Teaching and Language Corpora (TALC) '98, Oxford, UK, July 25-27, 1998

Constructing a New Corpus of Spoken American English

Hongyin Tao Linda R. Waugh
Cornell University

1. Purpose

The Department of Modern Languages at Cornell University plans to conduct a joint project with Cambridge University Press to build a spoken corpus of North American English. The main goal of the project is to collect a large amount of naturalistic North American (American and Canadian) English data. The purpose of the project is to provide a sound basis for research into the English language and for data-driven language teaching and learning. In addition, the project aims to provide additional spoken data for comparative/contrastive research into aspects of British and North American English.

Since the project is still in its early planning stage, we would like to take this opportunity to outline some of the important aspects of our vision of the project, with the understanding that future developments of the project will not necessarily follow what we are describing here. That is, we view this gathering of corpus linguists and language teaching professionals as a great opportunity to solicit suggestions for the construction of the proposed project.

2. Corpus Design

2.1 Discourse genre

As with any corpus project, the kind of data to be collected is closely tied to the potential use of the database. In our case, since we intend our database to be useful for both research and language teaching, as well as for comparative studies between British and North American English, we would like to collect many varieties of spoken English in terms of discourse genres or speech situations. Our British colleagues at Cambridge University Press, in collaboration with corpus linguists at the University of Nottingham, have developed a framework where spoken texts can be classified in a relatively systematic way, with the understanding that any attempt to taxonomize spoken discourse is bound to meet with borderline cases. For maximum comparability, we would like our data collection generally to follow the range of text types as proposed by our British colleagues. These speech genres are discussed in Carter and McCarthy 1997, which we list below without further elaboration.

Narrative
Identifying
Language-in-action
Comment-elaboration

Service encounters
Debate and argument
Language, learning and interaction
Decision-making/negotiating outcomes

2.2 Data management

Our data will come from both audio and video recordings of natural interactions. Given the ease of accessibility of current computer technology, we intend our data management to take full advantage of the digital technology that is current available. More specifically, (1) all our recordings will go through a digitization process for backup, transcription, and further publication. This will ensure a high level of data storage and manipulation. (2) Information about individual speech events (including general descriptions of speech topics) and speakers will be entered into a relational database, and will further be linked to transcripts and audio files.

2.3 Transcription and mark-up

We intend to adopt a dual system in transcription. That is, we will apply a broad transcription system for all the recordings that we decide to transcribe. At the same time, we will apply a narrower system to transcribe a small amount of data that may better suit the needs of some discourse researchers. The narrow system will be a tailored system on the basis of Du Bois et al. 1993. With this system, particular attention will be paid to the following properties of spoken discourse.

1. Words	Standard orthography, except for some widely accepted contraction forms (going to -> gonna, fellow -> fella).
2. Speakers and turns	A: B: Mary: Joe: X: Unknown speaker
3. Intonation units	A chunk of speech under a single coherent intonation contour.
4. Speech overlap	At least a rough indication. B: ...I remember, ...( ) I used to help Benny, and I'd get twenty-five cents a week. R: ... ( ) [A ^week]! B: [Twenty] --
5. Intonation contour class	. Final, complete intonation , Continuing ? Question ! Exclamatory -- Truncated
6. Truncation (words and intonation)	wor- I thought --
7. Pauses	Short: .. It looks like, Intermediate and long: ...( ) It looks like, (without indicating exact length of pause)

8. Laughter	@@
9. Uncertain hearings	XXX (three syllables), XXXXX (five syllables)
	Optional Qualities
10. Accent	^: ^week!
11. Lengthening	=: Withou=t that it's impossible to tell
	Non-verbal Behavior
12.Hand gestures	Minimum, descriptive, only for the obvious ones.
13. Head movements	((Pointing to Joe))
14. Eye gaze	((Turn to refrigerator))
14. Body orientation	((Turn on TV))

We anticipate that at least a small portion of our transcripts will be marked up with a set of tags indicating information concerning text structure, speech events, as well as other discoursal and grammatical properties that we deem necessary. This will make the transcribed data more cross-platform and more useful for indexing and searching purposes. At the moment we have not finalized any system for data mark-up.

2.4 Accessing the data

For the authorized user, accessing the data can be done in a variety of ways. In addition to traditional methods of searching the data, for example, we would like to explore the possibility of accessing concordance data via the World Wide Web, and possibly be able to link the transcript to the audio/video file. The advantages of using the Web include 1) original data are centrally stored and are thus more secure; and 2) it makes remote access most flexible, and with a least degree of human intervention. This seems to us to be one of the most cost-effective and user-friendly ways of making use of the data.

3. Short term and long term goals

Our immediate goal will be to collect about 50 hours of recordings, extracts of which will be transcribed to establish a half million word textual database. We plan to expand our collection as the project moves along.

4. Further implications of the corpus

Currently, the Department of Modern Languages at Cornell University teaches about thirty languages in its curriculum, and many of our teaching staff feel that a natural

discourse based approach to language pedagogy provides an exciting new perspective for second language acquisition, and many instructors have already began to undertake similar projects for their own languages. We intend to include foreign language learners of American English, and American English learners of languages as diverse as Indonesian, French, Mandarin Chinese, and Swedish. Here at Cornell, for example, we have plans to collect French native speaker data this coming year. With the English corpus as a pilot project, we hope to be able to establish procedures and tools for the building of potentially more than a dozen (native and second language learners') language corpora in our department in the near future.

5. Conclusions

Although a great deal of spoken English data have been collected in Europe, especially in the UK, limited effort has so far been made in North America to document large quantities of natural speech (with a few exceptions such as the Santa Barbara Corpus of Spoken American English). We are confident that the proposed corpus will make a significant contribution to the corpus linguistics community and to discourse linguistics in general.

References

Carter, Ronald, and Michael McCarthy. 1997. *Exploring spoken English*. Cambridge: Cambridge University Press.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming & Danae Paolino. 1993. "Outline of Discourse Transcription". In Edwards, Jane A. & Martin D. Lampert, eds., *Talking Data: Transcription and coding methods for discourse research*. 45-89. Hillsdale, NJ: Lawrence Erlbaum Associates.