Computational Linguistics and Language Learning

A. Zampolli, M. Sassi

AILA-UNESCO
Bruxelles, april 1987

A. Zampolli, M. Sassi

Computational Linguistics and CALTL[1]

We agree with Prof. Decoo to subdivide the topic of computer applications in language teaching in two parts. He will deal directly with the uses of computers in language teaching. I shall discuss here the methods, components and data which computational linguistics can offer as input or support to those uses.[2]

---

1 Paper presented at the *AILA-UNESCO Bruxelles* **Meeting on Technology for Language Teaching** *(april, 1987)*.

2 Furthermore, the experiences at my Institute, officially encharged with the development of computational linguistics in the framework of national researches in Italy, has shown that language teachers, and humanities teachers in general, as a consequence of the diffusion of computers promoted by the government in schools of different levels, are faced with the problem of finding goals and methods for computer uses, which seem to be more evident and obvious in natural sciences.
Teachers seem to ask not only for exercises, tests, etc., but also for different systems of personal interaction between students and linguistic and literary data.

a) **Text corpora**

Textual corpora in machine-readable form have been collected for several languages. They vary considerably along several parameters, namely:

- language: spoken/written; general/technical; etc.
- size, chronological limits
- subdivision in genders (novels, technical writings, newspapers, etc.)
- levels of linguistic analysis (lexical, syntactic, etc.) incorporated in the texts
- nature of the texts: complete texts; periods or sentences selected randomly, etc.

Some recent developments of computational linguistics and computer technology must be taken into account:

- Textual corpora tend to be structured in the form of textual data bases, intended for on-line use. Access methodologies offer several functions, often coordinated in an "interrogation language". It is possible for example to search in the entire corpus or in selected subsets a string of characters, a word, a lemma, a sequence of words, a word pattern, etc. The computer shows or prints the frequency of the searched elements, their distribution in the subsets of the corpus, their various contexts, etc.
- The technological developments make widely available to individuals, at rapidly decreasing costs, computational power (in the form of PCs), and storage capacity (in various forms: e.g. various types of disks) largely sufficient for very large corpora which, up until now, have required the connection with a centralized main-frame;
- Specialized softwares are available, which can be used to process linguistic data in connection with the functions of the interrogation language. For example, statistical packages can 'discover' syntagms or

collocations or, in general, patterns of high frequency in the corpus, which can be offered as objects of retrieval;

- The research in the corpus can be assisted by various knowledge sources which extend the power of the questions formulated by the human users. For example, a machine dictionary can search in the text all the inflected forms of a lemma, all the synonys, the hyponyms, the members of the same lexical field, etc. More or less ambitious parsers can search syntactic patterns, the arguments of a predicate, the modifiers of a noun, etc.
- For several reasons, including the support to basic research for machine-assisted translation, monolingual corpora tend to assume bilingual or multilingual dimensions. Those dimensions can be achieved at different levels: parallel corpora constructed with the same criteria (text-types, sizes, linguistic analysis) for the two languages; a couple formed by a corpus of texts and the corpus of their translations; bilingual contrasted corpora, i.e. a couple of corpora in which the linguistic elements of a text in one language are linked via explicit cross-references to the 'translational equivalents' in the translation.

Textual corpora can be used in the framework of language learning and teaching in various ways: to construct frequency dictionaries of linguistic units at various levels; to study the variation of those frequencies among various subsets and sublanguages; to extract quotations to be used as "examples" of linguistic properties, or as material for the construction of various types of texts and exercises; etc.

We wish to stress here an application, which exploits the specific characteristic of the textual DBs, and which has proved to be very promising, specially in the framework of advanced language learning.

The advanced learner can access the textual corpus by using all the tools and facilities described above, and "navigate" across the texts according to his personal needs and interests. Among others, he can obtain a direct personal appreciation of the "norm" (as intended by Coseriu) of usage of different linguistic units and constructs in different registers and sublanguages.

b) **Monolingual and multilingual multifunctional lexical DBs**

1. The recent strong interest in the lexicon is due to a number of factors:

a. Theoretical developments within linguistics are placing increasing emphasis on the lexical component. It is proving to be a central source of semantic as well as syntactic information.

b. Demonstrations of the feasibility of applications of natural language processing are creating demands for large-scale systems in industry and in national and supranational organizations. For these systems to be practical they must deal with tens and even hundreds of thousands of lexical items.

c. The effort required to create comprehensive dictionaries for these purposes is substantial; it may prove to be the most costly and time consuming task in such developments. Currently, each system is building its own lexicon, and there is increasing recognition that the duplication of effort is enormously expensive. However, differences in the organization and content of these lexicons make it difficult or impossible to share linguistically relevant information across systems.

d. The computational linguistics community is becoming increasingly conscious of the extensive resources contained in published dictionaries, and explorations are underway to determine how that information in machine-readable form can be exploited to expedite system development.

e. Publishers are beginning to realize the potential of their dictionaries for commercial purposes. They are recognizing the value of establishing lexical data bases from which a variety of dictionaries can be derived. They are also becoming aware of the breakdown of the distinction between different reference works (dictionaries, fact books, encyclopedias).

f. Increased communication among lexicologists, lexicographers, linguists, computational linguistics, publishers, and commercial natural language processing software developers has led to a heightened awareness of common objectives and the complementarity of skills and knowledge.

g. Initial experiments have given support to the idea that it may be possible to construct "neutral lexicons" that can be shared, with different theories selecting relevant linguistic information through an appropriate interface.

2. **Reusability of lexicographic information**

The preparation of a machine dictionary, and in particular of a LDB is one of the longest and most costly enterprises in the construction of NLP systems.

The number of researchers asking whether and how the various sources of lexical information already in existence are reusable, is increasing. The possible sources considered are printed dictionaries in MRF, terminological data banks, machine readable dictionaries for linguistic researches, computational lexicons for NLP systems.

The nature and extension of the reutilization process depend on several factors: nature of the source, its reliability, level of formalization and generality of the data. The transfer of the data from the original source

to a LDB is never straightforward. Several details must be taken into account, which may require 'ad hoc' decisions and manual intervention. Even in the simple case of the reutilization of a list of lemmas and their part of speech codes, different criteria can come into play for the identification of the lexical units (homonymy versus polisemy; autonomy of derived words; expressions written as single or separate graphical units; etc.) and for the number and the denomination of the morphosyntactic categories.

A more interesting case is the utilization of the grammatical information of the type associated to the lexical units in advanced learned dictionaries such as the LDCOE and the OALD.

Various groups are designing semiautomatic procedures which extract from the LDCOE the information on the syntactic properties of the lexical units which is requested by their linguistic theories and/or NLP systems.

Unfortunately, owing to the different status of teaching/learning of different languages as second languages, these types of dictionaries exist only for very few languages.

Some research groups are using substantially similar methods for the reusability of the semantic information conveyed by the definitions of the MRDs they have available

The researches carried out by these groups are based on the acknowledgement that the logic-linguistic structures of some types of definitions, already consolidated in the traditional lexicographical practice, implicitly represent some relevant semantic relations between the units of a lexical system.

Those semantic relations are expressed by a limited number of expressions.

The most studied relation is that of hyponymy, which structures large subsets of lexical units in conceptual taxonomies. The groups try to recognize, through pattern matching techniques, the definitions structured in 'genus proximum' and 'differentia specifica'. By exploiting the highly specific linguistic patterns which are used in the lexicographic tradition, they try to locate the term representing the 'genus'. All the lemmas whose definitions contain the same 'genus' term, are connected to the latter in a hyponymical relation. Various types of thesaurus-like relations are simulated in the lexicon through a structure of pointers among the lexical entries. The resulting conceptual taxonomy is directly relevant for several uses. The human user (a linguist or a lexicographer, an every-day student) may "browse" the lexicon not only through single words, but also through concepts and 'family' of concepts.

If - as is the case in our Institute - a textual DB is accessible with the 'help' of a MR dictionary, the query of the user can be expanded by searching in the texts not only the word he has used in the query, but all the related terms in the taxonomy.

By using appropriate procedures, linguistic information (selectional restrictions, semantic features, etc.) associated to a superordinated term can be inherited by its hyperonyms and used by parsers and generators.

Of particular interest seems the research which tries to identify the semantic relation which connects the derivative to its base, expressed by the definition of derived words.

One of the gaols is to formulate formal rules which express the syntactic and semantic modifications induced by the derivation on the syntactic-semantic properties of the base.

3. A Neutral LDB

We have promoted a working group which will involve outstanding representatives of the major current "linguistic schools". The group will investigate in detail the possibility of representing the linguistic information frequently used in parsers and generators (e.g. the major syntactic categories, subcategorization and complementation, verb classes, nominal taxonomies, etc.), in such a way that they can be reutilized in the following theoretical frameworks: government and binding, generalized phrase structure grammar, lexical functional grammar, relational grammar, systemic grammar, categorial grammar. This group will work on various languages. We shall start by examining in detail the treatment which the foregoing theories will assign to a representative sample of English and Italian verbs.

Let us suppose we are describing the Italian verbs by using the criteria, tests, formal apparatus of a given theory.

At the end, we shall subdivide the Italian verbs in classes, regrouping in a class all the verbs with the same description. We consider as members of a same class those verbs which have received the same description. The intuition we wish to prove is that the aforementioned theories will classify the Italian verbs substantially in the same way; in any case, the different theories will identify the same number of classes having the same members.

Each theory will of course describe the syntactic behaviour of a class using its own formal and explicative apparatus.

If this is true, it would be possible to label the verbs of the ILDB by distributing them into classes. The interface between the ILDB and a given theory and its relevant computational systems would thus contain the description of the syntactic behaviour of the different classes according to this theory.

This is of course only an abstract scheme, and we should envisage a number of strategies for its correct application.
However, we feel that this intuition is the same as stating that the properties taken into account by the different theories are in large part the same, although differently described and explained.

4. **Multilingual LDBs**

We conceive a bilingual MLDB as a complex structure consisting, essentially, of the following components:

a) A MLDB for the language L1, structured according to the aforementioned principles;

b) A MLDB for the language L2, structured according to the same or similar principles;

c) A bilingual "bridge" connecting the two monolingual MLDBs = i.e., a set of relations and conditions connecting their elements;

d) A textual DB containing, along with a reference corpus for L1 and L2, also a set of (so-called) "contrasted bilingual texts"; By this expression we intend a structure including a text in one language, its translation into another, plus a set of cross-references explicitly indicating the relations of (translational) equivalence between the corresponding elements of the two texts (n 15a);

e) A set of procedures and software tools for the access to the data, both for programs and human users, which will allow the retrieval operations to start from every pertinent point in the entire structure.

In our opinion, a bilingual MLDB must be considered as:

- a source of data for contrastive and comparative linguistic studies;
- a component in a workstation for assistance to translation (15b);

- a source of information for the construction of lexical components for NLP systems which require some type of transfer between languages, such as machine translation, involving the two langauges, such as the searching, in contrast and corpora, of possible translational equivalents for bilingual lexicography;
- a tool for computer-assisted langauge teaching and acquisition.

The information used to discriminate the different "meaning units" in the bilingual dictionaries, have the function of giving to the reader the elements needed for the transition of the language. For various reasons, including the saving of space, this information is expressed in an ambiguous or incomplete way. The effort made to interpret, disambiguate and translate them into formalized rules by using interactive procedures, will not only make it easier to map on the monolingual entries, but will also help to formulate conditions and constraints to be used in the lexical transfer in MT systems.

At the very end of this process, the information extracted from the bilingual MRD will be transformed into links mapping the "meaning units" of the bilingual MLDBs, and into constraints formally expressing the conditions of applications to the contexts of the texts to be translated.

As we have already said before, the "meaning units" of a multifunctional LDB will be connected within a network of various syntactic and semantic relations: synonymy, hyponymy, etc.

An interactive system will allow the user to "navigate" through the network starting from any entry and selecting the desired type of relations.

In the resulting MLDB, the user will access, starting from an entry in one of the two monolingual MLDBs, or from an element in the translation condition, a structure formed by three networks: the two monolingual

networks of syntactic-semantic relations, and the "bilingual" network linking, with "mapping relations" and "translation conditions", the two monolingual MLDBs.

The possibilities which are open to the human users seem to be obvious.

For example, the thesaurus-type of organization generated, in a monolingual MLDB, by representing the hyponymy relations, make it possible to access the bilingual MLDB not only through a single word, but also through a concept.

From a general point of view, the translator will be enabled to consider the correspondences not only from a word to its translation equivalents, but also from the "family of words" (e.g. a semantic field) in L1 to a "family" of words in L2.

This organization can also be useful to expand the information concerning the translation conditions ("semantic indicators" in various dictionaries) provided by a bilingual dictionary, to be reused in NLP systems involving two languages.

The "semantic indicator" is often a single word chosen to represent a "family" of words: e.g. a synonym, a hyponym, a typical subject or object of a verb, a typical noun of which an adjective can be predicated, etc. The monolingual MLDB can be used to expand this information given as a single word to the whole set of words to which it actually refers.

This expansion obviously implies previous disambiguation of the relation between the word given as "semantic indicator" and the set of words it represents. It must ne stressed that we are only at the beginning of this research, but the renewed interest for computer-assisetd machine-translation will certainly provide reusable data in the very near future.

5. **Possible uses in CALTL**

Various types of experiments have been conducted to explore possible uses of lexical data in machine-readable form for CALTL.
We quote the following examples:

- In connection with a morphological component, to create and test exercises on inflectional and derivational morphology;
- To create tests to assess the comprehension of the definitions, and, in general, of the linguistic information supplied by dictionaries oriented to language teaching and learning (see J. Miller);
- To create exercises and tests on individual aspects of lexical acquisition and competence;
- To allow the learner to "navigate" acrosse the lexicon of a language, following different grammatical and semantical realtions: derivation, synonymy, hyponymy, selectional restrictions, lexical fields, etc.;
- To retrieve exhaustive subsets of units sharing the same properties: parts of speech, suffixes, syntactical and semantical valencies, usage registers, technical sublanguages, etc. and, in general, to freely combine different dimensions to characterize subsets of units: parts of speech, rhymes, cognates, lexical fields, etc.
- To assist the learner in the exploration and use of bilingual lexicons: systematic treatment of collocations, idioms, and in general multiword units; to expand the research for the equivalent translations of a word through the network of semantic relations both in the source and in the target lexicon; etc.

c) **Formal models of different linguistic levels**

Systems for parsers and generation of sentences have always been the major chapter of computational linguistics, and include formalism for representing linguistic structures at different levels (morphological,

lexical, syntactical, semantical, etc.), rules to compute the structures, algorithms and strategies to apply those rules.

Several different trends can be recognized nowadays:

- Systems stressing the engineering aspect: in general, they deal with rather limited subsets of language, operate in very specific limited pragmatic domain, compute very deep linguistic, logic or pragmatic structures, and tend to reach a very robust, efficient performance, relying on specific procedural mechanisms and solutions.
- Systems whose interest is mainly concentrated on the formal properties of the grammar and the computational power of the algorithms. Experiments are conducted with different computational models, in order to comparatively assess their properties.
- Systems stressing the characteristics of their model as a tool for describing and exploring the characteristics of human cognitive components of the "faculty of language". These systems are usually strongly connected with the AI world, in particular with the field of knowledge representation languages
- Systems stressing the linguistic components. The computer is used in particular to collect, represent, retrieve large sets of linguistic rules, and to verify them with the projection on textual samples, or with the analysis and generation of sentences in interaction with the linguist.

This is the field which has more often interested the specialists of CALTL. Potential applications of those components and data seem very natural and direct.

However, very little has been done in this sense on a practical level, problably owing both to the extreme complexity, technicality, and specificity of the formal and computational models, and to the fact that

we are still unable to treat free samples of natural language in a completely automatic manner.

Let us quote three major types of applications:

- assistance to the teacher in the construction of syntactical exercises;
- automatic verification of exercises;
- study of linguistic theory rules; the student can interact with the parser or the generator during the analysis or synthesis of sentences, structures, parts of structures, and with the transfer component of MT systems, to evaluate contrastive aspects of language parsers.

ANTONIO ZAMPOLLI

L'elaboratore elettronico negli studi linguistici

Estratto dalla rivista IBM n. 2, maggio 1968