Tom Salsbury and Crista Crummer

# Using Teacher-Developed Corpora in the CBI Classroom

This article advocates the inclusion of teacher-generated corpora and activities that involve student analyses of corpora in content-based instruction (CBI). The content-based course discussed in this article was designed for engineering and architecture students at a large Mexican university. The corpus was compiled from textbooks as well as from representative journal articles and manuals that students use in their university content courses. We made this corpus available to the students enrolled in the class, and it became the corpus that students analyzed to reinforce lessons with grammar, vocabulary, and writing.

Such teacher-facilitated analyses of authentic language samples found in spoken and written language corpora help students recognize and identify patterns in language use that are different from how they intuitively use the foreign language and even from how the language is presented and taught in their textbooks. Thus, through a well-focused, teacher-directed look at these language patterns, students can become more target-like in their written and oral production. Moreover, corpus-based instruction is especially valuable in the case of CBI in an English as a foreign language (EFL) teaching and learning environment because the teacher's corpus could feasibly comprise authentic texts centered around one theme (in this case, engineering), which would allow students to increase their vocabulary knowledge and awareness of academic language use patterns in a specific content area or academic field.

## Corpus linguistics

To clarify, a corpus is a body of written and spoken language that teachers and researchers collect and analyze. From the corpus analysis of a text, researchers, teachers, and students are able to obtain word frequency counts as well as valuable information about how various words function and form patterns in the text in relation to other words. These linguistic associations (i.e., the patterns among words) fall into two categories: lexical and grammatical, which are referred to as *collocations* and *colligations*, respectively.

Lexical associations, or collocations, are how certain words co-occur in spoken or written texts. By performing a corpus analysis of the collocations of texts, language teachers and students are able to target key vocabulary in certain genres. For example, Stoller and Jones (2003) analyzed the use of the word *paper* in their corpus of chemistry texts, and found that it most frequently followed the phrase *in this*, to form the collocation of *in this paper*. Similarly, Stoller and Jones (2003) analyzed the word *shown* to demonstrate how it differs in use in the introduction, results, and discussion sections of peer-reviewed journal articles.

Grammatical associations, or colligations, are then derived from the lexical patterns that tend to occur in a text. By performing a corpus analysis of the colligations of texts, language teachers and students are able to confirm their own personal intuitions about grammar use in certain genres. Moreover, through the use of corpora, grammar patterns can be observed, introduced, and taught in the context of their actual use in oral and written discourse. For example, Stoller and Jones (2003) analyzed the frequency of passive voice in chemistry texts, in addition to the most common passive verbs, providing valuable information for students of chemistry on how to become a part of the chemistry discourse community. Stoller and Jones (2003) point out that the high frequency use of passive voice in chemistry texts contradicts the popular myth among English teachers that the use of passive voice contributes to weak writing practices.

### Authenticity of task and text

In teaching from corpora with second language learners, the task, not the text itself, is the most relevant component of authenticity (Mishan 2004). This is because the actual texts in a computer-based corpus are stripped of all formatting, making it impossible to discern titles, headers, and captions from the actual body of the text. Illustrations, which provide important context for comprehension and are essential in conveying a text's authenticity to the reader, are likewise absent from a text once it is made ready to be analyzed on the computer. Similarly, spoken texts in a corpus do not include questions and responses from other participants in the conversation, making much of the context and meaning of the spoken text incomprehensible. Indeed, the texts that comprise a corpus that is analyzed electronically, though authentic in word use, are void of the essential context that originally gave them shape.

The authenticity of a text depends on the reader's relationship to it (Widdowson 2000), and this relationship is created by the task that the student is engaged in with the text. Thus, for use in second language pedagogy, a corpus must be authenticated via the tasks assigned with it and the kinds of questions that students and teachers ask of the corpus (McCarthy and Carter 2001). The methodology of engaging second language learners in research tasks with a corpus is referred to as data-driven learning (DDL) (Mishan 2004). Students engage in corpus research through bottom-up, inductive reasoning as well as top-down, deductive reasoning via the traditional research method of forming hypotheses, experimenting, and drawing conclusions.

Content-based instruction draws heavily, if not entirely, from authentic texts. Our approach to authenticity parallels Wiesen (2000) in that we gave students science and engineering texts in their original format. Some of the texts were abridged but none were simplified. Our learners were motivated to see authentic texts in their English language course, and we found little of the initial resistance from students to theme-based CBI discussed in the literature (Stoller 2004; Wiesen 2000). Part of the enthusiasm for the course stemmed from the data-driven tasks that explored the science and engineering corpus that we compiled. Our approach to authenticity was multi-faceted. Students read authentic texts designed for high school and college students in the United States. They watched video recordings of American news specials and educational material related to the science and engineering themes chosen for the course. They engaged in academic reading and writing tasks using peer-reviewed journal articles. They participated in authentic speaking tasks via poster-session presentations, discussions and debates. Finally, they engaged in authentic research in language study using a corpus—the focus of this article.

Because we worked directly with the content area faculty, we were able to choose texts to analyze that were representative of what the students actually read in their major, and we received advice about the most common discussion topics and types of interactions in the field. Students were involved in text analysis, using the same concordancing programs that we employed as teachers. (See Reppen 2001 for a review of *MonoConc Pro and WordSmith Tools*.) Involvement in text analysis encourages students to be self-directed language learners, systematically analyzing academic uses of the second language. We wanted students to supplement dictionary work with searches of the corpus for content-specific uses of words and grammatical expressions.

### Compiling a corpus

The one-million word corpus for this project came from nearly 200 texts taken from engineering textbooks, articles from engineering journals, lab manuals from engineering classes, and transcripts from the Michigan Corpus of Academic Spoken English (MICASE). We were particularly careful to select from various disciplines within the College of Engineering, Architecture, and Design; thus, our corpus reflects written and spoken texts from electrical, mechanical, and industrial engineering as well as texts from industrial design and architecture.

McCarthy and Carter (2001) stress that the size of a corpus is less important than the texts the teacher chooses, how those texts are classified, and how the teacher (or students) analyzes the texts. We make the following suggestions to teachers compiling a corpus:

1. Choose texts or text types that students are likely to read. Since scanning texts can be time consuming, we suggest starting with key course readings from the first part of the course. Texts that are available in electronic format will facilitate the compilation of the corpus. Many corpora are available for free online (see Appendix); thus, a teacher may choose to scan just one chapter of a class reading and use online corpora to generalize findings.

2. Save the original electronic text as an Adobe Acrobat (*.pdf*) or Microsoft Word (*.doc*) file. Since a corpus decontextualizes language, it is important to have the option of re-contextualizing the data if necessary. (This option is generally not available for corpora found online.) Of course, the original format is essential if students are to read a text as an assigned reading. The text (*.txt*) files used in corpus analysis are not appropriate for class readings because they are comprised of uniform fonts and spacing that make it difficult to distinguish titles, headers, and captions from the body of the text. Illustrations and graphics are also absent.

3. Check the scanner settings when scanning paper texts to electronic format. If scanning to a *.pdf* file, the scan must be editable so that the file can later be saved as a text file. The scanner should be set to English if this is the target language; otherwise, the scanner will not recognize many words, which will greatly complicate the revision of each text. All scans of paper texts should be spell-checked for errors.

4. Save all electronic files (scans or originals) as text files, and include a short file name that refers back to the original text. Classify the files by text type—e.g., written language, spoken language—and store them in folders by type. Scans of entire textbooks should be divided by chapter, with each chapter stored as a separate file. In this way, students and teachers can cross reference the text file with the source text in easy-to-manage chapters, making it much simpler to go back to the original for context, content, and meaning.

5. Compress the corpus files and folders to a *.zip* file so that colleagues and students can easily download the corpus from a website, email message, or electronic learning platform. The corpus can then be extracted and saved to the hard drives of students' computers.

6. Take advantage of affordable and free corpus analysis tools available online.

Popular corpus analysis tools such as *WordSmith Tools* are available for free with limited functionality. This is a good way for teachers to experiment with the software before making a purchase. The free version of the software may be sufficient for the needs of the teacher and students. *WordSmith Tools* allows for 20 concordances, and it also returns totals for word searches (though it does not list the concordances beyond the first 20). We found the free version of the software sufficient for our students' data-driven learning tasks while we opted to purchase the software for our own computers.

7. Respect the learning curves of collaborating teachers and students. Using corpus analysis software can be difficult at first. Allow yourself and collaborating teachers plenty of time to become familiar with the software. We found it helpful to practice selecting texts and specifying word searches. Students were generally comfortable with the technology, yet to ensure that everybody knew how to navigate through the activities with the corpus, we spent two one-hour sessions in the first week of class making sure that students had access to the necessary files and software and were able to perform simple searches of their own. In working with our students, we prepared screen shots for each step in the process and modeled in class how to download the software and the corpus, select texts, specify words to search, and read the output.

### The case of *because* and *although*

One of our first analyses of the corpus involved comparing the written texts (journal articles, textbooks, lab manuals) to the spoken texts (MICASE). We examined the placement of subordinators according to whether they were in sentence-initial position, where the subordinate clause precedes the main clause, or in what we termed the middle position, where the main clause precedes the subordinate clause. Here we focus only on the subordinators *because* and *although*.

The oral data in our corpus has no occurrence of sentence-initial *because*; rather *because* functions in a subordinate clause to a previous statement (e.g., main clause, subordinate clause), as in the following examples from the corpus:

1. uh you can't do a swap in two statements *because* you lose one of the values

2. It turns out these don't come out at exactly a hundred and eighty degrees *because* there is the nu- the uh nuclear recoil happening.

3. we haven't been able to test them much because we don't even have model identification

In the written data, in contrast, *because* appears in both environments: in sentence-initial subordinate clauses as well as in the middle position (as in the oral data). The following examples from the data illustrate initial *because* (4) and sentence medial *because* (5).

4. *Because* they used quasi-static field theory and circuit method, their methods cannot be used any more when the frequency is higher than 1 MHz.

5. This is not a real sacrifice in performance *because* the necessary attention to transition details is an essential element of design.

Written data provides more examples of sentence-initial *because*, as seen in Example (4), but it is still not as common as the middle position, reflecting the same bias that is found in the oral data.

Interestingly, *although* has an inverse relationship to the data that we find for *because*. Biber, Johansson, Leech, Conrad, and Finegan (1999) found that *although* is more frequent in academic [written] prose than in oral conversation. Our data confirms this finding: *although* rarely appears in the oral data. However, in the written data, where it is found, it is most common in sentence-initial position, as in the following examples:

6. *Although* these projects were used in a computer architecture course, they could easily be applied in other courses.

7. *Although* the quantity of modes is usually unlimited, the higher modes have the smallest amplitudes and thus can be neglected.

We guided our students through similar bottom-up enquiry using the corpus. First, we enhanced an assigned reading on engineering ethics by highlighting all occurrences of *although* in yellow and all occurrences of *because* in light blue. Students were asked to work in groups to read the text, scanning first for the yellow highlighted words, and then writing down at least five example sentences, including punctuation, on the left-hand side of a T-chart. They repeated this procedure for the words highlighted in light blue, writing down example sentences on the right-hand side of the same chart. In pairs, students analyzed each sentence for the following: (a) the placement of the highlighted word in the sentence, (b) words that functioned as the subject of the sentence, (c) words that functioned as the verb in the sentence, and (d) observations about punctuation. As a class, we generated a grammatical rule to describe how *because* and *although* are used. This rule became our hypothesis.

Next, we tested our hypothesis with our entire corpus of spoken and written language in top-down enquiry. Students worked together with their laptop computers to search our corpus for all occurrences of *because* and *although*. Students saw firsthand the same patterns that we had found and came to the same conclusions that we had reached. We believe that such an exercise raises consciousness about the grammatical point in question and functions as a much more powerful instructional technique than the traditional presentation/practice model—which characteristically comprises a few artificial sentences from published teaching materials or teacher intuition.

### Modal auxiliaries—The case of *will* vs. *BE going to*

As in the previous activity, students started out with an assigned chapter to read which formed the basis for the inductive language work. The chapter dealt with the notion of teamwork, and the class sessions earlier in the week had been devoted to developing student comprehension of the content through various individual and small group activities.

Students searched the electronic version of the chapter for a list of modal auxiliaries and recorded the frequency of each. They ranked each modal in order of frequency from most to least frequent. Figure 1 lists the modal auxiliaries that the students found and their corresponding frequency in the corpus. *Will* was the most frequent modal auxiliary followed by *should* and then *can*. Modal auxiliaries that did not appear at all in the chapter were *shall, might, have to,* and *BE going to* (the last two being periphrastic or semi-modals).

**Figure 1. Sample exercise with modal auxiliaries and corpus analysis**

1. Search the corpus for the following modal auxiliaries. Indicate the frequency of each.

| | | |
|---|---|---|
| __*11*__can | __*25*__will | __*4*__must |
| __*4*__could | __*2*__would | __*0*__have to |
| __*0*__shall | __*3*__may | __*1*__want to |
| __*20*__should | __*0*__might | __*0*__BE going to |

2. Rank the modal auxiliaries in the text in order of frequency. Which are the most frequent modal auxiliaries? Which do not appear at all?

Students were then asked to look again at the colligations of the five most frequent modal auxiliaries in the text and to write down the three to five words that appeared to the right and left of the target word in the sentence, thus briefly reviewing the form of *modal auxiliary + verb* combinations as well as adverb placement (e.g., "we will *promptly* go").

Next, students focused on *will*, the most common modal auxiliary in the corpus. They searched for this word with *WordSmith Tools* and used the view function in the program to see the entire sentence for each instance of *will*. Working in small groups, students attempted to generate three distinct definitions of *will* depending on the context of each sentence. Thus, students engaged in guided exploration with the instructor to discover not only that *will* referred to future time (e.g., "It would be naive to think that everything *will* always go well with teams.") or that it appeared in conditionals (e.g., "If not, frustration *will* build up and lead to conflict.") but that it also carried a sense of volition or willingness to do some-

thing (e.g., We *will* be as open as possible but will honor the right of privacy."). In fact, *will* for volition was the most frequent use in this particular textbook chapter.

As a final activity in the class, students searched the entire written corpus to compare the frequencies of *will* and *BE going to*. What they found was interesting: *will* occurred 2,038 times in the corpus whereas *going to* with an inflected form of *BE* occurred only 30 times—that is, one occurrence of *BE going to* for every 68 occurrences of *will*. Then, students searched the oral data for the same modal expressions. This time, *BE going to* occurred 52 times and *will* occurred 141 times, or roughly one occurrence of *BE going to* for every three occurrences of *will*. From this exercise, students concluded that *will* is generally a more common expression for expressing future time (it occurs more frequently than *BE going to* in both the oral and written data). However, students also learned that as an expression of future time, *BE going to* is more likely to occur in spoken English than in written English.

### Vocabulary analysis—The case of *embodied*

Students were given opportunities to combine dictionary work with corpus analysis work. In addition to finding definitions of new words, students also discovered peculiarities of certain lexical items. One example comes from a search for the word *embodied*, which appeared in an introductory reading on ethics in engineering: "Professionals claim to be regulated by ethical standards, usually *embodied* in a code of ethics" (Harris, Pritchard and Rabins 2004, 13).

Students were instructed to use *WordSmith Tools* to find all collocates of *embodied* in the corpus. Figure 2 illustrates what students found in their searches of the corpus.

As shown in Figure 2, the corpus-search results of *embodied* show that a majority of collocates (11 out of 13) were followed by the preposition *in* (e.g., *embodied in*). Secondly, most examples of *embodied* from the corpus are the past participle (rather than the past tense). In fact, in every case where *in* follows *embodied*, the form of the verb is the past participle in a passive voice construction. We reviewed with students that the English passive voice is constructed with a form of the verb *BE + past participle*. Interestingly, only the first and the ninth concordances in Figure 2 follow this pattern. Nine out of the eleven passive constructions occur in reduced relative clauses (participial phrases), as in this example from the eighth concordance: "the concepts of industrial efficiency [that are] embodied in the discipline of IE…." The relative pronoun *that* and the auxiliary *are* have been omitted so that it reads *the concepts of industrial efficiency embodied in the discipline*. This was the most common type of occurrence of *embodied* in the data, contrasting with a prototypical account of the English passive voice as consisting of *BE + past participle*.

We designed multiple choice questions to guide students in their exploration of the form and meaning of this word. As a class, we searched for *embodied* in the corpus and found the definition of *embody* in the dictionary. We then carefully designed multiple choice questions to guide the students' investigations. Some sample questions are listed in Figure 3.

The questions in Figure 3 build on each other. The first item focuses student attention on the collocation with *in*. The next three items review verb forms. Students review past

**Figure 2. Example output for *embodied* in *WordSmith Tools***

| N | Concordance | | |
|---|---|---|---|
| 1 | modern history of suspension bridges begins in the early nineteenth century and is | embodied | in such masterpieces as Thomas Telford's Menai Strait Suspension Brid |
| 2 | et would eventually be proposed, which would have essentially doubled the span as | embodied | in the existing state of the art in one leap. This seems not to have bother |
| 3 | ance and the appearance of the winner, George Stephenson's Rocket. This engine | embodied | all the basic technical features later extended and developed in locomoti |
| 4 | s-comparison of the two can yield some interesting parallels. Though the concepts | embodied | in TQM may be relatively new, the concepts of industrial efficiency embo |
| 5 | any problem. This principle broadly translates into strategic quality management as | embodied | in BS 7850[3]. * The next principle refers to competent counsel or expert |
| 6 | ional autonomy. 5. Professionals claim to be regulated by ethical standards, usually | embodied | in a code of ethics. The degree of control that professions possess over t |
| 7 | fifth common characteristic of a profession: regulation by ethical standards, usually | embodied | in a code of ethics. We have already noted that profes |
| 8 | ncepts embodied in TQM may be relatively new, the concepts of industrial efficiency | embodied | in the discipline of IE are almost a century old. One of the most widely a |
| 9 | y done in collaboration, and thus all the elements of the process are not necessarily | embodied | in a single individual. Mainstone (in Pugsley, Mainstone, and Sutherland, |
| 10 | on for a better life. The predominant aesthetic ideal of the nineteenth century in fact | embodied | a vision of harmony between utility and beauty that was articulated at le |
| 11 | l and extraordinarily large scale, the lessons of the abnormal and extraordinary - as | embodied | in case histories of failures and how to use them - are much more relevan |
| 12 | idge (Fig. 7.4), completed in 1825. For all the beauty and technological achievement | embodied | in this structure's 580- foot span hung from wrought-iron chains, it was pla |
| 13 | tial information from her firm. For the public, the existence of professional standards | embodied | in codes of ethics enables a potential client or customer to make certain |

**Figure 3. Sample multiple choice items in a corpus activity with embodied**

1. Which of the following is the most important finding from the search?
   a. *Embodied* is almost always followed by *in*.
   b. *Embodied* appears in the search results.
   c. *Embodied* occurs in a technical context.
   d. *Embodied* is preceded by many different things.
2. The *Longman Dictionary* states that *embodied* is both a past tense and past participle of the verb *embody*. Which form of the verb *embody* is used in the original text?
   a. The past tense (e.g., "This engine *embodied* all the basic technical features.")
   b. The past participle (e.g., "The history of suspension bridges *is embodied* in such masterpieces as…")
   c. Both a and b
   d. None of the above
3. Which form of the verb *embody* is most common in the concordances? (HINT: all instances of *embodied in* are the same verb form.)
   a. The past tense
   b. The past participle
   c. Both a and b
   d. None of the above
4. You have probably learned in previous English courses that the passive voice in English is constructed with a form of **BE** + **Past Participle**. Given the results of the corpus search, how frequently is this the case with the passive voice construction of the verb *embody*?
   a. Zero occurrences
   b. No more than 2 occurrences
   c. At least 11 occurrences
   d. All 13 occurrences
5. The *Longman Dictionary* has two definitions for the word *embody*. Which of the following definitions is the closest in meaning to *embodied* as it is used in the original text?
   a. to be a very good example of an idea or quality
   b. to include something
   c. Both a and b
   d. None of the above

tense and past participle forms and apply this to the passive voice construction in the context of the new vocabulary word. Item 5 deals with the dictionary meaning of *embody* and asks students to choose the correct definition. A sixth item (not shown) pulled together the form, meaning, and use of *embodied*. Then we asked students to write two sentences using the verb *embody*. We could have elaborated further by asking students to illustrate in a paragraph how one thing is embodied in something else. Indeed, a corpus analysis can be a springboard to other meaningful language and content work.

### Additional ideas and activities

Mishan (2004) presents a way to teach phrasal verbs through corpus analysis of a single verb, such as get, and its most frequent collocations (e.g., *get in, get off, get on, get out*). Mishan also discusses how to engage students in exploring the use of *speak, say*, and *tell* in the corpus—three words commonly misused by learners of English. Such discovery learning exposes students to authentic language, and, most importantly, involves students in authentic uses of corpora in research tasks.

Proponents of CBI recognize that language is best taught using content that is interesting, relevant, and meaningful to students. Computerized tools to supplement the CBI approach, including corpus analysis tools, are readily available. However, these tools are by no means prerequisite to CBI course design. One alternative to electronic corpus analysis is a non-computerized, functional analysis of subordinators such as *when, after*, and *during*. (See Schleppegrell, Achugar, and Oteíza [2004] for a discussion of functional analysis.) Nonetheless, we believe that carefully constructed electronic searches of a much greater amount of text will better inform many language analyses. Teachers and students can learn much more about texts and the generalizability of this new information when those texts can be analyzed by computer software. This is particularly true when the source texts are available in their original format, as they will invariably be if teachers compile their own corpora from the same texts that students use in class.

Teachers and students can expand their analysis of lexical and grammatical features from a single passage or chapter of a text to the whole text as well as to other similar texts. For example, nouns, verbs, and conjunctions that contribute to a text's abstractness and ambiguity can be quickly searched with corpus analysis software. A corpus analysis is a particularly powerful aid to categorizing verbs, enabling students to explore the different ways words are presented and meaning is constructed. Schleppegrell et al. (2004) suggest that students put verbs into categories such as *saying and thinking verbs, relating verbs*, and other categories. Thus, the verbs *expressed* or *resent* can be easily explored by students in the entire text. The functional analysis of

participant roles in a text can also be supplemented by a thorough bottom-up analysis of an entire text. In a history text, for example, the word *government* will appear in many different contexts. Rather than analyzing only one context and a rather limited number of clause structures in which the word appears, students can explore how the word is used in many different passages, preferably from sections of the text that they have already read, but also as a preview to future sections and instructional units.

Helping students conceptualize metalanguage—the language used to talk about language—is another powerful use for a corpus and a compelling reason to engage students in its analysis. Learning metalanguage is an important component to any functional analysis of texts, electronic or otherwise. We argue that the authenticity of the language samples and research tasks discussed in this article greatly facilitate the conceptualization of the metalanguage necessary for most language analyses.

Figure 4 summarizes corpus-based ideas for classroom use of corpora. Teacher-compiled and professional corpora can provide opportunities for inductive and deductive learning. Sysoyev (1999) talks about student exploration as the first stage in integrative grammar teaching. As with bottom-up approaches to data-driven learning, integrative grammar teaching recognizes the value of allowing students to explore language and formulate rules inductively. A corpus, particularly one that is created by teachers and uses texts with which students are familiar, greatly increases the number of samples needed to induce a rule while at the same time maintaining the familiar context and content of the source text.

Mishan (2004) presents an innovative technique for student-compiled corpora using text sources from the Internet, particularly transcriptions of oral data. Learner-language corpora, typically compiled from students' written work but also possible with transcriptions of oral interviews, are extremely useful for teachers and scholars in longitudinal research with a small set of learners. Learner-language corpora can also be contrasted to native speaker corpora, which is particularly useful in comparing word choice in written essays. Unlike corpora compiled from class texts, learner-language corpora are designed principally for use as models in peer analysis of student writing.

Biber et al. (1999) present lexical bundles or formulaic chunks that occur in various registers. Formulaic chunks are simply collocations that occur together with very high frequency—so high, in fact, that second language learners may analyze them as a single unit or bundle. We have already presented the frequent collocation *in this paper* discussed by Stoller and Jones (2003) in scientific journal articles. Other examples are *would like, if I were, would you mind,* and *I think/don't think.* Students can explore these lexical bundles and other frequent collocations in both spoken and written discourse through analyses of corpora.

Mishan (2004), suggests that students take advantage of available transcripts on the Internet from popular TV shows in developing their own corpora. These transcripts are

**Figure 4. Ideas for using corpora in the classroom**

- Provide varied, authentic examples of the usage of vocabulary items or grammar structures already taught in class
- Guide students to draw conclusions about the meanings and/or usage patterns of vocabulary items or grammar structures found in a corpus
- Produce and use student-compiled corpus OR student learner language corpus
- Facilitate the contrastive analysis of language used in learner language corpus and native speaker corpus
- Compare and contrast spoken language corpus vs. written language corpus
- Locate formulaic chunks
- Raise awareness of speech acts
- Teach vocabulary collocations
- Teach grammar structure colligations
- Conduct a contrastive analysis between L1 and L2 for translation studies
- Experiment with the use of computer-based analysis (concordancer) vs. paper-based analysis

a good source for lexical bundles particularly as they occur in speech acts. Spoken language recorded in transcripts of science programs on TV (e.g., *NOVA*), can be contrasted to the written language used to discuss similar topics, and students can engage in a functional analysis of the language differences depending on register and language medium.

Bottom-up, inductive learning is facilitated by access to electronic corpora and the means to analyze those corpora. Students can explore myriad vocabulary collocations and grammar structure colligations in both bottom-up, inductive learning as well as top-down, deductive learning. In teacher or student developed corpora, the source texts should be available in their original format as we have already discussed. One of the key advantages to creating one's own corpus is that it allows for both computer-based and paper-based analyses. We emphasize that paper-based, functional analyses can be greatly supplemented by, but not replaced by, computer-based analysis.

## Conclusion

Corpus linguistics played an important role in the development of a content-based English course for students of engineering, architecture, and design. This article outlined the many ways that corpus analysis may be included in class activities to inform language study and complement functional analyses of texts for improved content learning. Because the students were involved firsthand in corpus analysis, they were invested in the process of learning about language. In this case, the language learning was particularly pertinent to the students as all of the textbooks that we used to compile the corpus came directly from student readings in their university-level content courses, and the journal articles that we chose were of the type that students would be expected to read in the upper-level classes towards the end of their undergraduate career. Focused language of this nature is much more representative of the kind of written and spoken language that students will encounter as professionals in their respective fields. Even the most experienced teachers cannot hope to intuit all the nuances of the specialized discourse within various disciplines.

Familiarity with corpus analysis tools becomes a powerful language learning aid for students, one that encourages self-directed learning. There are many lessons for teachers as well. Apart from learning more about language, teachers develop valuable new computer skills in learning to compile and analyze large amounts of language data. Using a scanner to convert paper texts to electronic texts, searching a library database for electronic journal articles, storing and managing large numbers of files, and organizing language data in an Excel spreadsheet are just a few examples of the computer skills that are fostered by a project such as this; these skills will translate into other professional and personal activities. It is both challenging and rewarding to hone new technological skills in the compilation of a language corpus and to discover something new about the language.

## References


Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman grammar of spoken and written English*. Harlow, UK: Pearson Education Limited.

Harris, C. Jr., M. Pritchard, and M. Rabins. 2004. *Engineering ethics: Concepts and cases*. Wadsworth Publishing.

*Longman dictionary of contemporary English*. 2003. Edinburgh Gate, UK: Pearson Education Limited.

*MICASE (Michigan Corpus of Academic Spoken English)*. University of Michigan, English Language Institute: http://micase.umdl.umich.edu/m/micase/.

McCarthy, M., and R. Carter. 2001. Size isn't everything: Spoken English, corpus, and the classroom. *TESOL Quarterly* 35 (2): 337–40.

Mishan, F. 2004. Authenticating corpora for language learning: A problem and its resolution. *ELT Journal* 58 (3): 219–27.

Reppen, R. 2001. Review of *MonoConc Pro* and *WordSmith Tools*. *Language Learning and Technology* 5 (3): 32–36.

Schleppegrell, M. J., M. Achugar, and T. Oteíza. 2004. The grammar of history: Enhancing content-based instruction through a functional focus on language. *TESOL Quarterly* 38 (1): 67–93.

Scott, M. *WordSmith Tools*. http://www.lexically.net/wordsmith/.

Stoller, F. L. 2004. Content-based instruction: Perspectives on curriculum planning. *Annual Review of Applied Linguistics* 24: 261–83.

Stoller, F. L, and J. Jones. 2003. *Using corpus tools for ESP materials development*. Paper presented at the TESOL International Convention, Baltimore, Maryland.

Sysoyev, P. V. 1999. Integrative L2 grammar teaching: exploration, explanation and expression. *The Internet TESOL Journal* 5 (6).

Widdowson, H. G. 2000. On the limitations of linguistics applied. *Applied Linguistics* 21(1): 3–25.

Wiesen, B. 2000. Content-based unit learning in English for Academic Purposes courses in teachers colleges. *Journal of Adolescent and Adult Literacy* 44 (4): 372–81.

TOM SALSBURY is an Assistant Professor in the College of Education at Washington State University. Previously, he directed the Department of Foreign Languages and Philology at the Tec de Monterrey in Mexico City, where he designed and taught content-based university English courses.

CRISTA CRUMMER lives in Querétaro, México, where she teaches English as a foreign language to high school students in the Tecnológico de Monterrey. She received her MA in Teaching English as a Second Language from Northern Arizona University in 2002.

## Appendix Websites for Coprora and Concordancers

***Using Teacher-Developed Corpora…*** • **Tom Salsbury and Crista Crummer**

**Bookmarks for Corpus-based Linguists**
David Lee, University of Hong Kong
**http://devoted.to/corpora**
A very complete website with links to corpora, concordancing programs, teaching resources, and academic papers about corpus linguistics

**British National Corpus**
University of Oxford
**http://www.natcorp.ox.ac.uk**
Allows users to search for specific words/phrases in the corpus and returns to random hits of those words/phrases

**Business Letter Corpus Online KWIC Concordancer**
**http://ysomeya.hp.infoseek.co.jp/**
An online concordancer that searches for words/phrases in a corpus of business letters

**Cobuild Concordance and Collocations Sampler**
Collins Wordbanks *Online* English corpus
HarperCollins Publishers
**http://www.collins.co.uk/Corpus/CorpusSearch.aspx**
Online concordancing program that allows user to do a limited search of the Cobuild corpus

**English Corner**
Resources of Interest for ESL Teachers
**http://www.angelfire.com/wi3/englishcorner/teachers/teachers.html#corpus**
A great source of links to websites about using corpus analysis and concordancing in the classroom

**HyperTextBooks Concordancer**
College of DuPage, Department of English
**http://papyr.com/applets/concordancer/**
An online concordancing program that allows you to upload your own texts in .txt format for analysis

**Michigan Corpus of Academic Spoken English (MICASE)**
University of Michigan English Language Institute
**http://www.lsa.umich.edu/eli/micase/index.htm**
The online corpus is available for browsing and searching for specific words/phrases; includes some discourse analysis activities using MICASE; provides links to other corpus sites

**Online Concordancer (v.4)**
**http://www.lextutor.ca/concordancers/concord_e.html**
An online concordancing program that allows users to select specific corpora and search them for collocations

**W3-C Corpus Linguistics: List of Corpora**
University of Essex
**http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/index2.html**
A list of different corpora available online

**Web Concordancer**
Virtual Language Center Hong Kong
**http://vlc.polyu.edu.hk/concordance/**
An online concordancing program that searches for words/phrases from a selection of corpora

**WebCorp**
Research and Development Unit for English Studies
**http://www.webcorp.org.uk/**
An online concordancing program that searches for words/phrases on webpages (i.e., the World Wide Web is the corpus)