SESSION 10b: CORE NL LEXICON AND GRAMMAR

Mark Liberman, Chairperson

Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104-6305

ABSTRACT

The value of generally-available speech and text corpora in facilitating research is clear—these resources are simply too expensive for each site to gather individually, and fair comparative evaluation of approaches requires shared training and testing material.

Similar considerations apply in the case of lexicons and grammars for broad-coverage NLP applications. Lexicons and grammars are very expensive to build, and it seems wasteful for every site to have to build them over again, especially in cases where no new invention is intended. Also, fair comparative evaluation of other components in modular systems seems to require combining them with a shared lexicon and grammar.

However, it is unclear whether the community of researchers can agree that a particular design is appropriate, and that a large-scale effort to implement it will converge on a result of fairly general utility. This session aimed to raise these issues, and to begin a discussion that we hope will result in a recommendation for action in the near future. The newly-formed Linguistic Data Consortium has a mandate to serve the community's needs in this area, and it will start from the ideas presented in this session by the panelists (Jerry Hobbs, Ralph Grishman, Paul Jacobs, Bob Ingria, Louise Guthrie, and Ken Church) and the audience.

1. Summary of Panelists' Presentations

Six panelists made brief individual presentations. Three of these dealt with lexical issues.

- Louise Guthrie (NMSU) presented a suggested from Yorick Wilks for a "core lexicon" that would be application-independent and easy to adapt to a wide range of parsers.
- Bob Ingria (BBN) discussed his experience with the similarities and differences in the form and content of the lexical entries used by different systems.
- Ken Church (AT&T) , playing devil's advocate, argued that the effort to turn lexical raw materials into a shared computational lexicon might better be spent on more raw materials.

Three other panelists dealt with grammatical issues.

- Jerry Hobbs (SRI) argued for a project to produce a "National Resource Grammar," an extensible, efficient broad-coverage English grammar that could be distributed generally to the DARPA community and to other researchers.
- Ralph Grishman (NYU) suggested that we need to know where our grammars and parsers stand, and that an on-going program of comparative quantitative evaluation would define the state of the art, and should also lead to significant improvements.
- Paul Jacobs (GE) discussed his experience in combining components from earlier systems implemented at GE and CMU, doing the computational-linguistics equivalent of putting a Ford engine into a GM chassis; the fact that such "transplants" are feasible increases the credibility of of an effort to produce common lexical and grammatical components.

2. Discussions and Conclusions

The panelists and the audience touched on a spectrum of resources ranging from "raw materials" such as corpora to "finished components" such as computational grammars and lexicons. In between are such things as annotated corpora, pronouncing dictionaries, specialized word lists such as gazeteers, thesaurus-like structures such as George Miller's WordNet, part-of-speech taggers, lexical materials such as lists of verb subcategorization frames, and partial parsers of various kinds.

The proposals from Guthrie/Wilks, Hobbs, and Grishman evoked both sympathetic responses and skeptical ones. One point that emerged was that the DARPA community should have a Written Language Coordinating Committee, to complement the Spoken Language Coordinating Committee, and to enhance interchange among DARPA WL contractors. The WLCC has since been organized, with Ralph Grishman as chair, and will provide a forum for further discussion and action on these issues.