In the Image of Language

Corpus Construction as Discourse Representation

John W. Du Bois
Linguisics Department, University of California, Santa Barbara
Director, Santa Barbara Corpus of Spoken American English

Kyoto, December 2011

ABSTRACT

Linguists don't often ask themselves what language is, or what it looks like, but perhaps they should. When the time comes to create an image of language, which will be contemplated as an object of scientific study, the question becomes more urgent. What is the object that we trying to represent? And how can we represent it in a way that captures its unique qualities, in a way that best supports our efforts to understand it? The question of representation is two-fold. The selection of materials for inclusion in a corpus is designed to "represent" the language of a certain group of people, or a certain ways of using language by that group. That is, the corpus is expected to be "representative" of the language or variety in question. At the same time, if the corpus materials include audio or video recordings of spoken language in use, there is the additional task of "representing" in writing what is happening on the recordings. We thus find that representation is a critical component of both the design and the transcription of a corpus of language in use.

But the real meaning of a corpus remains elusive. We can start by defining a corpus as a systematic, unified, representative body of observational data on a language. Or more vividly: a corpus is a slice of language. We cut a small slice in an attempt to represent the larger body of a language, whether English or French or Japanese. But this confronts us with the question: what is language itself, the object of our representation? Traditional corpus linguistics views language, and hence the corpus, as a collection of words in structured sequence -- nouns and verbs and other grammatical elements which it studies in order to arrive at generalizations about units, structures, rules, meanings, and frequencies. This is fine as far as it goes. But what if language also includes the meaning of silences, the pragmatic thrust of a conversation, the interactional goals of its participants, the social consequences that propel the talk from one construction and one activity to the next? I will suggest that the best way to take a slice of language is to take a slice of life; and this is what it means to represent a language. The approach to discourse representation that I propose is essentially ethnographic in its intent, seeking to capture the full scope of human life as represented in the way a group of people use language. But if the goal is to allow us to answer the larger questions about how and why speakers use language as they do, I would argue that such a corpus is of value not only for the sociocultural linguist but also for the scholar of linguistic structure -- the grammarian who really wants to understand grammar. This approach to creating a portrait of discourse, an image of language, seeks to put language in a new light, and to open up new questions, including questions about the relation between structure and use. I present examples of audio recordings of conversations and other speech events drawn from the Santa Barbara Corpus of Spoken American English, transcribed according to a method that I have been developing over a number of years. The goal is to show how an ethnographically sensitive approach to representing spoken discourse can create a corpus that has meaning for a diverse audience ranging from grammarians to language teachers to pragmaticists to social scientists.