Introduction to the Special Section on Machine Translation

ANTONIO ZAMPOLLI
Istituto di Linguistica Computazionale, Pisa, Italy

Foreword

When the use of electronic data processing techniques[1] on linguistic data began, two lines of research were, quite independently, activated:

- Machine translation (MT).
- Lexical text 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 few contacts.[2]

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).[3]

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.[4]

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

The Prague International Conference 'Les machines dans la linguistique' ratified 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 – LLC), 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*,[5] 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).[6]

In spite of ALPAC statements,[7] 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 corpora analysis and quantitative data, which, on the other hand, were attracting much attention at that moment in the LLC area, due, among other things, to projects for national historical dictionaries[8] and for frequency dictionaries.[9]

On the other hand, the LLC delayed taking advantage of the know-how, methodology, and tools produced from the very beginning by MT in the field of automatic lexica. MT not only developed research on specialized hardware,[10] storage, access techniques, inflectional and derivational morphological analysis, but certain projects had already begun the collection of large sets of monolingual and bilingual lexical and terminological data.

Very few exceptions can be reported in the LLC field, all primarily motivated by attempts to automatize the lemmatization of texts for the production of lemmatized indices and concordances. To my knowledge, the first experiments are related to Latin (CAAL, Gallarate and LASLA, Liège). These two systems[11] were presented and compared at the Pisa 1968 meeting 'De lexico electronico latino', during which was also presented the first proposal for a multifunctional lexicon, DMI: Italian machine dictionary conceived not only for lemmatization, but also as a repository of lexical knowledge both for computer programs (parsers, 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 restricted to small toy-lexicons of a few dozen words.[12]

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

As local organizer of the 1973 Pisa COLING, I endeavoured to include in the call for papers, and to promote in the Conference, sections explicitly dedicated to topics which could delineate 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, at 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 a definition and a disciplinary identity.[13]

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

The situation has changed only in the last two years. A variety of concurrent factors have contributed to finally establishing increasing contacts between LLC and CL. The awareness of the several relevant areas of common interest and needs is gaining ground on both sides. Some

Correspondence: Professor A. Zampolli, Istituto di Linguistica Computazionale (ILC), C.N.R., via della Faggiola 32, I-56100 Pisa, Italy.

Literary and Linguistic Computing, Vol. 4, No. 3, 1989

cooperative projects are jointly formulated at an international level.

This convergence is, partly, the result of the activities of some Institutes[14] 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,[15] which should permit the construction of the necessary NLP 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, ACH, and ALLC have not only jointly organized panel discussions within their respective international conferences to discuss relationships and possibilities of cooperation,[16] but are jointly sponsoring international projects for the creation of the above-mentioned linguistic resources, and in particular corpora, lexicon, and encoding standards.[17]

The inclusion of a section dedicated to various aspects of current MT projects within this journal fits into this framework.[18]

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 concordances and indices production were performed not with 'electronic machines', but with 'punched card electrical accounting machines' (Busa (1951), 22).
2. For the history of the first years of MT, see Locke and Booth (1955), 1–23; Booth, Cleave and Brandwood (1958), 1–7; Vauquois (1975), 14–32; Nagao (1988).
3. In the Introduction to the 'Actes du Colloque International sur la Mécanisation des Recherches Lexicologiques' held in 1961 in Besançon, B. Quemada says: 'Un des buts de ce Colloque sera aussi de mettre en contact des chercheurs qui sans s'ignorer tout à fait, n'échangent guère d'informations alors qu'ils travaillent sur une matière commune: la langue, et plus particulièrement, le lexique dans diverses disciplines. Nous avons la chance d'accueillir ici à côté des lexicologues et des lexicographes français et étrangers, des spécialistes de la traduction automatique (vocabulaire de base, terminologies scientifiques, speciales, dictionnaires automatiques, homographes, synonymes) de la traduction "artisanale" (...) de la documentation automatique (...) de la pedagogie des langues vivantes.' And R. Busa, in an article with a very significant title (given the period) 'L'analisi linguistica nell'evoluzione mondiale dei mezzi di informazione', in a debate on 'the two cultures: the fracture between sciences and humanities', says that 'the development of linguistic automation is triangular: lexical analysis, information retrieval, mechanical translation', Busa (1961), 117.
4. M. Kay (Kay, 1964), reporting on an informal meeting on Formats for Machine Readable Text at the end of the IBM-sponsored Literary Data Processing Conference (Yorktown Heights, 1964), and in an article in the fifth issue of the *Computers and Humanities* (Kay, 1967), explicitly stressed the common interest of MT and humanities researchers on this topic. In the same issue, only two MT projects are reported in the Directory of Scholars Active, of a total of 120 projects in the section Language and Literature, both directed by well-known linguists, B. Pottier and W. P. Lehmann.
5. 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*, 2). And at page 29, 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 miniature models of language have proven seriously deceptive in the past, and one can come to grips with real problems only above a certain scale of grammar size, dictionary size, and available corpora' (*ALPAC*, p. iv).
8. See the *Proceedings of the Table Ronde sur les Grandes Dictionnaires Historiques* (Firenze, 1973).
9. See, for example, the series of frequency dictionaries of romance languages of Juilland, published by Mouton in 1961 (Spanish), 1965 (Romanian), 1970 (French), 1973 (Italian).
10. See, for example, the optical disk developed by IBM as a storage medium for bilingual dictionaries.
11. The Gallarate Latin machine dictionary was made up of an alphabetical list of forms, progressively accumulated from processing the texts of St Thomas Aquinus. The *Liège Dictionary* was based on a list of stems, extracted from the Forcellini lemmas, and an associated morphological analyser (see Busa, 1968).
12. This situation is still true today, 'A recent workshop on linguistic theory and computer applications (Withelock *et al.*, 1987) reports an informal poll to establish the average size of the lexicon used by the prototypes dicussed... the average size was about 25 (words)' (Boguraev and Briscoe, (1989), 10).
13. The article, 'The Field and Scope of Computational Linguistics', of D. Hays in the *Proceedings of the Budapest COLING 1971* is particularly relevant, and it is interesting to observe the evolution towards a 'puristic definition' of CL in the opinion of the author in respect to his chapter on 'computational linguistics' in the *Encyclopaedia of Linguistics, Information and Control* (1969).
14. For example, the Istituto di Linguistica Computazionale, Pisa (Zampolli, 1983), the Institut für Deutsche Sprache, Mannheim, Språkdata, Göteborg; 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 ... informationware. In other words, the central problem will be how the informational signals sent by human beings will be mechanically processed, transmitted, stored, and then recalled 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 span will be the language industry' (Nagao (1989) 4). See also Walker and Zampolli (1989).

16. As examples we can quote the ACL-sponsored sessions at the joint ALLC/ACM Conference, June 1989, 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 1989.
17. Significant examples are the following projects:
    Text Encoding Initiative. An international project promoted by ACL, ACH, ALLC and sponsored by NEH and EEC, which aims at developing guide-lines for encoding and standards for exchanging a broad range of different classes of texts and dictionaries, to facilitate exchanges and further cooperation in humanities and in language industries.
    Data Collection Initiative (sponsored by ACL). The initial goal is to acquire at least 100 million English words in machine readable form. The project will cooperate with the group established by the Council of Europe, which operates on corpora of English, Italian, German, French, Spanish, Swedish, Serbo-Croatian. In this area, D. Walker and A. Zampolli are promoting a survey of textual and lexical resources in machine readable form, sponsored by ACH, ACL, ALLC, EURALEX, eventually cooperating with the 'Center for Machine-Readable Texts in the Humanities', to study the feasibility for which a grant has been awarded to Rutgers and Princeton.
18. The most talked about MT translation project is probably EUROTRA, the international cooperative initiative promoted by the EEC among the member countries, which aims at the creation of a preoperational prototype of a multilingual translation system between the nine official languages of the EEC. We have not included an article on EUROTRA, because a general overview, by B. Maegaard, has been already published in Vol. 3, no. 2 (1988) issue of this journal.

Bibliography

Actes du Colloque International sur la Mécanisation des Recherches Lexicologiques, Besançon 1961. *Cahiers de Lexicologie*, 3, 1961.

Almanacco Letterario Bompiani 1961. Milano, 1961.

Automatic Language Processing Advisory Committee. Language and Machine-Computers in Translation and Linguistics. Washington, 1966.

Boguraev, B. and Briscoe, T. (1989). *Computational Lexicography for Natural Language Processing*. Longman.

Booth, A. D., Cleave, J. P., and Brandwood, B. A. (1958). *Mechanical Resolution of Linguistic Problems*. London.

Busa, R. (1951). *Sancti Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum*. Milano.

—— (1961). 'L'evoluzione linguistica dei mezzi di informazione, Almanacco ...', 103–17.

Busa, R. (ed.) (1968). *Actes du Seminaire International sur le dictionnaire latin de machine, Calcolo*. Supplemento n. 2 al vol. v.

Hays, D. G. (1969). 'Computational Linguistics: Introduction', Meetham and Hudson (eds.). 49–51.

—— (1976). 'The Field and Scope of Computational Linguistics'. Papp and Szepe (eds.). 21–6.

Locke, W. N. and Booth, A. D. (1955). *Machine Translation of Languages*. MIT Press.

Maegaard, B. (1988). 'EUROTRA, The Machine Translation Project of the European Communities', *Literary and Linguistic Computing*, 3, no. 2, 61–5.

Meetham, A. R. and Hudson, R. A. (1969). *Encyclopaedia of Linguistics, Information and Control*. Pergamon Press.

Nagao, M. (1989). *Machine Translation—How Far Can It Go?* OUP.

Papp, F. and Szepe, G. (eds.) (1976). 'Papers in Computational Linguistics', *Proceedings of the 3rd International Meeting on Computational Linguistics*. Budapest.

Quemada, B. (1961). 'Introduction, Actes du Colloque ...', 13–18.

Smith, J. (1973) 'Ideals Versus Practicalities in Linguistic Data Processing', in Zampolli and Calzolari (eds.), v. ii. 2, 895–8.

Table Ronde sur les grandes dictionnaires historiques. Florence, 1972.

Vauquois, B. (1975). *La Traduction automatique à Grenoble*. Paris.

Walker, D. and Zampolli, A., Foreword, in Boguraev and Briscoe (eds.), pp.xiii–xiv.

Whitelock, P., Wood, M., Somers, H., Johnson, R., and Bennett, P. (eds.). (1987). *Linguistic Theory and Computer Applications*. Academic Press: New York.

Zampolli, A. (1968). 'Projet pour un lexique electronique de l'italien', in Busa (ed.), 109–26.

Zampolli, A. and Calzolari, N. (eds.) (1973–7). 'Computational and Mathematical Linguistics', *Proceedings of the International Conference on Computational Linguistics 1973*. Firenze.