Charles Fillmore
Lancaster University

## Pie-in-the-sky corpus tools to support semantic parsing

The FrameNet lexicon-building project at the International Computer Science Institute (http://www.icsi.Berkeley.edu/~framenet) has been devoted these past seven years to collecting and displaying lexical information, based on corpus evidence and expressed in terms of a frame-semantic descriptive vocabulary, that emphasizes the syntactic and semantic combinatory possibilities for English lexical units. Since the work has for the most part been carried out frame-by-frame, rather than by word-class, the product is partly an inventory of paraphrase patterns that cross part-of-speech categories.

In recent years we've been contracted to provide frame-semantic annotations for doing semantic parsing of full texts. Now, instead of picking and choosing good corpus examples to illustrate the combinatorial facts we're discovering, we have to account for every word in each selected text, and in a way that will facilitate the creation of meaning-representations for the whole. (So far, the texts have mainly been articles from the Wall Street Journal or documents on international treaties or on state and non-state use of WMDs.) In the course of thinking about how to do a good job on these more complex tasks, I have been daydreaming about the kinds of computational facilities we need.

Briefly, we need a stand-off annotation facility with an arbitrary number of "layers" indexed to character-strings in the texts being analyzed: character strings rather than space-separated "words" because some unspaced stretches are syntactically complex, and some word sequences function as syntactic/semantic units. In addition to text-annotation itself, we need access to information about the sub-structure of morphologically or syntactically complex lexical units, information about semantic frames evocable by each lexical unit, and information about complex grammatical constructions that are not expressible with ordinary constituency or dependency representations. For display purposes, all of this inventoried information can be aligned with the annotation when needed. We imagine our database integrated with vast ontological resources having various kinds of encyclopedic knowledge and common-sense inferencing abilities. We need layers for morpheme-identification, lexical-unit identification (linked to frame and valence information), named-entity recognition (linked to relevant category labels), the recognition of idioms and special grammatical constructions, and syntactic parses.

Numerous obstacles stand in the way of such a pie-in-the-sky program. The most obvious is that, because of the recursive property of language, each unit at a given level might have a complex internal structure that requires its own layering. Registrally, grammatically or lexically licensed gaps and discontinuities provide their own challenges, for both discovery and representation. Essential to the software of my dreams is the ability to modify the information at each layer, in places where the algorithms fail, and to decide which of the modifications should provide feedback to a learning mechanism. This will include relabelings and reattachments of parses, additions to inventories of idioms, compounds and named entities and exophoric links to metadata.

Much of the layering technology is in place, and is a part of our regular operations; and the existing database structures provides places (still empty) for some of the kinds of information we need. Collaboration with institutions equipped with efficient named-entity recognizers and ontologies is under way. The dream is inspired by Adam Kilgarriff's "Web as Corpus" ambitions and Adam Meyers' "Ultimate Annotation" goals.