Kwansei Gakuin University
Humanities Review
Vol. 18, 2013
Nishinomiya, Japan

Corpus and Its Role in Teaching

Daniel PARSONS*

Introduction

One of the goals of language teaching is to raise students' awareness of and provide practice in natural language use in a variety of registers. This implies that language learning tasks and learning materials should give consideration to the source of their language content. One way a teacher can make principled decision about tasks and materials is through the use of corpora. In this article I will examine what corpora are and show how they can be used to develop authentic language content for the classroom. I will review the corpus of science textbooks currently being built at the Living Language Laboratory in the Faculty of Science and Technology at Kwansei Gakuin University. I will further examine ways that this corpus can be used for data driven learning in the classroom.

Definition, Compilation and Use of Corpora in Research

A corpus is defined as a collection of naturally occurring language texts which are carefully chosen to represent a state, variety or a set of registers of a language, and compiled for the purposes of analysis (Sinclair, 1991; Biber et al. 1998; Gries, 2009; Flowerdew, 2012; McEnery and Hardie, 2012). However, there is sometimes contention over what counts as a corpus.

The idea of representativeness is part of what distinguishes a corpus from a simple database of texts. The texts chosen for a corpus are sampled based on specific criteria set in place before sampling begins. These criteria may be informed by research questions, and Sinclair (2004 a) lists ten basic principles for design. Biber *et al.* (1998), on the other hand, take a statistically informed approach to sampling, by showing that samples which are too small may over-represent or

* Instructor of English as a Foreign Language, School of Science and Technology, Kwansei Gakuin University

under-represent particular linguistic features of a register. Biber (1988) further argues that sample size should be determined by the internal variation of a register. Of course, it is still necessary to choose the wider texts which need to be sampled from, which highlights the issue of balance within the corpus. Once again, clear research questions are essential for maintaining balance (Reppen, 2010).

Early corpora such as the Brown corpus and the Lancaster-Oslo/Bergen (LOB) corpus, were designed as American English and British English counterparts, using printed sources (Johansson *et al.* 1978). While they both stand at a million words, they claim representativeness of a general English due to stratified sampling of relevant categories and sub-categories of texts. The sample sizes are small at 2000 words from each text. In contrast, The British National Corpus (BNC), includes sample sizes of up to 40,000 words. This choice is based on the notion that internal variation may not be uniformly distributed across texts (Nelson, 2010). The Corpus of Contemporary American English (COCA) is divided equally across spoken, fiction, popular magazines, newspapers and academic journals and is considered as a monitor corpus for change in the English language over time (Davies, 2011).

Concordance software tools provide the means to explore corpora. Well known tools include Wordsmith Tools and Antconc, and both tools provide a concordance lines generation function. Sorting functions allow easy identification of commonly occurring associations of words with other words (collocation) or grammar patterns (colligation). Investigations into collocation and colligation have helped revise theories of language and natural language use. Indeed, Hoey's theory of lexical priming claims to be a new theory of language based on the psychological concept of priming and evidence of language in action from corpora (2005).

Another approach to the analysis of large corpora is the multi-dimensional analysis carried out by Biber (1988). He employed a statistical method known as factor analysis to the Lancaster-Oslo/Bergen corpus and the London-Lund corpus of spoken English. This method calculates how particular linguistic features co-vary and consequently classifies the particular features into different varieties of text. In this way, Biber was able to describe and classify the variation of English in those corpora in terms of how the linguistic features varied. Recently, Biber's research has been expanded to include investigations into the variation of linguistic features in university spoken and written discourses (Biber, 2006).

The above approaches demonstrate how corpora and the tools used to explore corpora have been used in research to shed light on theories of language and provide better descriptions of authentic language use. This has implications for various arenas in education, and I will now turn my attention to the use of corpora in education.

Corpora in Education

Corpora have been used in a number of educational arenas. These include data driven learning in the classroom, using corpora in teacher education, creating teaching materials and writing dictionaries for language learning, and use by teachers for curriculum development and research (Sinclair, 2004 b; Aston, *et al.,* 2004).

Data driven learning brings authentic language directly into the classroom (Gilquin and Granger, 2010; Higgins and Johns, 1984). It is a form of discovery learning which encourages students to use concordance tools and observation techniques and form hypotheses about authentic language in much the same way as the lexicographer might. Data driven learning focuses the students on inductive learning and gaining insight about language use, which is in contrast to habit forming exercises (Johansson, 2009). Activities in the classroom may be teacher led or student led. Teacher led activities might involve the teacher preparing a series of gap fill activities based on concordance data, which, it could be argued, focus the students on forming habits. On the other hand, student led activities might involve them forming questions about language use and querying a corpus to find the answer. In fact, Gilquin and Granger (2010) suggest the use of a sleeper corpus in the classroom. A sleeper corpus is one which is accessible when students have a question about language use, in much the same way that a dictionary is queried.

There have been a number of studies into the effectiveness of data driven learning on student learning, and the majority of these studies have been with university students (Flowerdew, 2012: 203). Chujo *et al.* (2012) demonstrated that both paper based and computer based concordance exercises in the classroom are equally effective in helping university students learn about noun phrases. However, they cautioned that students may be spending time studying about a language as opposed to spending time using the language. Another study by Tian (2004) showed that groups using data driven learning made significant gains in understanding word usage, but there were no significant gains in grammar compared with a control group not using data-driven learning. Sun and Wang (2003) showed that high school students using an inductive approach with concordancers improved significantly more when studying collocations than students using a deductive approach.

Researchers also point out the impediments to data driven learning. One of these impediments is the availability of suitable corpora and the tools to access them. (Cheng, 2010). Another is the lack of training teachers have received in using corpora and concordance tools, making it difficult for teachers to feel confident using the data driven learning approach. Boulton (2010) discusses the loss of power which some teachers may experience due to a sense that their position of authority in the classroom may have been undermined. Further issues include the complexity of the data which students face (Chujo, 2012). Students may be able to form questions about language, but they may lack confidence in their own concordance searches, or feel uncertain about the insights they gain. Without proper support and training, both teachers and students are potentially left bewildered by the tools and complexity of the data in the classroom.

The Corpus of Science Textbooks

Compiling the Corpus of Science Textbooks began approximately two years ago, and is currently being reviewed at the Living Language Laboratory in the Faculty of Science and Technology at Kwansei Gakuin University. The purpose of the corpus is to provide data on scientific language use for the students in the faculty and to further assist researchers in understanding scientific writing in English. It is seen that the corpus will help students learn the scientific writing skills necessary to become successfully published in their future careers.

With this in mind, the corpus was compiled to represent the main areas of learning within the faculty, including life science, chemistry, physics, informatics and mathematics texts. Samples were taken from textbooks which were considered to represent topics and fields which undergraduate and graduate students would study should they have been studying in English. Stratified samples of 2000 words each were compiled and loaded into a file ready for use with a concordance tool. The corpus is still under construction and stands at approximately one million words.

Many of the files were originally photocopied from the textbooks and scanned using an optical scanner to create electronic versions of the documents. However, this led to many copying and scanning errors in the text, and it has in fact been claimed that optical scanners can reduce the accuracy by as much as 40% (Nelson, 2010). In the Corpus of Science Textbooks, for example, the two words “of the”, which occurs 10,674 times have become “ofthe” 540 times out of the one million words of the corpus. Furthermore, a simple computer program was written to cross reference each occurrence of a word with a dictionary of 320,000 entries. In a 12,000 words sample of the corpus, 555 errors were found. However, it should be noted that a small number of these errors were due to mathematical equations, which were not contained in the dictionary.

To confound the errors, an analysis of sample titles and brief scanning of selected files has revealed that the corpus is not balanced. Currently, 67% of the corpus is represented by life science texts, with the other subjects represented equally across the remaining 33%. This was not surprising, since one of the original purposes was to trial data driven learning with life science students. Furthermore, we recognized that the corpus is still being compiled. Nonetheless, this result demonstrates the need for keeping metadata.

It has been concluded that the Corpus of Science Textbooks needs an overhaul. This includes using files originally in electronic format to reduce the impact of the errors described above. In addition, it will be useful for research purposes to keep metadata about the files, and document sampling techniques and choices along the way. In particular, Burnard (2005) suggests that keeping metadata is good practice in building corpora.

In spite of the potentially huge impact of errors, a number of analysis techniques have been applied to the corpus for the purposes of demonstrating the potential this corpus has for being a valuable resource. This can be justified from a number of perspectives. Firstly, these analysis techniques are not being applied in an attempt to provide any conclusive evidence about language or teaching. In fact, they are being applied simply as a demonstration of the tools which are available, and to highlight more specifically what new tools might potentially be developed from now on. Secondly, when reading concordance lines from the corpus, most of the errors are translatable by human readers. This implies that there is still a lot of use to be gained from the corpus, especially for teachers who wish to use the corpus to develop materials for the classroom.

While nothing conclusive can be validated from the Corpus of Science Textbooks in its current state, it is still useful to demonstrate the tools and techniques available for future analyses on a revised, more accurate, well documented corpus.

Profiling the Corpus of Science Textbooks

Early uses of corpora in language teaching led to word lists based on frequency of use, which helped provide useful information about language content for course materials (Walsh, 2010). The top 20 words from the Corpus of Science Textbooks is shown in Table 1, and demonstrates how the most common words are not content words, but words that have grammatical functions. There is very little difference in this respect between the Corpus of Science Textbooks and any other corpora, as can be seen by the top 20 words in the Corpus of Contemporary American English (COCA), also in Table 1.

This is not particularly useful information in light of the original purposes for constructing the Corpus of Science Textbooks. In order to get an overview of the features of the corpus, a technique called frequency profiling, demonstrated by Rayson and Garside (2000) was implemented.

Table 1 The top 20 words in the Corpus of Science Texts and COCA

Rank	Science Corpus	COCA
1	the	the
2	of	be
3	and	and
4	a	of
5	in	a
6	to	in
7	is	to
8	that	have
9	are	to
10	The	it
11	by	I
12	DNA	that
13	for	for
14	as	you
15	be	he
16	with	with
17	or	on
18	from	do
19	on	say
20	which	this

Frequency profiling calculates a log likelihood score for each word in the corpus, which is a measure of the significance of the word occurring in our corpus. In corpus linguistics terms, frequency profiling is also known as a keyness test, and is a way of searching for key words in the target corpus. Either a comparison corpus or data about the frequency of words in a comparison corpus is required to calculate the log likelihood score. This is because the significance has to be calculated in relation to a standard population.

We chose to perform the frequency profile of our corpus in comparison to the top 5000 words in COCA. In other words, COCA was the standard population. COCA is large, well balanced and representative, implying that a comparison of the Corpus of Science Textbooks is against a general English. Furthermore, by restricting ourselves to the top 5000 words in COCA we are filtering out field specific technical words within our corpus. This is useful for revealing key words which may also be common daily words and can therefore have more time focused on them in undergraduate courses.

A computer program based on the equation (1) for calculating log likelihood scores was written. Table 2 shows the top 20 most significant words within the Corpus of Science Textbooks and their log likelihood scores.

$$(1)\ \mathrm{G}^2 = \sum O_i \ln \frac{O_i}{E_i}$$

Table 2 Most significant words in the Science Textbooks Corpus

Log Likelihood Score	Significant Words	Log Likelihood Score	Significant Words
29102	be	7610	have
22691	protein	7004	his
22663	genetic	6919	acid
13812	you	6729	molecule
11391	sequence	6447	the
10601	cell	5971	known
8615	say	5995	bacteria
8470	do	5585	found
8359	used	4922	magnetic
8153	of	4891	go

It is the prominence of life science vocabulary in the word list in Table 2 which adds weight to evidence that the corpus is over-represented by life science texts. The word "you" is calculated to be the fourth most significant word. This implies that writers in science textbooks may tend to address their readers directly. The words "of" and "the" are also significant, which may imply the heavy use of nominalized forms. This is an aspect of scientific writing which has been described by Halliday (2004). Halliday's work also helps to provide an explanation for the significance of "have", which could be attributed to explication of technical taxonomies. In other words, scientists often describe the features and characteristics of the things they are studying or measuring. The significance of "be" could be attributed to a tendency to use passive voice, possibly in conjunction with modal verbs, in scientific writing.

Ultimately, these insights are useful only for the demonstration of the investigative techniques used to provide the insights, due to the large number of errors in the corpus. Furthermore, lessons have been learned on good practices for compiling a corpus.

Current and Future Developments

The Corpus of Science Textbooks is being used in two main areas of teaching. The first is the development of course materials for first year undergraduates, and the second is data driven learning in the classroom. These uses are still under development and so a brief report is provided here.

The development of course materials is being facilitated by examining lexical bundles in the Corpus of Science Textbooks. Biber (2006) characterizes lexical bundles in the following way: their identification is purely frequency based, although the cutoff point may be arbitrary; they are not idiomatic units; they are not complete structural units but rather bridge two structural units. Examples of the lexical bundles are “a wide range of”, “as a result of”, “can be expressed in”, “for the case of”. Like Biber (2006), 4-word lexical bundles were chosen to limit the scope and return a manageable data set. In all 115 significant 4-word lexical bundles were isolated from the corpus.

Concordance lines for each lexical bundle have been generated from the corpus and are currently being analyzed using a six step approach, outlined by Sinclair (2003): initiate, interpret, consolidate, report, recycle, report. The “report” steps involve reaching and refining a hypothesis about how the lexical bundle bridges the structural units to the left and right of the bundle.

The structural units either side of the lexical bundle can be classified in terms of Halliday’s description of the language of science (2004). Combining the exploration of lexical bundles with Halliday’s classification demonstrates a technique for eliciting the context and corresponding linguistic features of the scientific method from the corpus, and the analysis is ongoing.

The second application of the Corpus of Science Textbooks is data driven learning in the classroom. The goal is to introduce second year students to the language of science through awareness raising activities. These activities involve looking at everyday language, such as “have” and “take”, considering the frequency of these words, and examining concordance lines to see the types of words that co-occur. Students begin to notice that technical vocabulary co-occurs with everyday language.

Future developments for the Corpus of Science Texts involve building a web based concordance tool for the corpus, applying Biber’s (1988) multi-dimensional analysis to the corpus, and developing new corpora to meet curriculum demands. A web based concordance tool will help solve the problems associated with easy access to corpus files and provide a user-friendly interface for ease of use in the classroom, as well as address the barriers to use perceived by teachers. Biber’s multi-dimensional analysis will enhance materials creation from the corpus in that it has the potential to reveal the distribution of linguistic features across various registers in the corpus. Finally, the compilation of corpora for transcribed presentations and learner essays may help meet the demands of the curriculum for first and second year students.

References

Aston, G., Bernardini, S. and Stewart, D. (eds.) (2004). *Corpora and Language Learners.* Amsterdam: John Benjamins B.V.

Biber, D. (1988). *Variation Across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (2006). *University Language: a Corpus-Based Study of Spoken and Written Registers.* Amsterdam: John Benjamins B.V.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Boulton, A. (2010). Data-Driven Learning: On Paper, In Practice. In T. Harris and M. Moreno Jaen (eds.) *Corpus Linguistics in Language Teaching.* Bern: Peter Lang, (pp.17–52).

Burnard, L. (2005). Metadata for Corpus Work. In M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice,* (pp.35–52). Oxford Text Archive. Available at: http://ahds.ac.uk/creating/guides/linguistic–corpora/index.htm.

Cheng, W. (2010). Corpora and Language Teaching. In A. O'Keefe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics,* (pp.319–32) London and New York: Routledge.

Chujo, K., Anthony, L., Oghigian, K. and Uchibori, A. (2012). Paper-based, computer-based, and combined data-driven learning using a web-based concordancer. *Language Education in Asia* 3(2), 132–45.

Davies, M. (2011). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25: 447–65.

Flowerdew, L. (2012). *Corpora and Language Education.* Palgrave Macmillan.

Gilquin, G. and Granger, S. (2010). How Can Data-Driven Learning Be Used in Language Teaching? In A. O'Keefe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics,* (pp.359–70) London and New York: Routledge.

Gries, Th. S. (2009). Quantitative Corpus Linguistics with R: A Practical Introduction. Routledge.

Gu, Q. (2010). Variations in beliefs and practices: teaching English in cross-cultural contexts. *Language and Intercultural Communication,* 10(1), 32–53.

Halliday, M. A. K. (2004). The Language of Science. London: Continuum.

Higgins, J. and Johns, T. (1984). Computers in Language Learning. Longman ELT.

Hoey, M. (2005). Lexical Priming: A New Theory of Words and Language. Oxford and New York: Routledge.

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. In K. Aijmer (ed.), *Corpora and Language Teaching,* (pp.33–44). Amsterdam: John Benjamins.

Johansson, S., Leech, G. N. and Goodluck, H. (1978). Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers, Department of English, University of Oslo. Available at: khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM

Kilgariff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics,* 29(3): 333–47.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics.* Cambridge: Cambridge University Press.

Nelson, M. (2010). Building a Written Corpus: What are the basics?" In A. O'Keefe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics,* (pp.53–65) London and New York: Routledge.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings* *of the Workshop on Comparing Corpora,* 1–6.

Reppen, R. (2010). Building a Corpus: What Are the Key Considerations? In A. O'Keefe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics,* (pp.31–7) London and New York: Routledge.

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

Sinclair, J. (2003) Reading Concordances: An Introduction. Pearson Education Limited

Sinclair, J. McH. (2004 a). Corpus and Text – Basic Principles. In M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice,* (pp.1–21). Oxford Text Archive. retrieved from: http://ahds.ac.uk/creating/guides/linguistic–corpora/index.htm.

Sinclair, J. McH. (ed.) (2004 b). How to Use Corpora in Language Teaching. Amsterdam: John Benjamins B.V.

Sun, Y. and Wang, L. (2003). Concordancers in the EFL classroom: cognitive approaches and collocation difficulty. *Computer Assisted Language Learning* 16(1), 83–94.

Tian, S. (2004). Data-driven learning: do learning tasks and proficiency make a difference? *Proceedings of the 9th Conference of Pan-Pacific Association of Applied Linguistics,* (pp.360–371).

Walsh, S. (2010). What Features of Spoken and Written Corpora Can Be Exploited in Creating Language Teaching Materials and Syllabuses? In A. O'Keefe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics,* (pp.333–44) London and New York: Routledge.

Woods, D. (1998). *Teacher Cognition in Language Teaching.* Cambridge: Cambridge University Press.