100 CAMBRIDGE ENGLISH CENTENARY 1913–2013
Cambridge English Centenary
Symposium on Speaking Assessment

Corpus evidence and the lexicogrammar of speaking

Ute Römer, Georgia State University

The past few decades have witnessed a massive increase in corpus research activity in a range of linguistic subfields, including strands within Applied Linguistics. Corpora are increasingly accepted as powerful tools that help us gain insights into language structure and use, and help inform language teaching and testing practice (see, for instance, Flowerdew 2012, Hawkins and Filipović 2012, Reppen 2010, and Römer 2011). This paper discusses the importance of considering corpus evidence in highlighting central aspects of spoken language and addresses the question "How can corpus tools and techniques help us shed light on the concept of speaking?"

Since spoken language is not a uniform phenomenon but varies considerably depending on the context of use, the paper does not attempt to describe speech 'in general'. Instead, it focuses on one particular, more specialized type of language: spoken English produced in a US research university setting. This type of language is captured in MICASE, the Michigan Corpus of Academic Spoken English (Simpson, Briggs, Ovens and Swales 2002), a collection of 152 transcripts and 1.8 million words, based on 200 hours of recordings of speech events from across the University of Michigan in Ann Arbor.

The paper starts out with a brief analysis of frequency word and keyword lists of academic speaking (compared to academic writing), including observations on Zipfian profiles (Zipf 1935), and then focuses on phraseological items (variably referred to as n-grams, formulaic sequences, lexical bundles, clusters, etc.) that are particularly common in speaking and carry important discourse functions. Software packages for corpus access and analysis are used to extract lists of contiguous word sequences (n-grams, e.g. _you know_, _a lot of_) and non-contiguous word sequences (phrase-frames, e.g. _a * of_, _I don't * so_) of different lengths from MICASE. The resulting lists are filtered manually for items that play a central role in academic speech and appear to have a particularly high communicative value.

The final section of the paper reviews rating scales of a selection of high-stakes speaking tests and discusses in how far these rating scales capture central aspects of spoken language as highlighted by corpus analysis. It then discusses implications of our MICASE-based findings for (academic) speaking assessment. In the light of corpus findings, the paper challenges the prevalent separation of vocabulary and syntax in assessment criteria. It questions whether scoring criteria such as "Grammatical Resource" and "Lexical Resource" (UCLES 2012: 64) can and should actually be kept separate in assessing speaking proficiency. Overall, the paper provides evidence for the interrelatedness of vocabulary and grammar in academic speech and stresses the importance of phraseology as a core, rather than a peripheral aspect of language (see Ellis 2008), adding to a growing body of existing work in corpus research on phraseology (see e.g. Biber 2009, Hoey 2005, O'Donnell, Römer and Ellis 2013, Römer 2009, 2010, Sinclair 2008). It demonstrates how corpus analysis can contribute to a better understanding of the real-world spoken lexicogrammar and how it helps us uncover the patterned nature of speaking.

References

Biber, D (2009) A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing, *International Journal of Corpus Linguistics* 14(3), 275–311.

Ellis, N C (2008) Phraseology: The periphery and the heart of language, in Meunier, F and Granger, S (Eds.), *Phraseology in Language Learning and Teaching* (pp. 1-13), Amsterdam: John Benjamins.

Flowerdew, L (2012) *Corpora and language education,* London: Palgrave Macmillan.

Hawkins, J A and Filipovic, L (2012) *Criterial features in L2 English,* Cambridge: Cambridge University Press.

Hoey, M (2005) *Lexical priming: A new theory of words and language,* London: Routledge.

O'Donnell, M B, Römer, U and Ellis, N C (2013) The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm, *International Journal of Corpus Linguistics* 18(1), 83-108.

Reppen, R (2010) *Using corpora in the language classroom,* Cambridge: Cambridge University Press.

Römer, U (2009) The inseparability of lexis and grammar: Corpus linguistic perspectives, *Annual Review of Cognitive Linguistics* 7, 140-162.

Römer, U (2010) Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews, *English Text Construction* 3(1): 95-119. [Reprinted in Biber, D & Reppen, R (Eds.) (2012) *Corpus linguistics. Volume I: Lexical studies,* London: SAGE Publications.]

Römer, U (2011) Corpus research applications in second language teaching, *Annual Review of Applied Linguistics* 31, 205-225.

Sinclair, J M (2008) The phrase, the whole phrase, and nothing but the phrase, in Granger, S and Meunier, F (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407-410), Amsterdam: John Benjamins.

Simpson, R C, Briggs, S L, Ovens, J and Swales, J M (2002) *The Michigan Corpus of Academic Spoken English,* Ann Arbor, MI: The Regents of the University of Michigan.

UCLES (2012) *Cambridge English: Advanced. Certificate in Advanced English (CAE). Handbook for teachers,* Cambridge: University of Cambridge ESOL Examinations.

Zipf, G K (1935) *The Psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: The M.I.T. Press.