Corpora in the classroom: Data-driven learning for Freshman English

Simon Smith

Foreign Language Center, National Chengchi University
smithsgj@nccu.edu.tw

Abstract

In recent years, Western ELT scholarship has emphasized student-centered learning, focusing in particular on Computer-Aided Language Learning (CALL) and the use of linguistic corpora, including Data-Driven Learning (DDL). In the universities of Taiwan, and many Asian nations, those teaching English to non-majors have to overcome certain hurdles if they are to successfully use DDL techniques. Class numbers tend to be high, with up to 70 students in the room, and computer workstations for individual use are not normally available. The present paper, therefore, shows how to make selective use of salient corpus examples before the whole class, for example when explaining language points in warm-up or wrap-up sessions. We focus in particular on the importance of making students aware of collocational patterns.

DDL tasks are often thought of as part of a student-centered research approach, and some experienced teachers believe that many non-major students lack the motivation or proficiency to engage with such tasks. The present paper compares that view to that of some European and Korean studies, where it is claimed that students at even low proficiency levels can benefit from DDL.

In an approach partly inspired by Content-Based Instruction, we try to motivate our students to learn from corpora by assigning real-world tasks. The tasks vary in nature, but often require the students to comment on a software or corpus usability issue, or to use (or develop) their own computer skills to present vocabulary or other corpus information in an effective way.

Introduction and background literature

It comes as no surprise, in this digital age, that many approaches to language teaching and learning rely on the use of the computer, and that countless computer-based tools for language learning have been developed over the last few decades.

Computer-assisted learning (CALL) is now of great significance and importance in the acquisition of second languages – especially English – in Taiwan and all over the world. There are entire journals devoted to research on the topic, including for example Computer Assisted Language Learning, published by Taylor and Francis. At nearly all language teaching institutions, the use of online resources is now commonplace in the language classroom, and listening labs are often computerized. Ellis (1995) notes that CALL has a particularly important role to play in the acquisition of vocabulary, because this is the part of language study to which the student can most usefully turn his attention in private. Thus, teacher contact hours can be devoted to more communicative activities that cannot so easily be practised alone. There is, indeed, a great variety of applications available on the web for students to use in private study. Many ESL textbooks now have accompanying websites with
interactive exercises, and there are CALL sites targeting grammar, such as WERTi (Metcalf and Meurers 2006; http://prospero.ling.ohio-state.edu/WERTi). Lextutor (http://www.lextutor.ca) offers vocabulary building exercises, while VISL (Bick 2005; http://beta.visl.sdu.dk/visl2) features both lexical and grammar-based tasks and games, including one where the learner is asked to “shoot the adjective”!

Our own TEDDCloG (Smith et al, forthcoming) system teaches and tests collocational knowledge through cloze exercises, and is based on the Sketch Engine suite of corpus query tools.

The use of linguistic corpora in language learning is often described in terms of data driven learning (DDL), described by Hadley (2002). In a clear parallel to data-driven computational algorithms, DDL attempts to impart linguistic knowledge by making available samples of authentic language, and inviting language learners to tease out the grammatical patterns for themselves. Other, more traditional, approaches teach the grammatical patterns explicitly, and make use of non-authentic materials, on the analogy of rules-based algorithms.

Johns (1991) likens the language learner (on the DDL model) as a researcher, analyzing target language data and becoming familiar with the language through the regularities and consistencies encountered. He notes that it is “intelligent, sophisticated, and well-motivated” students that are in the best position to benefit from DDL.

Hadley (2002) mentions resistance to the use of DDL in an Asian context, and notes his own initial uncertainty about using the approach in his Japanese ELT classroom, on the grounds that some Japanese are uncomfortable about being taught grammar, especially by a non-Japanese teacher. This concern probably applies to the Taiwan context also: university students here sometimes feel that they have already in some sense acquired English grammar, and all that remains for the perfection of their knowledge of English is to learn more vocabulary. As many practising teachers at university level will testify, this is in fact very far from the truth. Furthermore, DDL techniques can be just as effectively applied to the acquisition of vocabulary as to that of grammar.

Regrettably, the way that target language vocabulary is generally acquired in Taiwan flies in the face of modern pedagogy. A highly traditional approach is normally taken, where the student memorizes items from a list: this is probably the only learner-centred task in which the average university student engages. No special merit seems to be attached to wide reading or inferring of meaning from context, such as might lead to long-term retention. It is commonly assumed that almost every word of English has an exact Chinese equivalent to which its meaning may be mapped; the perceived goal is to figure out which words are likely to come up in a future test and memorize these mappings.

Another potential objection is that learning through introspection and reflection does not seem to fit too comfortably into the received model of Asian pedagogy. Most stakeholders (students, parents, university administrators, education officials and unfortunately many teachers) have a fairly traditional understanding of the learning process, where essentially the teacher delivers content and the students somehow absorb it (or the students practise production, and the teacher corrects them where necessary). This is despite the best efforts of
Taiwan universities, both private and public, to encourage and reward student-centred learning.

The problem is exacerbated in situations where class sizes are large, and students are relatively unmotivated. Even if it could be readily demonstrated that DDL yields better acquisition results than traditional approaches (that is, DDL-taught students are more likely to acquire written and spoken fluency in the target language), this would not necessarily prove that the needs of Taiwanese university students were being met. Communicative competence is only one of a range of such needs; others can include passing proficiency tests (which may only make a partial assessment of such competence), meeting a stipulated number of teacher contact hours, socializing with classmates from other academic departments, and learning English vocabulary for use in Mandarin-English code-switching exchanges, quite common in Taiwanese professional life. If the teacher is not Taiwanese, cross-cultural communication needs may be addressed, and skills acquired.

Despite all this, Hadley (2002) and Lee (2007, a Korean high school study) reported some success with the approach in Asian classrooms. Also, Boulton (2007) described an experiment in which relatively unmotivated French students reacted positively to a DDL treatment of English phrasal verb instruction. The lack of motivation that Boulton identifies is not dissimilar to that often encountered in Taiwan; namely, that the English course being taken is to fulfil credit requirements only, and the student is unlikely to need English skills in future professional life.

Boulton’s experiment consisted of three phases. First, there was a pre-test, where students’ ability to choose between pick or pick up, and look or look up was examined. Next, the students were provided with a concordance sample of 25 lines featuring the same four lexical items. They were given 10 minutes to peruse the data. Finally, a post-test similar to the pre-test was administered, and it was found that the performance of all levels of learner did indeed improve. Boulton is very cautious about his findings, and does not conclude from them that DDL is successful as an approach. Rather, he hopes that other teachers will conduct related experiments with their own students, and that ultimately the research community as a whole will reach firmer conclusions about the usefulness of DDL.

Another point that needs to be made about Boulton’s work is that the whole procedure took only half an hour of class time; and that of that, only 10 minutes was spent on reading the concordance. Poring over this rather dry data for a short time might be acceptable for most students, but any more than that and boredom is likely to set in, with affective filters locking securely up in all but the most motivated students. As Kilgarriff et al (2008) put it, “The bald fact is that reading concordances is too tough for most learners. Reading concordances is an advanced linguistic skill.”

The Sketch Engine

Kilgarriff’s Sketch Engine (SkE) corpus query toolset goes some way towards “softening” concordance data up for language learners. Concordances themselves are enhanced by making available a sentence mode, as well as the traditional KWIC mode, so that more may be gathered from the context. Concordance lines can also be ranked by quality: a “good” example sentence is defined by Kilgarriff et al as one which is neither too short nor
too long, which doesn’t contain a lot of rare words or anaphors (which can sometimes only be resolved by looking outside the sentence), and is constrained by a few other parameters specified by the team. The first page of concordance sentences for the word *corpus* is shown in Figure 1, in default order, while Figure 2 shows the “best” examples first. The results speak for themselves.

**Figure 1** Concordance output for *corpus*, default order

Russell does not deny the existence or the importance of the *corpus delicti*. For further information please visit the Centre for Corpus Research website. An overview of some of the areas where *corpora* have been used can be found on the Research areas page. You may have used *corpora* extensively in the past, or you may simply have been intrigued by seeing demonstrations of their use. It is already used for coding large monolingual *corpora* (for example, the British National *Corpus* of 100 million words).

Introduction to Corpus Linguistics This includes: the use of corpus access techniques; interpretation of *corpus* data; corpus-based theories of language.

The English Department of the University of Birmingham is one of the leading centres for the study of Corpus Linguistics in Britain. Introduction to Corpus Linguistics This includes: the use of *corpus* access techniques; interpretation of corpus data; corpus-based theories of language.

It is already used for coding large monolingual *corpora* (for example, the British National Corpus of 100 million words).

Introduction to Corpus Linguistics This includes: the use of corpus access techniques; interpretation of corpus data; corpus-based theories of language.

However, *Corpus* Linguistics has developed considerably in the last decades due to the great possibilities offered by the processing of natural language with computers. *Corpus* Linguistics is the study and analysis of data obtained from a *corpus*. A research library needs to retain material so as to provide future researchers with a rich *corpus* of source material. But he does stress the need to prove the *corpus delicti*.

All students are expected to be computer-literate, to have an interest in language, and to be aware of what *corpora* are. Learn more if you want to learn more about *corpora* and *corpus* linguistics you can use the links below.

*Corpus* Linguistics is now seen as the study of linguistic phenomena through large collections of machine-readable texts: *corpora*. The Departments of German (Bill Dodd) and Italian (Jacqueline Vicenzi) also have interests in *corpora*.

**Figure 2** Concordance output for *corpus*, best-first order
SkE offers a number of other features which were devised principally with the lexicographer in mind, but which nevertheless could make a language learners DDL experience richer and more fun than reliance on raw corpora. These include Word Sketches, one-page, automatic, corpus-derived summaries of a word’s grammatical and collocational behaviour; a distributional thesaurus, which shows which words commonly occur in the same context as a user-supplied keyword, and are likely to be near synonyms, and the sketch differences module, which shows how the collocational behaviour of two user-supplied keywords differs. All of these were described in some detail at the DAE conference two years ago, in Smith, Huang, Kilgarriff, & Chen (2007).

**Corpora in a Taiwan classroom**

Over the last two years, the author has been using corpora in Freshman English classes at two Taiwanese universities, one a private institution, the other public. On the whole, standards of motivation and attainment are expected to be higher at a public university, and that expectation has been borne out by impressionistic results. At the private university, I only ever used corpora in the classroom itself, finding that the meaning and collocational patterns of a word could be quite well illustrated by projecting a Word Sketch, for example, on the in-class screen.

In a similar way, I often project Google image search results, and definitions from online dictionaries, on the screen. I don’t think of Word Sketch display as any closer to DDL than these two procedures, however.

Even though response times on the SkE server are consistently quite fast, it does seem to take time to call up the displays, and it interferes with the pacing of the class. I also found it difficult to figure out the right amount of exposure to corpora for students: if too little, they will perhaps not be familiar enough with the corpus tools to learn much from the displays. If too much, there is a risk that the lesson will turn into a computer walkthrough and some students will lose interest.

At the public university, I have continued to make limited in-class use of Sketch Engine output. I have also given some homework assignments which required students to make use of corpora, asking them to produce word webs similar to those suggested in the reading skills textbook I use, Anderson (2008). Among the assignment types I have used, it is this that most closely resembles a DDL task.

![Word web from Anderson (2008)](image)
As seen from Figures 4 and 5, where students had to choose keywords from a BBC news story of their I modified the Anderson format, such that synonyms and antonyms were not required, but a translation was, as well as example sentences from the concordance. I hoped that students would find highly salient collocations from Sketch Engine word sketches, and then click on the concordance function to find sentences which included those collocations. Figure 4 was turned in by student who understood that part of the task, while Figure 5 contains example sentences which do not illustrate the collocations given.

1. Lessons about personal, social and health matters including sex and relationships will be compulsory in all England’s schools from ages five to 16.
2. Education is compulsory for children in most countries.
3. Whilst contents insurance isn’t compulsory, it is strongly advisable.
4. First, ASWs took responsibility for decisions diverting individuals from compulsory admission.

Figure 4 Student word web demonstrating understanding of collocation examples
In an earlier task, the students had been asked to draw similar word webs by hand and hand them in on paper; there was great variance among the students’ artistic skills. For this second task, an important feature was in fact the production of an acceptable computer graphic. Students were given no guidance on the file format to use, and had only the example from the textbook (Figure 3) as inspiration for their choice of layout. Thus, they were required to rely on their own initiative to complete the task. The task also afforded some training in computer text and graphics manipulation, which will serve them well in future university and professional life.

**The second semester: future plans for DDL**

In the upcoming semester, I shall be devoting more class time to DDL and corpus linguistics: about one out of two hours total contact time per week. At least two of the weekly classes will take the form of workshops, where each student will have a computer available, and they will be able to practise generating concordances and using Sketch Engine and other corpus query tools under fairly close supervision. At the first workshop, students will complete a task sheet which I have already used successfully with linguistics postgraduates; by the end of the semester, they will have to turn in a corpus-based project, either discussing language patterns that they find interesting, or comparing two or more corpora or corpus interfaces.

The reader may note some similarity between my planned approach and Content-based Instruction (CBI), defined Crandall and Tucker (1990:187) as “an approach to language instruction that integrates the presentation of topics or tasks from subject matter classes (e.g.,
math, social studies) within the context of teaching a second or foreign language.” In this case, the subject matter would be corpus linguistics. On another definition (Davies 2003), CBI means “learning about something rather than learning about language”. Clearly, learning about language is one of the goals of the plan. It has been argued by the above scholars and many others that learners are more likely to acquire language if the subject matter is of genuine interest to them; I believe that while Taiwanese learners are not drawn with enthusiasm to the memorization of vocabulary and grammatical patterns, but that they are interested in language itself, and will enjoy learning about words, collocation and meaning through the analysis of the authentic texts found in a corpus.

Bibliography