Presentation

The Laboratoire d'Automatique Documentaire et Linguistique (LADL) is a research laboratory of the University Paris 7 (Department of Computer Science) also supported by the Centre National de la Recherche Scientifique (CNRS (Linguistics)). The size is about 30 people, including graduate students. The research programme of the LADL is centered on the construction of fundamental tools for natural language processing. These tools are essentially:

- linguistic components: electronic dictionaries and grammars, mainly for French and English, but also for other languages (e.g. Spanish, Korean);

- computer programmes: algorithms that apply dictionaries and grammars to corpora, in order to locate and lemmatize meaningful units of texts, the main applications being: automatic indexing of texts, research of information in full texts, helps for translation.

LADL has close links with the Centre d'Etudes et de Recherches Linguistiques of the Research Institute Gaspard Monge (University of Marne-la-Vallée) and the Laboratoire d'Informatique Linguistique of the University Paris XIII (Prof. Gaston Gross). LADL is the central node of the network RELEX of laboratories involved in the construction of the same tools for other languages and using standardized software for natural language processing.

A. The linguistic data bases at LADL

1° A set of dictionaries defined according the complexity of their entry items:

- a lexicon of about 90,000 simple word entries for French (French DELAS),
- lexicons of about 60,000 simple word entries for English and Spanish.

Such lexicons (DELA systems) are morphological dictionaries, the inflectional properties of their entries are systematically encoded, so they can be automatically inflected. Thus, from the French DELAS, a lexicon of about 900,000 word forms with their grammatical attributes is derived (DELAF);

- a lexicon of phonetic transcriptions (DELAP), for the French DELA system,
- a lexicon of French compound words:
    - 7,000 compound adverbs (M. Gross 1990),
    - over 100,000 compound nouns.

2° Lexicon-grammar of French

Every verb has a specific set of arguments (i.e. subject and complements), to the point that often, this set is unique. Hence, the syntactic properties of verbs, or rather of the elementary sentences defined for each verb, have to be systematically described. No system predicting sentence forms from semantic features can be thought of. The

systematic description consists in matrices whose rows are verbs (i.e. elementary sentences) and columns are sentence forms into which verbs may enter or not. The sentence forms are the usual transforms of unary sentences, often simple declarative forms. Matrices are binary: a "+" sign appears at the intersection of a given row and a given column when the verb in the row enters the structure represented in the given column, a "-" sign appears in the opposite situation.

A lexicon of the 12,000 main verbs of French has been subdivided into about 50 classes (C. Leclère 1991). Each class has a specific matrix. The sentence forms are about 400, including various of pronominalization, passivization, sentential complement reductions, and nominalizations by support verbs.

A lexicon of 25,000 elementary sentences with at least one frozen argument. Their representation by binary matrices follows the same principles.

Partial lexicons of sentences with support verbs (*être, avoir, faire,* etc.*)* and predicative nouns have also been built (J. Giry-Schneider 1978, 1987, A. Meunier 1977).

3° Local grammars

Various sets of utterances are best described as variants of each other:

- a technical term, its synonyms and its abbreviations;
- a frozen sentence, with lexical variants in the frozen arguments;
- utterances belonging to a narrow and specialized technical language.

In such cases, finite automata can be shown to be remarkably adequate to the description of equivalent or related utterances.

Besides isolated families, full-fledge grammars have been built for domains such as: the expression of dates, durations, temperatures, of the language of Stock Market reports.

B. Algorithms for corpus processing

Corpus processing starts with a recognition procedure of words. Most natural language processing is based on a formal notion of word: a word is a sequence of characters found in a text and delimited by two consecutive separators (i.e. blanks and punctuation marks). A different notion of word is used in all applications developed at LADL: a word is a sequence of characters recognized, by a look-up procedure, that is, found in a dictionary. This approach has various consequences:

- precise information can be attached to any word in a dictionary, starting with its different meanings. Hence, many words become ambiguous, requiring procedures that eliminate ambiguities;
- many sequences of characters are not found in dictionaries: proper names, acronyms, numerical values, etc. To the extent it is practical, lexicons or grammars for these items should be built, even if they have been ignored in the domains of classical lexicography

and linguistics.

Applying large dictionaries to large corpora involves high performance algorithms. This subject is an ongoing research, whose theoretical aspects are developed at the Institut Gaspard Monge.

The main procedures in current use are the following:

- dictionaries have been compacted in the form of finite automata. For exemple, the dictionary DELAF has a size of 1.3 Mo. Different approaches to high speed access to dictionaries are studied;
- indexing of texts is being developed (J. Senellart 1996) to increase the speed of recognition of complex utterances (e.g. local grammars);
- INTEX, an integrated system of corpus analysis has been built by M. Silberztein 1993. This system lemmatizes texts and analyses contexts of words in order to solve ambiguities. The contexts that have been formalized and put to use are:
- dictionaries of compound words,
- local grammars,
- lexicon-grammars (E. Roche 1994).