Second Language Corpus Studies

R Reppen, Northern Arizona University, AZ, USA

Corpus linguistics has contributed to several areas of applied linguistics. In addition to core contributions in the areas of lexicography and grammar, corpus linguistics has also provided insights into the areas of register variation (e.g., spoken versus written language, across academic disciplines, stylistic variation), language change over time using historical or diachronic corpora, studies of gender differences, and, more recently, the area of second language studies (Reppen, 2001; Granger *et al.,* 2002; Granger, 2003). By using large, principled collections of naturally occurring language, corpus linguistics can accurately explore and describe linguistic characteristics and patterns associated with language use in different contexts (e.g., talking among friends, giving a formal speech, writing a friend, writing a research paper), across different speakers, and how language varies regionally. These descriptions can then be used to accurately describe patterns of variation and can also be used to inform pedagogy for second-language learners.

Corpora consist of large collections of spoken and/or written texts, are typically stored on computers, and are often grammatically annotated and/or marked up for certain text features (e.g., Biber *et al.,* 1998; Meyer, 2002; Reppen and Simpson, 2002). Because of their large size, often well over a million words, it is essential to have tools that allow users to effectively and efficiently search the corpora. There is a variety of computer programs available, ranging from concordancing software (e.g., MonoConc, WordSmith) that can generate word lists and identify specific words or combinations of words, to sophisticated programs that can perform comparisons that track features across a range of texts. Most users will interface with corpora through the use of concordancing software, most of which can be used with either an unannotated corpus or one that is annotated for grammatical or text features. A concordancing program allows users to generate word frequency lists, see target words in context, look for expressions, and also search for particular combinations, such as verb plus preposition, or what verbs frequently occur with complement clauses (if using a grammatically tagged corpus). Concordancing programs are useful tools for both language researchers and language students and teachers.

The development of learner corpora (e.g., Granger, 2003) has enhanced the ability of corpus linguistics to make contributions to the areas of second language acquisition and language pedagogy (Partington, 1998; Conrad, 1999, 2003; Burnard and McEnery, 2000; Biber and Reppen, 2002; Granger *et al.,* 2002; Meunier, 2002; Granger, 2003). Rather than relying on information from case studies or single examples, researchers are able to use corpora from second language learners to describe and explore the linguistic patterns of second language learners. As more second language corpora are developed, they will become powerful resources for cross-linguistic comparisons of different first language speakers producing different target languages. Researchers will be able to accurately describe the linguistic patterns of second language learners. This information will help shape teaching and language pedagogy to more accurately address the needs of second language learners.

In spite of second language or learner corpora only recently beginning to take hold, information from corpus linguistic research on English corpora is being used to inform second language instruction and assessment. Rather than relying on native speaker intuitions about how language is used, material developers and language teachers are now able to base pedagogical decisions on information from corpus-based research. Using information from detailed corpus studies based primarily on native English speakers (e.g., Biber *et al.*, 1999, 2004), teachers and material developers are able to produce lessons and instructional materials that target the needs and goals of the language learners. These corpora based on native English speakers contribute valuable information to the areas of second language research and pedagogy in two ways. First, they allow teachers and researchers to identify patterns of language use while also providing real language in context for second language learners to interact with in addition to textbooks and other course materials. Second, the native speaker corpora provide a starting point for cross-linguistic comparisons between the learner corpora and the native speaker corpora. These types of analysis will provide valuable insights into areas of similarities and differences for a variety of language learners (e.g., many different first languages) across a range of different contexts.

In addition to teachers using corpora to inform their teaching decisions, language teachers are also bringing corpora into the classroom and learners are interacting with corpora to explore questions of language use (Gavioli, 1997, 2001). For example, learners in an English for academic purposes (EAP) class can use a corpus to examine how certain phrases are used or to learn specialized vocabulary in their area of study. The MICASE (Michigan corpus of academic spoken English) site has over a million

words of academic spoken language. This corpus, and the companion concordancing program available online, represents a range of academic disciplines and academic settings (e.g., lectures, advising sessions, class discussions). Lessons and activities based on the MICASE corpus can be of great use to advanced learners of English preparing to study at English-speaking universities. Another valuable EAP resource is the research done on the T2K-SWAL corpus – a corpus of over 2 million words of academic spoken and written academic English (Biber *et al.*, 2004). Although the T2K-SWAL corpus is not available for general use as is the MICASE corpus, research based on the corpus is available and provides insights as to the linguistic patterns in spoken and written academic English. In addition to these two corpora, which are valuable resources for academic English, there are also other available corpora that cover a wide range of aspects of English (e.g., American National Corpus, British National Corpus, ICAME). Involving learners in exploring language through the use of corpora in the classroom can serve as a strong motivator and also help promote autonomous language learning.

The role of corpus linguistics in the area of second language studies is just gaining momentum. The widespread availability and use of computers in language classrooms and increasing availability of corpora should serve as a catalyst for the development of additional tools and for corpora that can be used for both research and pedagogical purposes. With the development of more second language corpora, both spoken and written, the fields of second language studies and language pedagogy will change significantly over the next decade. The types of analysis and the insights gained as to cross-linguistic variation in different contexts will help to shape the areas of language research and pedagogy.

Bibliography

Biber D & Reppen R (2002). 'What does frequency have to do with grammar teaching?' *Studies in Second Language Acquisition 24,* 199–208.

Biber D, Conrad S & Reppen R (1998). *Corpus linguistics: exploring language structure and use*. Cambridge: Cambridge University Press.

Biber D, Johansson S, Leech G, Conrad S & Finegan E (1999). *The Longman grammar of spoken and written English*. Harlow, Essex: Pearson Education.

Biber D, Conrad S, Reppen R, Byrd P, Helt M, Clark V, Cortes V, Csomay E & Urzua A (2004). *Representing language use in the university: analysis of the TOEFL 2000 spoken and written academic language corpus*. MS #25. Princeton, NJ: ETS.

Burnard L & McEnery T (eds.) (2000). *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang.

Conrad S (1999). 'The importance of corpus-based research for language teachers.' *System 27,* 1–18.

Conrad S (ed.) (2003). *TESOL Quarterly 37,* 3.

Coxhead A (2000). 'A new academic word list.' *TESOL Quarterly 34,* 213–238.

Donley K & Reppen R (2001). 'Using corpus tools to highlight academic vocabulary in SCLT.' *TESOL Journal 10,* 7–12.

Gavioli L (1997). 'Exploring texts through the concordancer: guiding the learner.' In Wichmann A, Fligelstone S, McEnery T & Knowles G (eds.) *Teaching and language corpora*. London: Longman. 83–99.

Gavioli L (2001). 'The learner as researchers: introducing corpus concordancing in the classroom.' In Aston G (ed.) *Learning with corpora*. Houston, TX: Athelstan. 108–137.

Granger S (2003). 'The International Corpus of Learner English: a new resources for foreign language learning and teaching and second language acquisition research.' *TESOL Quarterly 37,* 538–546.

Granger S, Hung J & Petch-Tyson S (eds.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Meunier F (2002). 'The pedagogical value of native and learner corpora in EFL grammar teaching.' In Granger S, Hung J & Petch-Tyson S (eds.) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins. 121–141.

Meyer C (2002). *English corpus linguistics*. Cambridge: Cambridge University Press.

Partington A (1998). *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam: John Benjamins.

Reppen R (2001). 'Elementary student writing development: Corpus-based perspectives.' In Simpson R & Swales J (eds.) *Corpus linguistics in North America: selections from the 1999 Symposium*. Ann Arbor: University of Michigan Press. 211–225.

Reppen R & Simpson R (2002). 'Corpus linguistics.' In Schmitt N (ed.) *An introduction to applied linguistics*. London: Arnold. 92–111.

Relevant Websites

http://americannationalcorpus.org – American National Corpus (ANC). The first release of 11.5 million words is available. The site has samples of the corpus format and links to papers related to the ANC project.

http://info.ox.ac.uk – British National Corpus (BNC). 100-million-word multiregister corpus of spoken and written British English. The site also has links to many corpus-related resources.

http://www.hit.uib.no/icame.html – ICAME: International computer archive of modern and medieval English.

http://www.hti.umich.edu – MICASE: Michigan corpus of academic spoken English.

http://www.athel.com – MonoConc.

http://oup.com – WordSmith.