Introduction to the Special Section on Machine Translation

ANTONIO ZAMPOLLI
Instituto di Linguistica Computazionale, Pisa, Italy

Foreword

When the use of electronic data processing techniques on linguistic data began, two lines of research were, quite independently, activated:

- Machine translation (MT).
- Textual analysis (French: dépouillement (LTA); production of indices, concordances, frequency counts, etc.).

While MT was promoted mainly in 'hard-science' departments, LTA was developed mainly in humanities departments and, probably for this reason, the two lines had very little contact.

At the beginning of the 1960s, the perception of a possible reciprocal interest was explicitly manifested, in particular through the invitation of MT researchers to LTA conferences, like Tübingen (1960) and Besançon (1961).

The topics quoted were, in particular, text encoding systems for different alphabets; detection of frequency of linguistic elements in large corpora; and automated dictionaries. But, in effect, real cooperation was very rare if not totally absent.

The year 1966 was particularly important for both lines of research, but for opposing reasons.

The Prague International Conference 'Les machines dans la linguistique' enlisted the international acceptance of the LTA as an autonomous disciplinary field, and its extension to a broader area (which included new dimensions of processing - phonology, historical linguistics, dialectology, etc., and called literary and linguistic computing - LCC, whereas the ALPAC report (see Automatic Language ... 1966) brought about an abrupt arrest in the majority of MT projects throughout the world and the beginning of the so-called 'dark ages' of MT.

Following, de facto, the recommendations of the ALPAC report, basic research on natural language processing occupied the area characterized so far by MT activities, and computational linguistics emerged as a new disciplinary activity (CL). 1

In spite of ALPAC statements, CL focused mainly on the development of methods for the utilization of linguistic models - in particular formal grammars - in the analysis and generation of isolated sentences, in an almost exclusively monolingual framework. A distorted (I believe) interpretation of the Chomskyan paradigm, led to an almost complete disinterest in corpus analysis and quantitative data, which, on the other hand, were attracting much attention at that moment in the LTA field, due, among other things, to the projects for national historical dictionaries and for frequency dictionaries.

On the other hand, the LCC delayed taking advantage of the know-how, methodology, and tools produced from the very beginning by MT in the field of automatic texts. MT not only developed research on specialized hardware, storage, access techniques, artificial and derivational morphological analysis, but certain projects had already begun the collection of large sets of monolingual and bilingual textual and terminological data.

Very few exceptions can be reported in the LCC field, all primarily motivated by attempts to automate the lemmatization of texts for the production of lemmatized indices and concordances. To my knowledge, the first experiments are related to Latin (CAAL, Galleratz and LASLA, Liége). These two systems were presented and compared at the Pisa 1968 meeting 'Le lessico elettronico latino', during which was also presented the first proposal for a multifunctional lexicon, DMR: Italian machine dictionary conceived not only for lemmatization, but also as a repository of textual knowledge both for computer programs (parson, generators, phonological transcription, etc.) and human uses (qualitative and quantitative researches on the structure of the Italian lexicon).

The CL activities which came after MT, almost completely neglected the development of large lexica, practically resuming to small toy-lexicons of a few dozen words.

For several years the problem of the relationship between LCC and CL was practically ignored.

As local organizer of the 1973 Pisa COLING, I endeavored to include in the call for papers, and to promote in the Conference, sections explicitly dedicated to topics which could determine the area of common interest.

The attempt was successful in terms of joint participation, and it was probably not just by chance that J. Smith presented there, as an international level, the newly founded ALLC (Smith, 1973).

But in those years a (so to speak) 'puristic' approach characterized the general reflections of CL, which was searching for definitions and a disciplinary identity.

It can be interesting in this respect to read the Foreword of H. Karlgreen, chairman of the Scientific Commitee, and in my Introduction to the Proceedings of COLING 1973 (Zampolli and Caltolani, 1973).

The situation has changed only in the last two years. A variety of constraint factors have contributed to finally establishing increasing contacts between LCC and CL.

The awareness of the several relevant areas of common interest and needs is gaining ground on both sides. Some
cooperative projects are jointly formulated at an international level.

This convergence is, partly, the result of the activities of some Institutes whose activities programmatically and institutionally cover both fields, but above all it is aided by a new framework. Supranational Organizations and International Associations are paying increasing attention to the potentiality of the so-called language industries at industrial, social, and cultural levels. This term designates a variety of practical applications of computational systems embedding components for natural language processing: office automation, full text information retrieval, man-machine communication, speech analysis and synthesis, and, of course, MT, etc. The development of language industries requires the development of an adequate language technology, which should permit the construction of the necessary LE components. It is of crucial importance that the nature of this development requires the convergence of know-how and experience, developed both by LLC and CL, and the creation of resources, methods, tools which are relevant for both. The creation of multifunctional large monolingual and bilingual lexical knowledge bases, reusable in a variety of applications, and the collection of large linguistically annotated corpora, for the study of the qualitative and quantitative characteristics of various sublanguages and specific domains, are priority tasks.

ACL, ACM, and AILLC have not only jointly organized panel discussions within their respective international conferences to discuss relationships and possibilities of cooperation, but are also jointly sponsoring international projects for the creation of the above-mentioned linguistic resources, and in particular corpora, lexicon, and encoding standards. The inclusion of a section dedicated to various aspects of current MT projects within this journal fits into this framework.

We are planning a second issue dedicated to lexical knowledge bases, which seems to be an area in which LLC, CL, and humanities computing will naturally cooperate in the immediate future, because of the central role of lexical knowledge both in automatic and computer-assisted activities in all three fields.

Notes

1. In fact, the first experiments of computerized and indices production were performed not with electronic machines, but with "punched card electrical accounting machines" (Bosu 1951, 22).


3. In the Introduction to the "Acce de Colloque Internationale sur la Metropole des Recherches Lexicologiques" held in 1964 in Besancon, B. Quenault says: "Un des buts de ce Colloque sera Plato de mettre en contact des chercheurs qui sans se gever tout a fait, s'engage en d'informations alors qu'ils travaillent sur une matiere commune: la langue, et plus particulierement, le langage des diverses disciplines. Nous avons la chance d'accueillir le ciel des lexicographes et des lexicographes Français et Etrangers, des specialistes de la traduction automatique (vocabulaire de base, termologies scientifiques, specialises, dictionnaires automatiques, lexiques, synthese) de la traduction Artificielle (..) de la documentation automatique (..) de la pedagogie des langues vivantes."

4. M. Kay (1964), reporting on a formal meeting on "Formal Machines Readable Text" at the end of the IBM-sponsored Library and Processing Conference (Yokohama Heights, 1964), and in an article in the 56th issue of the Computer and Humanities (Kay, 1965), explicitly discussed the common interest of AT and humanities researchers on this topic. In the same issue, only two MT projects are reported in the Directory of Scholarly Activity, of a total of 120 projects in the section Language and Literature, both directed by well-known linguists, R. Poole and W. F. Lehman. But not, I think, inspired by it.

6. The Chairman of the Committee on Science and Public Policy, in a letter to the President of the National Academy of Science, stated "the support needs for computational linguistics are distinct from automatic language translation" (ALPAC, 26). And as page 35, one reads: "work toward machine translation, together with computational linguistics work that has grown out of it."

7. We quote from the recommendation: "Small scale experiments and work with machine models of language have proven seriously deceptive in the past, and one can come to grips with real problems only Shows a certain scale program size, dictionary size, and available corpora" (ALPAC, p. 44). 

8. See the Proceedings of the Tableau Rendu sur les Grands Dictionnaires Historiques (Ferrara, 1973).

9. See, for example, the series of frequency dictionaries of romance languages of Jullif, published by Mouton in 1961 (Spanish), 1965 (Romanian), 1970 (French), 1972 (Italian).

10. See, for example, the optical disk developed by IBM as a storage medium for bilingual dictionaries.

11. The Gallician Latin machine dictionary was made up of an alphabetical list of forms, progressively accumulated from processing the texts of St Thomas Aquinas. The Latin Dictionary was based on a list of stems, extracted from the Pericope Eneasian, and an associated morphological analyzer (see Bass, 1964).

12. This situation is still true today. A recent workshop on linguistic theory and computer applications (Weishopt et al., 1987) reports an informal poll in which the average size of the lexicon used by the prototypes discussed, the average size was about 25 (words) (Riegler and Brinse, 1989, 40). 


14. For example, the Institut de Linguistique Computationnelle, Pau (Zampolli, 1983), the Institut fur Deutsche Sprache, Mannheim, Sorbiskat, Gimpel, etc.

15. "...computer systems will undoubtedly enter every corner of future society. When that day arrives, the most
important technology will be specifically concerned ... with ... information/writing. In other words, the central problem will be how the informational events seen by human beings will be mechanically processed, transmitted, stored, and then retrieved in a form which can be interpreted by other human beings. Linguistic information and the techniques for processing it will be at the heart of the information society. Such technology might be called language engineering, and the industry which it will spawn will be the language industry' (Nagao 1989:4). See also Walker and Zampoli (1990).

16 As examples we can quote the ACL-sponsored sessions at the 1980 AAAI/ACM Conference in Toronto: "The Use of the Lexicon in Humanistic Research", "Computational Linguistics and Humanistic Research", and a similar panel discussion at the 27th Annual Meeting of the ACL, Vancouver, June 1980.

17 Significant examples are the following projects: Text Encoding Initiative. An international project proposed by ACL, ACH, ALLE and sponsored by NEH and EEC, which aims at developing guidelines for encoding standards and for the creation of text and dictionaries, in order to facilitate exchanges and further cooperation in humanities and in language industries.

Date Collection Initiative (sponsored by ACL). The initial goal is no lesser than creating at least 100 million English words in machine readable form. The project will be coordinated by the Council of Europe, which operates on corpora of English, Italian, German, French, Spanish, Swedish, Dutch, Greek, and Latin. D. Walker and A. Zampoli are preparing the survey of textual and lexical resources in machine-readable form, sponsored by ACH, ACL, ALLE, EURALEX, eventually cooperating with the Centre for Machine-Readable Texts in the Humanities, in order to facilitate the exchanges for which a grant has been approved at Rutgers and Palermo.

18 The most notable among MT translation projects is probably EUROTRA, the international cooperative initiative promoted by the EEC among the member countries, which aims at the creation of a transnational prototype of a multilingual translation system between the major official languages of the EEC. We cannot include an article on EUROTRA, because it is a general overview, written by B. Moogard, who has been already published in vol. 3, no. 2 (1988) issue of this journal.

Bibliography


Walker, D. and Zampoli, A., Foreword by both, in Bolognesi and Briscoe (eds.), 81--84.

