COCA-SPOKEN

In the British National Corpus, the Spoken texts account for 10% of the corpus (or about 10 million words). In COCA, we wanted Spoken to be more than that, and about 12.5% of the corpus (~127 million words) comes from spoken American English. It would have been impossible, however, to create a corpus that size by tape recording lectures, conversations, etc – especially since the corpus was created by one person, in about 3-4 months, with a budget of approximately $0.

The only option was to use transcripts of conversations, which were already in electronic form. Therefore, we obtained transcripts of unscripted conversation on TV and radio programs like All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Today Show (NBC), 60 Minutes (CBS), Hannity and Colmes (Fox), Jerry Springer (syndicated), etc.

There are several questions, of course, regarding the use of transcripts like these. Perhaps the three most important ones are:

1) Do they faithfully represent the actual conversations?

2) Is the conversation really unscripted?

3) How well does it represent "non-media" varieties of Spoken American English?

Regarding the first question, we feel confident that the transcripts do represent very well the actual spoken conversation. Look at this video from the Larry King show on CNN (it comes on after the ad) and then compare it to the transcript. Another example is from the "Talk of the Nation" show on NPR. Compare the audio recording (click on "Listen Now") with the transcript (as contained in our corpus). Our sense is that the transcripts do an excellent job transcribing the conversation, including interruptions, false starts, and so on.

The second question is whether the conversation is really unscripted. In the Larry King interview above, there are a handful of "formulaic / scripted" sentences like "Welcome to the program", "We'll now go to a commercial break", etc. But probably 97-98% or more of the conversation is unscripted. In the NPR transcript, there is a bit more scripted material -- a paragraph or two at the beginning of the show, and some announcements for upcoming commercial breaks. But about 95% or so is still unscripted. The question is whether you would rather have an 80+ million word spoken corpus with about 5% scripted material (but still leaving more than 75 million words of unscripted material), or a "completely pure" corpus that is so small (1-2 million words) that it is unusable for many types of research. We opted for the former.

In terms of the third issue (naturalness), there is one aspect of these texts that does make them somewhat unlike completely natural conversation. That is of course the fact that the people knew that they were on a national TV or radio program, and they therefore probably altered their speech accordingly -- such as relatively little profanity and perhaps avoiding highly stigmatized words and phrases like "ain't got none". In terms of overall word choice and "natural conversation" (false starts, interruptions, and so on), though, it does seem to represent "off the air" conversation quite nicely. But no spoken corpus (even those created by linguists with tape recorders in the early 1990s) will be 100% authentic for real conversation -- as long as people know that they're being recorded.

Finally, it is possible to do some quick searches that show the overwhelming "spoken" nature of these texts. The following phrases are ones that we would expect to occur much more frequently in spoken (American) English than in other genres. Click on them to see how common they are in spoken and the other genres:

and I'm like ,

Do you think

, you know ,

so not ADJ

I guess that

. Well ,

. Sure .

One other note:

In the transcripts there is text indicating who the speaker is, or codes referring to "voice-overs" or other notations made by the transcriber. For example:

SUMMIT It should be a C-note. Mr. CARY ANDERSON That's it. Mr. ANDERSON Oh, very good. See, you didn't have to get nervous, Mr. Cronick. You were really very good at it. SUMMIT All right. Mr. ANDERSON You -- you were coming fast and furious here. It was great. I could sleep. (Laughing) I appreciate it.

We were able to "separate out" most of these (shown in gray above), although there is a bit more of this in the transcripts from 2008. Where they have been separated out, they are not included in the overall word count, nor can they be searched for (this is on purpose, since they're not "spoken"), but they do appear in the Keyword in Context displays.