TESL Reporter 36,1 (2003), pp. 23-29

Use of Corpora in Language Learning

Jackie F. K. Lee
The Hong Kong Institute of Education

The Corpus

Within linguistics, a corpus is defined as a collection of texts, spoken and/or written, which is generally assumed to be representative of a given language or a variety of a language for linguistic analysis. With the advent of computers which make it possible to store, scan, and classify large masses of material, there has been a rapid expansion of corpus linguistics in the last four decades.

The turning point in the development of corpus study was the Survey of English Usage, which began at University College London in 1958 under the supervision of Randolph Quirk (Rundell and Stock, 1992). The aim of the Survey was to examine English usage, particularly grammar. By today's standard, the corpus was small. It consisted of one million words of running text, which were recorded manually and stored on index cards. Half of the texts collected were spoken and half were written English. Because the data were not originally stored computationally, the limitation of the Survey is that physical access was not easy.

The first machine-readable corpus was developed at Brown University in the United States. The Brown Corpus, completed in 1964, consisted of 500 written samples with a total of one million words of running text. Since the corpus was designed to be entirely synchronic, all of the texts chosen were first published in the United States in a single year (1961). The text categories chosen ranged from scientific writings and newspaper reportage to westerns and romances. In order to achieve a representative balance of the different genres, extracts of 2,000 words only were made from each selected sample. The Brown Corpus was thus carefully balanced and provided valuable information about word frequency and related statistical data.

In 1978, the Lancaster-Oslo/Bergen Corpus (LOB), a computer corpus of British written English designed to match the Brown Corpus, was completed at the University of Lancaster, with the assistance of Stig Johansson at Oslo and the Norwegian Computing Centre for Humanities at Bergen (Collins, 1987). Similar to its American counterpart, the LOB Corpus was composed of 500 written texts of about 2,000 words each. The year of publication (1961) and the sampling composition were the same as those of the Brown Corpus, though there were some inevitable differences in text selection.

Whereas the Brown Corpus and LOB were mainly used for exploration of linguistic patterns (Kjellmer, 1987; Meijs, 1988), nowadays, applied corpus linguistics is being developed with the aim to modify pedagogic materials. Corpora are being used to compile dictionaries and grammar books. COBUILD has made extensive use of the Bank of English, which amounted to 450 million words in January 2002, to write dictionaries, grammar and usage books to dispel the myths about the English language which were based on the intuitions of some "armchair" writers (Collins COBUILD). The widespread availability of computing facilities these days also enables teachers and learners to consult large collections of electronic texts on-line. They can search for word combinations, check word frequencies and see examples of how particular words are used. A good deal of the corpus-based work (e.g., Butler, 1991; Johns, 1991; Johns, 1994) focuses on the production of teaching and testing materials so as to develop students' inductive learning strategies and their learning autonomy.

An effective learning approach as suggested by a number of researchers (e.g., Littlejohn, 1985; Cotterall, 1995) is to increase learner autonomy, which is characterised by students' taking significant responsibility for their own learning. Littlejohn (1985) comments that an outcome of promoting learner autonomy may be an increase in enthusiasm for learning. Using corpora in teaching is a way to increase learner autonomy because instead of telling students why a certain structure is unacceptable, the teacher can abandon the role of expert and say, "Let's read the corpus examples and find out the answer together." By searching corpora, students can do their own research and discover the usages of contemporary English at their own pace.

In this paper, I would like to demonstrate how to use corpora for error correction and vocabulary building in a language classroom. There are some publically available corpora which anyone can use for free. Users only need access to the internet to be able to perform corpus searches. The World Wide Web Access to Corpora Project (W3-Corpora), which aims to provide free access to existing linguistic corpora under the Gutenberg Project via the web to students and researchers in linguistics and related disciplines, is run by the Department of Language and Linguistics at the University of Essex. The corpus data can be used in a number of ways, including the following:

a. Comparing two similar words
b. Building vocabulary
c. Examining debatable usages

The following are some illustrations showing how to use the search engine.

Comparing Two Similar Words

Previous studies (e.g., Lockhart, 1996; Sengupta and Falvey, 1998) reveal that teachers are concerned with grammar and mechanics in their rating of students' essays. The traditional method used by many teachers is to mark every error students make. This method, however, has long been criticised as "ineffective" since marking every error is time-consuming for teachers and humiliating for students, who find the "bloody" corrections discouraging. As a consequence, writing and marking compositions are often regarded as the most unpopular tasks for students and teachers respectively (Chen, 1996/1997, p. 29). What is worse, despite the hard work of teachers who work as "marking slaves" or "marking robots", many students seem not to be learning from error correction since they keep making the same mistakes. A more effective learning approach is to increase learner autonomy by incorporating the idea of research in the classroom and asking students to find out the correct usage for themselves. The following example shows how to let students realize the difference between the two easily confused adjectives: *live* and *alive*.

In the *Longman Dictionary of Contemporary English* (1995), the following definitions are given for *live* and *alive*.

Live: not dead; living

Alive: still living and not dead

It seems that the two words *live* and *alive* mean the same. By looking at the KWIC (keyword-in-context) concordance data,[1] however, one can find that the two words cannot be used interchangeably. Take a look at the following examples:

Live:

a person, a real <u>live</u> person, who would be fond of me

He had never seen any <u>live</u> boys, but he had seen pictures of them

Still there must be many <u>live</u> creatures in the world besides caterpillars.

Alive:

They have not the joy of being <u>alive</u> which is a kind of earnest

the criminals will still be <u>alive</u>; but when he cuts off their

the weak anthrax virus would be <u>alive</u> in the anthrax-infected field,

reason, I am sure that you remain <u>alive</u>: it is impossible that you should

The corpus data show that *live* is an attributive adjective used before a noun, whereas *alive* is a predicative adjective, which is used after stative verbs such as *be* and *remain.* By studying the output from the searches, students can see how similar words are used in real context, and work out the correct usages by themselves.

Vocabulary Building

Doing corpus searches also helps students to develop vocabulary by learning the central words and how words are formed with prefixes and suffixes. Take an example of the string "separa". At the time of writing this paper, the match frequency found for this string was 2,027 in all the corpora under the Gutenberg Project. To examine what those words are, one can go to the Lexical Frequency page, which provides the following information:

separate	736	*inseparably*	23
separated	554	*separations*	17
separation	293	*separators*	10
separately	111	*separator*	5
inseparable	86	*inseparables*	4
separating	68	*separable*	3
separates	36		

The Lexical Frequency table above shows that the words given are all semantically related to "separate". By looking at the frequencies of each word, we find which word is the central one, which might be useful, for example, when deciding which word a foreign learner of English might want to learn first. We see that the most frequent words here are separate and separated, with 1,290 instances, which is about 64% of the total. These statistics suggest that *separate* and *separated* are the central items to learn. The second most frequent is the noun *separation,* with 293 occurrences, which is approximately 14% of the total, and the third most frequent is the adverb *separately*, with 111 occurrences (5%).

By studying the frequency list, we also see some examples of word-formations including the string "separa". We can see how some suffixes can be used to mark different parts of speech, for example *-ion* and the less common *-ble* for nouns, and the inflectional endings *-s* marking the plural form and *-ed* marking the past tense and past participle. The common prefix *in-* is found with the string, forming words such as *inseparable* and *inseparably.* For advanced learners, activities can be designed to let students guess the meanings of new words in sentence concordances. For example, by

studying the following sentences taken from the corpora in the Gutenberg Project, students can be asked to work out the meanings of each of the words underlined.

- *She was facing courageously the three inseparables, Hagar, Viney, and Lucy, squatted at the top of the steps, and she was speaking her mind rapidly and angrily.*

- *Two—this exactly like the first, except that those inseparables, Hagar, Viney, and Lucy, whom Miss Georgie had inelegantly dubbed "the Three Greases", appeared, silent, blanket-enshrouded, and perspiring, at the office door in mid-afternoon.*

- *A ship could not be spared to convey him to England; he therefore travelled through Germany to Hamburgh, in company with his inseparable friends, Sir William and Lady Hamilton.*

Examining Debatable Usages

Since corpus data reflect authentic usage, some linguistic myths and distortions that originated from the intuitions of some 18th-century purists such as Samuel Johnson, Bishop Lowth and Lindley Murray, and perpetuated from generation to generation via dictionaries and grammars can be refuted. Take the Latin plural *data* as an example. Some concordances for this word are given below:

Few of these data were ever actually used, however;

almost all the numerical data are largely guess work. It will

so many of the data, whether for hope or fear, were

The observational data are not yet sufficiently accurate

where a great deal of data may be found. Edison says with

free flight, to get as much data as possible regarding the conditions

gave a good deal of useful data for the construction of later vessels

the Details. With this data as a guide it should be

Ross-Smith's flight valuable data was gained in respect of reliability

Although the legitimacy of treating the foreign plural *data* as a singular noun is disputed, it does exist as a singular noun in actual practice. In the corpora studied, *data* is used in two constructions: (1) as a plural noun with a plural verb and certain plural modifiers (e.g., *these, many, few*); (2) as a mass noun, taking a singular verb and singular modifiers (e.g., *this, much, a great deal of*). By examining the concordances, students can realise that the prescriptivists' teaching that *data* is a foreign plural and must precede a plural verb is without empirical foundation in real English.

Conclusion

Apart from the three uses of corpora suggested in this article, Stevens (1991) and Johns (1994) recommend other ways of learning vocabulary and grammar by means of concordance-based inductive learning strategies. Experience in using concordance data has indicated that it is a powerful stimulus to student enquiry in a learner-centred classroom and it helps students become better language learners and researchers outside school. An appropriate role for the teacher is no longer an authority figure but a research organiser who provides a stimulating context in which the learner can develop strategies for self-discovery. In Johns' (1991, p. 2) words, "we [teachers] simply provide the evidence needed to answer the learner's questions, and rely on the learner's intelligence to find answers."

Notes

1. A computer-based concordance program will find all the instances of a particular item (morpheme, word, expression) in the texts selected. KWIC is a way of displaying concordances; that is, it prints the item in the middle of the screen, with a fixed number of characters of context to the left and to the right.

References

Butler, J. (1991). Cloze procedures and concordances: The advantages of discourse level authenticity in testing expectancy grammar. *System, 19* (1/2), 29-38.

Chen, J. (1996/1997). Issues in the teaching of writing in Hong Kong. *Hong Kong Polytechnic University Working Papers in ELT and Applied Linguistics, 2*, 29-38.

Collins COBUILD. The Bank of English. Retrieved December 6, 2002 from http://titania.cobuild.collins.co.uk/boe_info.html.

Collins, P. (1987). Computer corpora in English language research: a critical survey. *Australian Review of Applied Linguistics, 10*, 1-19.

Cotterall, S. (1995). Developing a course strategy for learner autonomy. *ELT Journal, 49*, 219-227.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. *English Language Research Journal, 4*, 1-16.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on Pedagogical Grammar*. Cambridge: Cambridge University Press.

Kjellmer, G. (1987). Aspects of English collocation. In W. Meijs (Ed.), *Corpus Linguistics and Beyond*. Amsterdam: Rodopi.

Littlejohn, A. (1985). Learner choice in language study. *ELT Journal, 39*, 253-261.

Lockhart, C. (1996). Teachers' beliefs about writing in Hong Kong secondary schools. *Perspectives: Working Papers, 8*(2), 45-79.

Meijs, W. (1988). *All But* and *If Not* in Brown and LOB. In M. Kyto et al. (Eds). , *Corpus Linguistics, Hard and Soft*. Amsterdam: Rodopi.

Rundell, M., & Stock, P. (1992). The corpus revolution. *English Today, 30*, 9-14.

Sengupta, S., & Falvey, P. (1998). The role of the teaching context in Hong Kong English teachers' perceptions of L2 writing pedagogy. *Evaluation and Research in Education, 12*(2), 72- 95.

Stevens, V. (1991). Concordance-based vocabulary exercises: A viable alternative to gap-filling. *English Language Research Journal, 4*, 47-61.

University of Essex. Welcome to the W3-Corpora site! http://clwww.essex.ac.uk/w3c/

About the Author

Jackie F. K. Lee is a lecturer in the Department of English at the Hong Kong Institute of Education, where she is responsible for teaching language courses to pre-service and in-service student teachers.