Biber, D. (2020). Corpus analysis of spoken discourse. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching conference*, ISSN 2380-9566, Northern Arizona University, September 2019 (pp. 5–7). Ames, IA: Iowa State University.

CORPUS ANALYSIS OF SPOKEN DISCOURSE

Douglas Biber, Northern Arizona University

Although the majority of corpus-based linguistic studies have focused on written discourse, there has been a surprisingly long research tradition focused on the corpus-linguistic description of spoken discourse. Thus, while the earliest electronic corpora of written texts (the Brown Corpus and LOB Corpus) were constructed in the 1960s and early 1970s, work on a major corpus of natural spoken texts – the London-Lund Corpus – was already underway by the mid-1970s. Similarly, corpus-linguistic studies of lexico-grammatical features in speech can be traced back to this same time period (see, e.g., Aijmer, 1984; Stenstrom, 1986).

Once collections of spoken transcripts have been transcribed, there is no reason why they cannot be analyzed as a corpus using the same techniques as corpus-based analyses of written discourse. And in fact, there have been many studies of this type. For example, the *Longman Grammar of Spoken and Written English* (Biber et al., 1999) – based on corpus analyses of the full range of lexico-grammatical features in English – documents the patterns of use for both conversation as well as written registers like newspaper prose. The book *University Language* (Biber, 2006) similarly tries to undertake a comprehensive linguistic description of spoken as well as written registers found in American universities, again based on a large-scale corpus analysis of transcribed-spoken and written texts. And Appendix A (by Barbieri & Wizner, 2019) in a recent textbook on register analysis (Biber & Conrad, 2019) catalogs dozens of corpus-based studies focused on the distinctive lexico-grammatical characteristics of spoken registers.

The more controversial question, though, is whether corpus-analysis techniques can be applied to the study of phonetic and prosodic patterns in a spoken corpus? The two main issues here are: 1) the nature and availability of spoken corpora, and 2) the research goals of corpus-linguistic analyses.

Regarding the corpora, there are currently numerous publicly available corpora with orthographic transcription of spoken discourse (e.g., the BNC, MICASE, BASE, and COCA). Although these corpora provide amazing resources for the study of lexis and grammar in speech, they are of no help for the study of prosodic or phonetic patterns. In that regard, there are three major publicly-available corpora with orthographic transcriptions plus annotation of prosodic features (such as pausing, prominence, and tone choice): the London-Lund Corpus, the Hong Kong Corpus of Spoken English, and the C-ORAL-ROM (which is actually a collection of five different spoken corpora for each of the major Romance languages). In addition, there are private spoken corpora that have been carefully annotated for prosodic features, such as the Nurse-Standardized Patient (NSP) Corpus developed by Staples (2015). Thus, there are reasonable corpus resources available for researchers who hope to analyze prosodic patterns across speakers and situations of use.

In contrast, there are no phonetically transcribed corpora currently available to the public. This is a strong assertion, based on a technical understanding of what a 'corpus' is: a large and principled sample of texts designed to represent a target domain of language use (e.g., a language, dialect, or register; see Egbert, Gray, & Biber, forthcoming). Thus, according to this conceptualization, phonetically-transcribed speech collections like the Speech Accent Archive and the so-called TIMIT Corpus are archives but not corpora. That is, these are collections of texts, but because

there is no particular design that motivated the collection, and no intention to represent any domain of language use, they are simply collections (or archives) and not corpora.

Another way to think about this difference is from the perspective of the research goals of the analysis. The goals of a linguistic analysis of a corpus are generalizable patterns – i.e., a discovery of patterns of language use that can be generalized to the domain represented by the corpus. In contrast, the goals of analyzing speech excerpts in an archive are to illustrate differences across speakers – with no attempt to discover generalizable patterns. This is not to say that speech archives have no value. Just the opposite is the case: speech archives are a wonderful resource for illustrating different ways in which speakers phonetically realize the same words and expressions. But that application is not the same as a corpus-linguistic investigation of generalizable patterns.

One major reason that phonetically-transcribed corpora are not publicly available is that it would require an incredible amount of work to develop one. First, spoken texts would need to be collected and recorded in a way that represented a domain of use. Second, those texts would need to be transcribed orthographically. And third, the texts would need to be transcribed phonetically and linked in a parallel manner to the orthographically-transcribed corpus. The motivation for the third step is clear: we need phonetic transcriptions if we are going to analyze phonetic patterns! But the motivation for the second step might be less obvious. The reason that we would need a parallel orthographically-transcribed corpus is that all linguistic searches of words and grammatical constructions would be based on that version. To take a simple example, it would not be possible to analyze all of the different phonetic realizations of the word *horse* if there was no automatic way to identify occurrences of the word *horse*!

Thus, to date, there have been no major corpus-based analyses of phonetic patterns in a large generalizable corpus. There have, however, been model corpus-based analyses of prosodic patterns. Three of those studies are briefly described here: Cheng et al. (2008), Staples (2015), and Biber and Staples (2014). In the first of these, Cheng and her colleagues analyzed prosodic patterns in the 1 million-word Hong Kong Corpus of Spoken English (HKCSE), which was designed to represent the English language use of native speakers of English (NES) versus native speakers of Cantonese (HKC), as it occurred in four spoken registers: academic teaching, business interactions, casual conversations, and public conversations. One focus of the investigation was the comparison of NES and HKC speakers, finding, for example, that HKC speakers are much more likely to express prosodic prominence on personal pronouns than NES speakers.

In the second of these studies, Staples (2015) compared the prosodic discourse styles of internationally-educated nurses (IENs) and US-educated nurses (USNs), focusing especially on prosodic differences in the realizations of specific speech acts by IENs versus USNs. This study was based on analysis of the prosodically-annotated NSP Corpus, with c. 80,000 words of text from 102 nurse-patient interactions. The study uncovered several major ways in which the typical prosody employed by IENs differed from that employed by USNs. For example, in their empathetic responses to the patient (e.g., *Oh, well, I'm sorry to hear that*), USNs typically employed a much greater pitch range than IENs, while IENs tended to produce flat tone units with little variation in pitch. USNs also used falling tone to a much greater extent than IENs, while IENs tended to rely on level tone.

Finally, the Biber and Staples (2014) study is also based on the HKCSE, comparing the prosodic realizations of stance adverbials by NES versus HKC, across three different spoken registers. One general pattern of use across all three of these studies is that non-native speakers of English are much more likely to employ prosodic prominence than native speakers. For example, in the Biber/Staples study, roughly 90% of occurrences of the stance adverb *actually* in conversation were realized with prominence when they were produced by HKCs, versus only 50% of *actually* tokens produced by NESs.

In summary, corpus-based analysis is a promising research approach when applied to spoken corpora, uncovering generalizable patterns of use that were not anticipated by casual observation. The major limiting factor for such analyses is the availability of appropriate corpora: at present, there are only a few available corpora that include prosodic annotation, and no phonetically transcribed corpora. Thus, the major take-away message is the need for research efforts to develop spoken corpora with prosodic annotation and phonetic transcriptions, providing the basis for future research with a much more generalizable basis than currently possible.

ABOUT THE AUTHOR

Douglas Biber is Regents' Professor of English (Applied Linguistics) at Northern Arizona University. His research efforts have focused on corpus linguistics, English grammar, and register variation (in English and cross-linguistic; synchronic and diachronic).

REFERENCES

Aijmer, K. (1984). Sort of and kind of in English conversation. *Studia Linguistica, 38,* 118-28.

Biber, D., and S. Conrad. (2019). *Register, Genre, and Style [2nd Edition]*. Cambridge: Cambridge University Press.

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

Biber, D. (2006). University language: *A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, D. & Staples, S. (2014). Exploring the prosody of stance: Variation in the realization of stance adverbials. In T. Raso & H. Mello (Eds.), *Spoken Corpora and Linguistic Studies (*pp. 271–294). Philadelphia: John Benjamins.

Barbieri, F., & Wizner, S. (2019). Appendix A: Annotations of major register and genre studies. In D. Biber & S. Conrad (Eds.), *Register, Genre, and Style* [2nd edition], (pp. 318-349). Cambridge: Cambridge University Press.

Cheng, W., Greaves, C., & Warren, M. (2008). *A corpus-driven study of discourse intonation: The Hong Kong Corpus of Spoken English (prosodic)*. Philadelphia: John Benjamins.

Egbert, J., B. Gray, and D. Biber. (in preparation). *Towards Representativeness in Corpus Design.* Cambridge: Cambridge University Press.

Staples, S. (2015). *The discourse of nurse-patient interactions: Contrasting the communicative styles of U.S. and international nurses.* Philadelphia: John Benjamins.

Stenström, A-B. (1986). What does really really do? In G. Tottie & I. Backlund (Eds.), *English in speech and writing* (pp. 149-64). Stockholm: Almqvist and Wiksell.