Vignette 12b
Establishing Corpora from Existing Data Sources

Mark Davies

Corpora are searchable collections of spoken and written language (nearly always in electronic format) which can be used for linguistic analysis. Ideally, the texts come from sources where, at the moment of speech or writing, there was no understanding that the materials would later be used in a corpus for linguistic analysis, since this helps to preserve the "naturalness" of the language.

Until recently, the largest publicly available spoken corpus was the 10 million words of spoken English in the 100-million-word British National Corpus (BNC). Other important corpora of spoken English are the Cambridge and Nottingham Corpus of Discourse in English, the Cambridge North American Spoken Corpus, the Santa Barbara Corpus of Spoken American English, the Switchboard corpus, and the CallHome corpus. Unfortunately, with the exception of the BNC, most of these corpora either are not publicly available (they are just used for in-house materials development) or are prohibitively expensive for most researchers (costing $1,000 or more).

Because of the issues with pricing and (lack of) availability, some researchers might consider creating their own corpora. Unfortunately, it is almost prohibitively difficult for individual researchers to create large spoken corpora "from the ground up." It takes a very long time and a great deal of money to design a corpus and find speakers, record the speech, and (especially) to carefully transcribe the speech and then revise and correct the texts. The only reason it could be done in the cases of the spoken corpora listed above is that there was typically a large research team and robust funding for creation of each corpus.

As a result, the most realistic alternative for most researchers is to create corpora from existing resources. This was, for example, the process that was followed in the creation of the Corpus of Contemporary American English [COCA] (Davies, 2009; 2011). Although this is the largest and most up-to-date publicly available corpus of English, it was created by just one person in less than a year. The corpus contains 425 million words of text (including 85 million words of spoken language – eight to nine times the size of the spoken portion of the BNC). And unlike any other corpus of English, COCA continues to be expanded: 20 million words of text (including four million words of spoken English) continue to be added each year. The remainder of this vignette will focus on some of the issues raised in the creation of COCA from existing resources, for the benefit of others who might want to follow a similar path.

Perhaps the most obvious question is what resources to use to find spoken texts. One approach might be to consider texts that are not actual "spoken texts" per se but attempt to model natural spoken language, including scripts for television series, radio, movies, and plays. COCA has more than 15 million words of text from these types of sources, and there are probably hundreds of millions of words of text from such resources freely available online. (For example, in just a day or two we created another 70-million-word corpus of scripts from US soap operas.) The question with these "pseudo-spoken" texts, however, is how closely they in fact represent actual spoken language. Many different phenomena in COCA – lexical, phraseological, and syntactic – show that while these scripts are probably the most "spoken-like" of all of the non-spoken genres (fiction, magazines, newspapers, and academic journals), there is still a noticeable difference between these texts and those from actual spoken English. (For this reason, these texts are categorized as "Fiction" in COCA.)

Another possibility might be to find interviews online, as we did while we were compiling the spoken component of the 100-million-word Corpus del Español and the 45-million-word Corpus do Português. There are at least three issues involved in using these resources, however: First, some of the texts that are the easiest to find come from speech types that are probably overly formal, such as political press conferences, and that may only slightly resemble natural, conversational speech. Second, there is a question of how much post-interview "editing" and cleanup has already been done to the texts to eliminate things like hesitation, false starts, and backchanneling. Third, creating such a corpus may involve a great deal of manual editing to extract the interviews from thousands of web pages on hundreds of websites, each with its own formatting for headers, footers, ads, and comments.

Recognizing these limitations, perhaps the best source for spoken language are the transcripts of unscripted speech on television and radio programs such as *Oprah*, *Jerry Springer*, *Geraldo*, *Good Morning America* (ABC), *60 Minutes* (CBS), *Larry King Live* (CNN), or *All Things Considered* (NPR). As I have already noted, more than 85 million words of speech from such resources were used in the creation of COCA, and these 85 million words of spoken data are just a small fraction of what is available online. For example, CNN alone has freely downloadable transcripts of all of its programs from the past 10 years or so, representing more than 250 million words of text.

These transcripts typically do not have the shortcomings of some of the formal interviews discussed above. First, the transcripts cover a wide range of speech types and topics, such as interviews with politicians, actors, or sports figures, or discussions about parenting, hairstyling, new electronic devices, or any number of other topics. This means that the vocabulary is quite diverse, and the style is more informal and natural than that used in press conferences and similar speech types. Second, the transcripts used in COCA have minimal editing to remove features such as hesitation and backchanneling. Third, the page format for all of the tens of thousands of transcripts is typically the same or quite similar, which reduces the problem in processing the texts.

There are two limitations with such transcripts, however. The first concerns the naturalness of the language. The speakers knew that they were on national

TV or radio and were therefore probably on guard to avoid non-standard features like double negation (*She doesn't have no reason*), double modals (*They might could do it*), lexical items and constructions like *ain't* or *had went*, and profanity (which would, in any case, be censored by the television or radio program). Nevertheless, as is discussed in Davies (2009), for most linguistic phenomena these transcripts still model normal everyday conversation quite well. For example, colloquial features like quotative *like* (*He was like, I'm not going with her*), *so not* ADJ (*She's so not interested in him*), or even the common *you know* (*He's, you know, kinda worried about her*) are much more common in the spoken data in COCA than in the data from other genres (fiction, popular magazines, newspapers, and academic journals).

The second concern about using transcripts is the difficulty in coding them for demographic information, e.g., age, gender, ethnicity, or socioeconomic status. There are more than 40,000 spoken transcripts in COCA, with at least two and perhaps as many as 10 or 20 speakers in each transcript, and someone would need to find demographic information for each of the hundreds of thousands of speakers. For lesser-known participants on these programs, this would likely not be possible, and even for those where it is possible, it would be extremely time-consuming (perhaps 25,000-plus hours) and very expensive (hundreds of thousands of dollars). For a smaller corpus (e.g., 100,000–1,000,000 words), it might be possible to code for speaker variables, but then the corpus might only be large enough for it to be possible to look at very frequent linguistic phenomena, such as discourse markers or very high-frequency grammatical constructions.

One way around this problem of sparse demographic coding would be to focus on comparing the different television and radio programs, rather than all of the speakers on these programs. Obviously, this would not give the level of demographic encoding that most sociolinguists are accustomed to, but it is likely the only possibility for large corpora that are created from existing resources. For example, one could easily and quickly create a 5-million-word *Oprah* or *Jerry Springer* corpus (which is presumably fairly informal) and compare it to a 5-million-word corpus containing more formal conversation on a program like *Face the Nation* or the *Newshour* on PBS, with perhaps an intermediate corpus from programs like *All Things Considered* or *Good Morning America* in the mix as well.

In summary, there are a wide range of sources that are publicly available, which allow researchers with even very limited funds and personnel to create very useful corpora of contemporary (spoken) language.

References

Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159–190.

Davies, M. (2011). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*, 447–465.

Vignette 12c
Working with "Unconventional" Existing Data Sources

Joan C. Beal and Karen P. Corrigan

In this vignette, we share our experience of working with data collected at various times and according to varying methodologies to create the Newcastle Electronic Corpus of Tyneside English (NECTE) (for a fuller account, see Allen et al., 2007). In creating this corpus, we faced a number of challenges, some of which required us to devise new policies and protocols, albeit with advice from colleagues. Given the endeavors of sociolinguists working in the pre-digital age, there must be many important and useful collections of data languishing in cupboards, on shelves, or even under beds. We hope that this account of our experiences will inspire readers to rescue these data from "shedding the hard-won sounds of 20th-century speech in the constantly dispersing particles of ferric oxide of an obsolescent recording system" (Widdowson, 2003, p. 84).

The primary data behind NECTE were collected by two teams of sociolinguists, one working in the late 1960s and early 1970s on the Tyneside Linguistic Survey (TLS) (see Pellowe, Strang, Nixon, & McNeany, 1972, for the TLS methodology) and the other in the 1990s for the Phonological Variation and Change (PVC) project. The latter dataset posed fewer problems for the NECTE team, since it had been collected using what are still considered state-of-the-art methods and recorded in digital format (see Milroy, Milroy, & Docherty, 1997). We therefore concentrate on the challenges involved in processing the TLS data.

The first challenge was to find as many of the data and accompanying metadata as possible. The majority of the data had been left in the department of Newcastle University where the TLS team had worked. Unfortunately, the materials were not stored in controlled archival conditions but rather in unlabeled boxes in store-cupboards, in serious danger of deterioration. More data came to light only after our project began, when a former member of the TLS team brought back some recordings and index cards he had taken with him upon relocating. Although the original TLS data collection was carried out in accordance with the principle of random sampling, the NECTE team did not inherit the original random sample in this technical sense and instead inherited ad hoc remnants of it. Nevertheless, a majority of the interviews were, in fact, preserved. Moreover, the richness of the social data collected by the TLS team has ensured that NECTE users can make up their own balanced sample from the available material, as has already been done in publications such as Beal and Corrigan (2005a; 2005b). More recently, Barnfield and Buchstaller (2010) have sampled