SESSION 1: LEXICONS, CORPORA, AND EVALUATION

George A. Miller, Chair

Cognitive Science Laboratory
Princeton University
Princeton, NJ 08542

Our technologies for collecting, storing, and disseminating vast amounts of information have gotten ahead of our technologies for collating and analyzing it, and that situation has posed a serious challenge for human language technology. As a consequence, natural language processing has been moving rapidly towards large-scale systems addressed to real tasks. Demos that won't scale up are no longer interesting.

Large-scale systems are not feasible, however, without large-scale resources for development and evaluation. Toward this end, the Linguistic Data Consortium was created in 1992 with a combination of government and private funds. The Consortium's mandate is to create a repository of linguistic resources and to make them available for research and development in human language technology. Much of this session was devoted to a description of their progress toward that goal.

In order to deal with the range and variety of words encountered in real life communications, it has been necessary to obtain bigger and better lexicons, and the Linguistic Data Consortium has supported the creation a syntactic lexicon, Comlex. In their report on the Comlex Syntax project, Macleod, Grishman, and Meyers explain how they have added syntactic information to their lexicon, information far more detailed than is found in standard dictionaries.

In order to evaluate proposed systems, it has been necessary to obtain good spoken and textual corpora—large and balanced if possible, but certainly large. A series of three papers describe what the Linguistic Data Consortium has been doing to meet that need in a wide variety of languages. Corpora of telephone speech that are being collected at SRI and at the Oregon Center for Spoken Language Understanding (both with support from the Linguistic Data Consortium) are also described. It is important that these copora will be generally available. If systems are developed or evaluated on different corpora, there is no way to know whether differences in performance should be attributed to the systems or to the corpora.

In some quarters, the opinion seems to be that the singular value of the large corpora that are becoming available now is that they permit statistical analyses that were not possible before. I have nothing against statistical analyses—I have walked both sides of that street in my time—but it is important to realize that there are other good reasons for wanting to collect large corpora and make them generally available.

For example, a large textual corpus is an enormous aid in compiling the lexicons that we need. As a lexicon grows in size, the entries come closer and closer to the limits of a lexicographer's personal knowledge. Then it becomes important to be able to consult examples drawn from actual usage. But to get, say, 20 examples of a rare word, access to a very large corpus is needed.

Large corpora of spoken language are needed to assess speaker differences and to sample speech as it occurs under real conditions. What people say to one another is very different from the edited prose found in books or newspapers. Here again, as recognition vocabularies get bigger, increasingly large corpora are needed in order to have adequate samples of the spontaneous use of rare words.

Still another reason is that corpora are needed to test claims that are made for natural language processing systems. It doesn't matter whether a system is developed with hidden Markov models or with augmented transition networks; in order to test it, you need a representative corpus that has been processed in advance by human language users. Recent approaches to the evaluation of speech systems, and the results of the 1993 benchmark tests of spoken language systems are described.

Finally, although language technologists have little to say about it, another good reason to compile corpora is that the material merits preservation and study in its own right. Humanistic scholars are busily at work collecting local dialects or machine-readable text for their own purposes. It is probably the humanistic background of publishers that leads them to think that their machine-readable text has some intrinsic value, thus causing us legal problems when we try to get permission to use it.

There are, in short, many reasons to make lexicons and corpora readily available to the research community. This session, however, is not concerned to defend their usefulness, but rather to make sure that everyone knows what is available and how to get it.

Unfortunately, the amount and variety of work presented in this initial session left no time for the group discussion that such important topics merit.