TOWARDS A POLYTHEORETICAL LEXICAL DATABASE

Preprints

Walker, A. Zampolli, N. Calzolari
(eds.)

ILC - 1987

Istituto di Linguistica Computazionale

C. N. R.

CONTENTS *

* The final version of the Proceedings will also contain the papers by M. Antona and N. Calzolari, B. Ingria, J. Kegl and B. Levin, A. Zaenen.

INTRODUCTION

1. The creation of machine dictionaries is one of the major activities of the Istituto di Linguistica Computazionale (ILC) of the Italian National Research Council. The ongoing projects concern classical and medieval Latin and contemporary Spanish and Italian. In particular, we aim at restructuring in the form of a multifunctional lexical database, eventually connected to lexical databases of other languages through the use of machine-readable bilingual dictionaries, the present Italian Machine Dictionary (DMI).

The DMI was contructed from 1968 to 1972 with the following goals:

- to allow the semiautomatic lemmatization of contemporary Italian texts, and possibly their automatic phonological transcription;
- to provide information to parsers and generators of Italian sentences;
- to constitute a computational representation of the Italian lexical system, permitting a large number of studies and researches on its quantitative and qualitative structures.

In its first version, the DMI was composed by a list of approximately 120,000 lemmas, from which about 1,000,000 forms can be automatically generated.

At a successive stage, thanks to a contribution given to the University of Pisa by the Italian House of Deputies (House of Commons, Parliament), about 200,000 definitions were added in machine-readable form.

From the very beginning, we were forced to ask ourselves the question whether it was inevitable that the use of the DMI would be limited to computational systems encompassing a specific linguistic theory, or whether it was possible to conceive it in more general terms, "neutral" with respect to its utilisation in various linguistic theories.

Even the different researchers, using the DMI to semiautomatically lemmatize Italian texts, employ different sets of linguistic criteria to identify and classify the lexical units occurring in the texts to be processed for the creation of concordances, frequencies, lexicographical documentation, etc.

We have structured the DMI and the morphological look-up so that each researcher can specify certain options and thus automatically generate from our DMI a lemmatization system which is compatible with his personal criteria.

Clearly, the construction of a "neutral" lexicon becomes increasingly complex now that we have to insert into the DMI linguistic information at the syntactic and semantic level.

2. This problem may well be of a more general interest within the recent framework of the increasing attention which is being paid to lexicology and lexicography in linguistics and computational linguistics.

A large amount of work in computational linguistics (CL) is carried out on experimental lines, with consequently small-sized lexical prototype systems. Furthermore, emphasis is traditionally placed on the representation, organization and use

of linguistic knowledge as encapsulated and expressed by linguistic rules and procedures. Lexical data seem to be considered of secondary importance or, at least, easy to handle. At a recent workshop (UMIST 1985), an informal poll - constituting a good representative sample of today's computational linguists - was conducted among the invited speakers, to establish the average size of the lexicons used by their systems. With the exception of a prototype MT system, the average size was about 25 words. This observation may be extended to a large number of systems within the computational linguistics paradigm; MT systems being the only eventual exceptions.

Today we are forced to consider the following facts:
a) Our CL community has been recently faced with the request of large scale NLP systems, owing to the recent advances in CL technology which make such applications feasible, and the interest expressed by national and international public and private organizations.
b) For real world applications, it is of fundamental importance that a CL system is able to deal with tens of thousands of lexical items. The projects presently underway must be accomplished within reasonably fixed time limits. The preparation of dictionaries can be delayed no longer.
c) Various projects have been promoted for the same language. Up until now, it has been a fact that each system has its own ideas and conventions with regard to content, organization and structure of its lexicon. This makes it difficult or even impossible to share linguistically relevant information, across various NLP systems, for the same language.
d) Duplication of efforts may be a very "sad" fact. Building a comprehensive consistent NLP dictionary is probably the most costly and time consuming task in every NLP project. In this situation, it is natural that not only researchers and developers, but also the promoting and financing authorities should put forward the question as to whether it is possible to design a rich, powerful and flexible LDB, where different linguistic theories and CL systems can find the relevant lexical information required.

3. In the past two years, a number of initiatives at the international level have clearly shown that the creation of multifunctional lexical databases, the reusability of existing lexicographical resources, and cooperation among research groups, industry and publishing houses are considered to be issues of primary importance in an effective program of research and development on multilingual lexicology and lexicography.

In particular, a workshop "On automating the lexicon" was organised by N. Calzolari, L. Rolling, J. Sager, D. Walker, A. Zampolli in Grosseto (19-23 May 1986), sponsored by the European Community, the University of Pisa and the Institute of Computational Linguistics of the Italian National Research Council. It was also held under the auspices of the Association for Computational Linguistics, the Association for Literary and Linguistic Computing, EURALEX, and the AILA Commission on Computational Linguistics. The second in a series, it built on

its predecessor, which was held in April 1983 at SRI International, Menlo Park, California.

Where the first workshop examined the machine-readable dictionary from the perspective of the research community, the publishers and the emerging market intermediaries, the second was much broader. Its purpose was to explore research efforts, current practice, and potential developments in work on the lexicon, machine-readable dictionaries, and lexical knowledge bases with special consideration for the problems created by working with different languages. The intent was to identify the current state of affairs and to recommend directions for future activities.

To help in the realization of these objectives, a set of papers was solicited for the workshop under the following general headings: Research Areas, Core Problems, Application Areas, and Developing Research Resources. People were asked to prepare comprehensive surveys and evaluations of activities underway in the field. We also requested reports on national projects in related areas. At the end of the agenda a "Consolidation" session was scheduled to consider the following topics: the lexical entry as a basis for integration, cooperation and communication, priorities for research and development, and next steps. The workshop papers are in print (D. Walker, A. Zampolli, N. Calzolari, eds.).

The participants were chosen to bring together representatives from the different kinds of areas that we believed were relevant to the various problems associated with the lexicon. This led us to invite linguists, lexicographers, lexicologists, computational linguists, artificial intelligence specialists, cognitive scientists, publishers, lexical software marketers, translators, funding agency representatives, and professional society representatives.

One of the recommendations formulated at the conclusion of the Grosseto Workshop and circulated at the New York Workshop ("The Lexical Entry", July 1986) affirms that it is necessary to "compare and contrast lexical information, particularly in the form of 'lexical entries', as reflected in logical and linguistic theories, computational linguistic systems, machine-readable dictionaries, translation activities, and lexicographic practice in order to identify dimensions of similarities and differences; based on those dimensions, create a metaformat that subsumes the structure of the various types of information to be included and that can be used both as a reference frame for evaluation and exchanges and as a model of a computerized metalexicon from which lexicons for different research and applications may be derived".

Encouraged by the results of the Workshop, and under the urgent pressure of defining the content and representation of the semantic and syntactic information in the DMI, at the end of the Grosseto Workshop A. Zampolli invited B. Boguraev (Cambridge), B. Ingria (BBN) and F. Uszkoreit (Stanford) to form a working group to explore the feasibility of representing the lexical information and, in particular, the information frequently used in parsers and generators, so that it can be reutilized in various theoretical frameworks, and in different applications both by automatic procedures and human users.

Zampolli has presented this idea in other meetings; in particular, at the end of the New York Workshop "The lexical entry" (July 1986), and at the COLING86 panel discussion on multilingual lexical resources.

On these occasions, other researchers joined the group which met for the first time in Pisa (18-22 March 1987). At this first meeting, a brief exposition of the different theories was followed by a comparison and discussion of the ways in which a list of verbs, distributed in advance, was categorized according to each theory.

The syntactic requirements of each linguistic theory on the lexical entries were taken into account in order to formulate a list of syntactic patterns which could be directly accepted or indirectly translated into each different theory. This was done taking into account the more fine-grained distinctions, on which a consensus was reached. These subcategorization frames should, in a sense, represent the "facts" or the extensional classes which are the natural basis which each theory, at the end, has to describe. They correspond to the traditional constructions and are characterized in common words. They will then be mapped into the different theoretical frameworks by means of different set of "translation rules". A first provisional template was drawn up with a list of the types of information which have to be coded in the Lexical Entry.

The following linguistic theories were taken into account: GB, CFG, GPSG, lexical functional grammar, dependency grammar, systemic grammar.

Other important topics such as participant or thematic roles, selectional restrictions, and meaning-form relation (e.g. in transitivity alternations) have been left for discussion later on.

The group should meet again soon at Stanford. We hope that these preprints are useful as background/preparatory documents for this meeting.

The work of the group will also take into account the data offered by a large corpus of English.

Antonio Zampolli
Nicoletta Calzolari