Hadley, 2002; Johns, 1994). Research has shown that readers make use of different techniques in engaging with text. They may do intensive reading, which involves carefully working through a text word for word; extensive reading, where readers move more quickly, as in pleasure reading; skimming, in which they attempt to gather the gist of a text; and scanning, in which they hunt for a particular piece or type of information. A computerized concordance program itself works at the scanning stage, moving quickly through a text of any size, and collecting each occurrence of the word in question into a list in which each line shows one sentence using that word, with the key word highlighted. With this list in hand, a student has the data needed to proceed with his/her observations and to try out inductive strategies in order to sort out the senses and uses of the word. (While it seems intuitively useful for students to be able to pull together examples via concordancing, the assumptions that such concordance-based vocabulary learning is, in fact, effective is explicitly tested and verified in Cobb [2004] in a study using a group of first year Arabic-speaking college students who were learning English. He found that learners consistently retained vocabulary knowledge at higher rates when they gathered the meanings by means of multiple-context examples—as offered by concordance software—than they did by using other methods of mastering new words.)
While students have the goal of reaching conclusions based on contrasting output from a concordance program, teachers, on the other hand, need to choose the sources to be used as input, basing their choices on two areas: (a) the frequency and difficulty of the words for their students and (b) the interests of the particular students. Depending on the students’ proficiency, teachers may compile and concord materials for beginning students, then go on to train more advanced students to compile and concord their own material. Kennedy and Miceli (2001), however, in teaching a group of intermediate learners of Italian who were working with a corpus of contemporary written Italian, found that many of the students’ problems with the concordancing task stemmed not from lack of language proficiency but from their being inadequately prepared to “acquire certain attitudes and habits of researching” (p. 87). That is, Kennedy and Miceli suggest that teachers invest time up front marketing the observations that can come from mining such a language database.

The relevance of word frequency, the first consideration in choosing vocabulary material, has been extensively studied by Paul Nation and an affiliated group of scholars in New Zealand. Describing Nation’s 1990 book *Teaching and Learning Vocabulary*, Cobb (2002) summarizes the state of frequency studies:

Computer analysis shows that about 80% of the individual words in most written English texts are members of the 2000 most frequent word families, so that any second language reading course should ensure that its users meet and know these words. After roughly the 2000 mark, however, the pay off for direct learning trails off, and at that point learners should either rely on inferencing strategies or else move on to direct study of items that are frequent not in the language at large but in chosen areas of study or interest such as academic texts in general or domains of study like economics in particular. Either way, the goal is to arrive at a point where 95% of the running words are known in an average text, which a series of experiments show is the point where independent reading and further acquisition through inference become reliable. (p. 243)

For teachers seeking information on the words in a given text, Nation and Coxhead have produced the software Range (formerly Fwords and VocabProfile) and Frequency–Windows programs for analyzing lexical distribution in texts. Studies of the 2,000 most frequent words were
first set forth in a list of “general service” words compiled by West (1953); while more recently, lists of words in academic English have been compiled by Xue and Nation (1984) and Coxhead (2000).

The second aspect in choosing texts to input, which involves connecting to the reader’s background knowledge and interest in the topic, is also known to be important in motivating readers. Specifically, reading teachers are encouraged to activate readers’ “schema” (Carrell & Eisterhold, 1983; Carrell, 1984) by either bringing them up to speed on necessary background concepts, or first encouraging them to recall what they may already know on a topic. Either way, students are more inclined to keep acquiring new terms and increasing their reading speed if they have some initial interest in or need to know more about a given subject. Thus, teachers are encouraged to either have some say in the selection of appropriate reading materials, or even to let readers choose their own. This supports the idea of using specialized corpora from which to cull new vocabulary. If students are engaged in learning and writing outside of class on the topics of chemistry or folktales, for example, then readings on history or sports may not make the most motivating vocabulary sources, and vice versa. Accessing a general language corpus is also less likely to reveal the relevant uses of the terms that students need to acquire in their chosen areas of reading.

Many English as a Second Language reference books and teacher training texts have begun to acknowledge the need to incorporate concordancing and frequency tabulating technology, but often in a preliminary way. Three books are examined here, as being typical of the different types of areas that teacher texts focus on. The first, represented by Omaggio Hadley (2001), is a general methodology textbook for teachers of ESL. The author devotes two pages to a section discussing how computers can provide “richer content for language learning” (p. 163) but cites only two works, both from 1998, that present an overview of technology issues in language study. These highlight ways to use authentic material, including the Web, and caution about the need for teacher input, though no particular analytical tools are presented. Thus the issue of the pedagogical impact of technology is raised, but the interested teacher must seek details elsewhere. Anderson (1999), in a book that focuses just on reading skills, goes a bit further. This author discusses the potential benefits of using indexing software in vocabulary studies, mentioning the DOS program VocabProfile, and going on to include a suggestion that readers download MonoConc and try it on a text being used for a reading class (pp. 31-32). These are more specific starting points, but the coverage in the text is quite brief, and the soft-
ware mentioned is neither free, platform independent, nor current. Finally, Boswood (1997) edits a useful handbook of short, hands-on classroom techniques. He includes six concordancing selections by different authors, which look at such topics as collecting commonly co-occurring word pairs, developing vocabulary, creating corpora, and learning to do concordancing. This book is more focused, but assumes a familiarity with the software being discussed, as in these instructions for setting up a concordancing lesson on overused vocabulary words: “Procedure: 1. Instruct the students in the use of the concordancer if necessary. . . .” (p. 266). Despite such rudimentary coverage in general Teaching of English as Second Language books, many other works specifically focusing on issues of computers in language learning are available for those who are already intending to incorporate such technology (as seen in specialist subsets of professional organizations, such as the Computer Assisted Language Learning Interest Section [CALL-IS] of the professional group Teachers of English to Speakers of Other Languages or the Computer Studies in Language and Literature Discussion Group of the Modern Language Association). Many uninitiated teachers, however, may be surprised to find how easily such technology tools fit into their existing teaching goals.

FINDING THE TOOLS

Having decided to attempt a data-driven approach to vocabulary instruction, where does a teacher find the two necessary software types—the corpora themselves and the concordancing programs? Several issues are involved in acquiring these tools.

First, as far as the collection of texts to sample, many bodies of texts are available already, free or to purchase, on CD or via download. Of course, to end up with a corpus specific to a given class, one attractive option is to compile one’s own texts. Getting texts into electronic form is easier than it once was. Typing and downloading are the most likely methods, with scanning and optical character recognition, or inputting via voice recognition software as additional possibilities. (See Bowker & Pearson [2002] for details, however, as these latter methods can be more time consuming than they first appear, and generally require additional software as well.) Material on the Web, as mentioned earlier, can be mined for public domain sources such as the fiction being continually added to Project Gutenberg, selections of e-mail postings and chat, professional journalism sites, and so on, although another caveat is that
many of these will be copyrighted. Once a teacher has selected the appropriate texts to work with, another consideration is whether the corpus should be “tagged,” where tagged means that each word’s part of speech is labeled. Clearly, there are advantages to having a text that is already tagged for part of speech (so that, for example, students can hunt for just the verb *run*, rather than the noun *run*, or check for a suffix that attaches only to a certain class of word, such as the *-ly* that attaches to adjectives in order to form adverbs, as found in *quickly*, and *suddenly*, rather than the less frequent *-ly* that attaches to nouns to form adjectives, as found in *ghostly* and *friendly*). However, the likelihood is that most teachers and students will have available to them only untagged texts. In fact, the sorting of similar looking word forms is one of the tasks that students would then be able to profitably perform themselves. (Not to mention the difficulties in finding one set of tags that all users would agree on and be fluent in. Most text tagging is done by linguists rather than language teachers, and hence contains much finer, and often more cryptic, distinctions than the familiar “principal parts of speech” that students first encounter. When doing their first-hand tagging exercise, even computational linguistics graduate students are often initially overwhelmed by the options provided by the 65-180 members of various tag sets.)

After a corpus has been obtained, the second step is to try out some concordancing and indexing software tools on the texts. Many sources exist for specially designed software for concordancing and indexing texts. Mike Scott’s WordSmith Tools suite and Michael Barlow’s MonoConc, for example, are both full of features and well supported, though both only run in a Windows environment. For Macintosh users, SIL’s Conc and Mark Zimmermans’s old HyperCard stack FreeText are both also packed with features, and have the additional advantage of being free. Users with little time or money to invest in new software skills, however, may find it is possible to start by accessing some concordance-like tools they already have at hand. For example, for concordancing abilities, those familiar with the UNIX environment may recognize that the command “grep” will scan for and gather all instances of a given token from a designated text file. Stevens (1995) contains a brief discussion of DOS commands that do likewise (p. 4), thus command line users have options available. For Windows and Macintosh users, the newest free version of Adobe Reader (6.0) now has a concordancing feature in its “find” function, so that a basic word search of any PDF file now brings up a small keyword in context list for the word.
As for tools that count and sort words, separate modules of the four concordance programs mentioned above each present windows that give a list of every different word in a text—ordered either alphabetically or in some cases by frequency. However, teachers who already have a word-processing or spreadsheet program should check the “sort” feature (found, for example, under the Table menu of Word and under the Data menu of Excel) which will allow one to sort a list of words alphabetically as well. And free indexing tools on the Web (e.g., Cathy Ball’s Web Frequency Indexer) will also sort text that is pasted in, giving back a list sorted either by frequency or alphabetical order. By combining the output of these various free sorting programs, a teacher can readily get a list of high- or low-frequency words in a given text selection, and from this information, select relevant terms with which the students can run a concordance search to then try working out the meanings from contextual and linguistic clues.

The sections discussed here have pointed out the Type II aspects of using specialized text corpora and concordancing tools to help students improve vocabulary learning in a second language. This student-centered, data-driven learning method can be easily adapted for different levels of reading students. Students will benefit by getting practice in observation and critical thinking as well as language mastery itself. Teachers are encouraged to try out ways of acquiring electronic copies of texts containing specialized authentic language as well as finding sources of tools for sorting and concordancing these texts.

REFERENCES

CALL-IS. Computer Assisted Language Learning Interest Section of TESOL. http://www.uoregon.edu/~call/

Trados Freelance (Version 6.5) [Computer software]. WinAlign features detailed at http://www.trados.com/offer/prz/t6new.html


