The Consortium for Lexical Research

Rio Grande Research Corridor
Computing Research Laboratory
New Mexico State University
Box 30001, Las Cruces, NM 88003.

lexical@nmsu.edu
(505) 646-5466
Fax: (505) 646-6218

Work in computational linguistics has reached the point where the performance of many natural language processing systems is limited by a "lexical bottleneck". That is, such systems could handle much more text and produce much more impressive application results were it not for the fact that their lexicons are too small.

The Association for Computational Linguistics has proposed that a Consortium for Lexical Research (CLR) be established, and DARPA has agreed to fund this. It will be sited at the Computing Research Laboratory, New Mexico, USA, under its Director, Yorick Wilks, and an ACL committee consisting of Roy Byrd, Ralph Grishman, Mark Liberman and Don Walker.

The Consortium for Lexical Research will be an organization for sharing lexical data and tools used to perform research on natural language dictionaries and lexicons, and for communicating the results of that research. Members of the Consortium will contribute resources to a repository and withdraw resources from it in order to perform their research. There is no requirement that withdrawals be compensated by contributions in kind.

A basic premise of the proposal for cooperation on lexical research is that the research must be "precompetitive". That is, the CLR will not have as its goal the creation of commercial products. The goal of precompetitive research would be to augment our understanding of what lexicons contain and, specifically, to build computational lexicons having those contents.

The task of the CLR is primarily to facilitate research, making available to the whole natural language processing community certain resources now held only by a few groups that have special relationships with companies or dictionary publishers. The CLR would as far as is practically possible accept contributions from any source, regardless of theoretical orientation, and make them available as widely as possible for research. There is also an underlying theoretical assumption or hope: that the contents of major lexicons are very similar, and that some neutral, or "polytheoretic," form of the information they contain can be at least a research goal, and would be a great boon if it could be achieved. A major activity of the CLR will be to negotiate agreements with "providers" on reassuring and advantageous terms to both suppliers and researchers. Major funders of work in this area in the US have indicated interest in making participation in the CLR a condition for financial support of research.

The Computing Research Lab (CRL) already has a range of machines appropriate for advanced computing on dictionaries (including the construction of large-scale matrices): DARPA-supported access to a Connection Machine, a large Intel Hypercube, a Sequent Symmetry, and an IBM-ACE parallel machine, as well as a range of conventional hardware. The Consortium would expect to have access to an appropriate range of large-scale storage machines, as well as capacities accepting and providing materials by network, tape and CD.

Resources and Services of the Consortium

The following lists of lexical data and tools seem to provide a reasonable starting content for the repository. We will continually solicit and encourage additions to this list.

Data

1. word lists (proper nouns, count/mass nouns, causative verbs, movement verbs, predicative adjectives, etc.)
2. published dictionaries
3. specialized terminology, technical glossaries, etc.
4. statistical data
5. synonyms, antonyms, hypernyms, pertainyms, etc.
6. phrase lists

Tools

1. lexical data base management tools
2. lexical query languages
3. text analysis tools (concordance, KWIC, statistical analysis, collocation analysis, etc.)
4. SGML tools (particularly tuned to dictionary encoding)
5. parsers
6. morphological analyzers
7. user interfaces to dictionaries
8. lexical workbenches
9. dictionary definition sense taggers

Services

Repository management will involve cataloging and storing material in disparate formats, and providing for their retransmission (with conversion, where appropriate tools exist). In addition, it will be necessary to maintain a library of documentation describing the repository's contents and containing research papers resulting from projects that use the material. A brief description of the services to be provided is as follows:

a. CRL will provide a catalog of, and act as a clearinghouse for, utilities programs that have been written for existing online lexical data.

b. CRL will compile a list of known mistakes, misprints, etc. that occur in each of the major published sources (dictionaries etc.).

c. CRL will set up a new memorandum series explicitly devoted to the lexical center.

d. CRL will also be a clearinghouse for preprints and hard-to-find reprints on machine-readable dictionaries.

e. CRL also expects to conduct workshops in this area, including an inaugural workshop in 1991.

f. CRL would provide a catalog for access to repositories of corpus-manipulation tools held elsewhere.

An annual fee will be charged for membership. We suggest beginning with $2,000 for commercial organizations, and $200 for educational institutions. An even lower fee for individual members may be considered. It is intended that after an initial start-up period, the Consortium become self-supporting. Membership and withdrawal fees may need to be adjusted in order for that to happen.

Anyone interested in participating *even in principle* as a provider or consumer of data, tools, or services should send a message to **lexical@nmsu.edu**.