AILA Scientific Commission for Applied Computational Linguistics
Report on Activities and Future Prospects

Antonio Zampolli

Chairman of the AILA Commission on
Applied Computational Linguistics

In the two meetings of the Commission held during the AILA International Conference in Brussels, the following points were discussed:
- directions for future scientific activities
- participation in the AILA Conference to be held in Sidney
- relationship with the other AILA Commissions.

I. Scientific Activity

The relationship between ACL and other "adjacent" subfields of computational linguistics (theory of linguistic computation, computational psycholinguistics, computation in service to linguistics, text processing, etc.) has been a frequent source of discussion (cf. e.g. Zampolli 1973, Karlgren 1973, Thompson 1983).

For certain important aspects, the approach of the North-Americans differs considerably from that of the Europeans and Japanese.

Nevertheless, it is precisely in ACL that the most obvious prospects for the integration of the various activities emerge. In fact, the realization of specific applied tasks requires the activation of a number of knowledge sources and a collaboration between all those techniques and tools, produced in the various subfields of computational linguistics, which can be used to achieve the purpose. In this way, certain tasks require the joint utilization of statistical information obtained from the quantitative analysis of corpora of large dimensions, of theoretical models of competence and execution, of mono and multilingual dictionaries and lexicons, of generative and parsing algorithms, of AI techniques and methodologies, etc.

Furthermore, today, a number of international scientific associations are operating in applied sectors which effectively require the utilization of theories, methods and tools of computational linguistics (e.g. ALLC, ACH, ACL, EUROLEX, TERMIA, etc.)

It is easy to observe that there is a natural overlapping in the membership of these Associations. In fact, the multi-disciplinary character of computational linguistics has been evident from the very beginning.

It thus appeared important to choose lines of activity for the immediate future which:
- allow the specific characterization of the activities of the Commission;
- stress relationships with other AILA Commissions;
- coincide with themes and problems of great actual interest in today's so-called "information society";
- respond to the normal expectations not only of scholars and research workers, but also of those "developer communities" which are now evolving around the rapid diffusion of new technological tools (telematics, personal and home computers, etc.)
- are compatible with the actual knowhow and interests of the members of the AILA Commission.

On the basis of these points, the chairman proposed the following sectors of activity:

I.1. Computer Assisted Language Instruction and Learning

The chairman reported that Professor Decoo, who had not been present at the last meeting, had accepted his invitation to collaborate with him in this subsector. The chairman also reported on the contacts which had been made with CALICO (Computer Assisted Language Instruction Consortium). An article of the President of this association will be published in the AILA Bulletin, and the possibility of a joint sponsorship of congresses, schools and other activities will be examined.

Two major possible trends for activity were generally recognized:

a) the study and development of systems and programs explicitly designed for language teaching;

b) the identification and development of the possibility of applying didactically knowhow, methods and tools, which are produced in the Computational Linguistics area without specific reference or finalization for teaching or learning. For example:
- the techniques for dynamic multiple access to the treasury of data contained in mono- and bilingual dictionaries compared with the traditional method of static access to alphabetically ordered data. (See Zampolli et al. in this Bulletin);
- the utilization of morphologic and syntactic components of natural language processing systems for the production of language exercises;
- the access to large corpora of texts, representative of a language, through on-line interactive searching systems, as an aid in becoming familiar with the frequency and norm of usage of particular features or of different registers of the language.

Particular attention will be given to the search and organization of suitable material for the demonstrations at the Congress in Sydney.

I.2. Natural Language Processing Systems (NLP)

The situation in this sector is to some extent confused. On the one hand we have linguistic levels for which the construction of fairly generalized components for language analysis or generation could be hypothesized. For example, some experts think that we could even predict, for the near future, the construction of "a chart parser for GPSG on chip, which you could buy together with a reasonably extensive grammar to plug" into your NLP system. On the other hand, it is not clear whether a satisfactory domain-independent semanticis can be supplied.

Discussions are underway to decide which of the commonly proposed applied tasks in the field of NLP can actually be realized, or deserve to continue a research and development effort, in consideration of the present state of knowledge.

For example, some scholars are sceptical whether natural language is the best vehicle for computationally naive access to databases. "A well-designed interactive system based on a high-resolution display, menus and a printing device would out-perform a natural language interface for most applications" (Thompson, 1983).

Another example concerns the relationship between language translation and the computer. After the negative period following the ALPAC report, this sector is now undergoing a new revival. Side by side with commercial systems such as SYSTRAN, Logos and ALPS, research and development projects are flourishing. Mention can be made of the Japanese national project and, in particular, of the EUROTRA project, promoted by the Commission of the European Community, which involves a number of European researchers.

Discussions between supporters of three different approaches are extremely lively in this sector (automatic translation, computer assisted human translation, and human assisted computer translation).

Significant resources and considerable funds are available also for speech production and understanding, partly thanks to the European ESPRIT project. The same may be said for the treatment of natural language in office automation systems. Procedures ranging from spelling checkers to grammar and style correction have been launched.

The ACL Commission intends to promote a number of surveys on existing tools and components for computational linguistics, and to attempt to identify common features among the different sub-fields mentioned above. Prof. Geens has agreed to collaborate with the chairman in this subsector.

I.3. Computational Lexicography and Lexicology

The following types of activity can be identified in this subsector:

a) Text Processing to produce Indexes, Concordances, etc.

This type of analysis can be almost considered as a routine activity today, at least in specialized Centres.

The following types of action can be undertaken:

- coordination of those operating in this sector, so as to avoid duplicates and in order to guarantee collaboration in the acquisition of corpora in machine readable form.
- the creation of a framework (metalanguage, minimal norms, etc.) within which projects, programs, corpora, and lexical analyses can be clearly defined so that the results and data of different Centres can be compared and exchanged.

b) Dictionary Production

The computer can be employed in three separate and fundamental stages:

(i) Acquisition of the necessary lexical documentation: indexes, concordances, citation slips, etc.

The observations of point a) remain valid for this stage.

(ii) Computer Assisted Editing of Lexical Entries.

This mainly consists in the creation of workstations to access different knowledge sources (pre-existing dictionaries, textual archives, monographic documentation, etc.) with which the lexicographer can interact in the editing of lexical entries. This is a typical sector for research and development.

The first steps which should be undertaken in this sector seem to be:

- to encourage the exchange of ideas and experiences between lexicographers from the academic world and those working in industry. At the present, these two groups are examining the problem tentatively, from partial and incomplete viewpoints;
- to improve the formation of lexicographers by the preparation of guidelines and manuals and by the institution of specialized schools such as the Pisa Summer School;

(iii) Printing by photocomposition.

This is now a routine operation. The immediate actions to be undertaken are:

- a survey of photocomposed dictionaries available in machine readable form;
- the study of their possible exploitation, by means of suitable procedures, in order to extract lexical components which can then be used in automatic language processing (see point d).

c) Terminological Data Bases, Thesauri, etc.
These tools can be considered as natural objects of the activity of the AILA Commission on Terminology. Nevertheless, they also interest several activities which fall within the scope of the ACL Commissions. For example, terminological data are essential for computer assisted translation systems and must, in some way, be accessible in the context of interactive lexical databases. On the other hand, we are now experimenting methodologies for the creation of thesauri starting from the taxonomies implicit in the definitions of the normal printed dictionaries.

d) Automatic Lexicons, Machine Dictionaries, Multifunctional Lexical Databases
Automatic lexicons and machine dictionaries are essential core components in a whole series of activities belonging to applied computational linguistics. They supply essential information for parsing, generation, and transfer between different languages, in a variety of frameworks: models and theoretical studies of natural languages as a means of communication; educational activities; socio-linguistic researches; content-analysis; literary and philological text-processing; office automation; and all the previously mentioned systems of natural language processing. A number of scholars and researchers working in different disciplines, (e.g. linguistics, computational linguistics, artificial intelligence), are beginning to recognize that the intelligent processing of language using computer technology requires careful attention to the details of descriptions on a large scale, and that for those goals the availability of large quantities of lexical information constitutes a major priority. The creation of large automatic lexicons requires a considerable effort in terms of times and costs, owing to the amount of information to be acquired, and the question whether it is possible to conceive so-called "neutral", multifunctional lexical databases is increasingly posed. Neutral means constructed in such a way that the lexical components required by different specific utilizations, for different applications and within different theoretical frameworks, may be easily extracted.
The first action to be undertaken in this direction seems to be the promotion of a survey:
- to review the different lexical resources now in existence in machine readable form (see, for example, the typology in the article of Zampolli et alii)
- to obtain a description of the data contained in these resources, of their structure and of their internal organization;
- to compare their typology, structure, representation, classification, etc.
- to identify similarities, differences and analogies, in order to propose possible areas of progressive convergence and harmonization, and to suggest recommendations on the possibilities of gradually creating common, multifunctional lexical data bases.

The AILA Commission intends to sponsor, possibly in collaboration with other associations and organizations, a survey of this type, and also a workshop on lexical resources in machine readable form, which should be held in Pisa (in 1985 or 1986), with the participation of European, American and Japanese researchers belonging to the different disciplines involved.

II. Participation in the Sydney Conference

The absence of sessions explicitly devoted to applied computational linguistics in the Scientific Program of the Brussels Congress was much regretted. It was remembered that during the first AILA congress at Nancy in

1964, when AILA first appeared on the international scientific scene, approximately half of the communications presented concerned applied computational linguistics. Thus, the exclusion of this sector from the activity of the AILA congress means depriving AILA of one of its two original constituents. This seems ever less desirable as the utilization of computational tools in today's so-called "information society" becomes increasingly wide-spread.

It is true that, as already mentioned above, many specialized International Associations are already working in this sector and that there is an inevitable overlapping of the various memberships.

It is, however, also true that the activities of each of these Associations focusses on particular aspects, themes and methods. The AILA Commission must stress the specificity of its own framework. (cfr. the Presidential Address by A. Zampolli at the 1984 ALLC International Symposium).

The chairman was then asked to contact the organizers of the Sydney Congress, in order to:

- recommend the inclusion in the Scientific Program of the Congress of sessions dealing with applied computational linguistics;
- offer the collaboration of the Commission for the scientific organization of these sessions;
- offer the collaboration of the Commission for the preparation of computational facilities and for the organization of demonstrations of programs, procedures, tools in the sector of Applied Computational Linguistics.

Contacts have already been made with the organizers of the Congress, and with Associations such as the ALLC, the ACH and the ALC which have expressed their intention to collaborate.

III. Relationship with other AILA Commissions

The interdisciplinary relationships with other AILA Commissions was evidenced, and in particular with those Commissions which are more directly co-interested in the items chosen for the activity of the next years.

In particular, the chairman of the Commission was invited to join the Commission for Lexicography, in the last part of the session in Bruxelles, during which it was decided to start a collaboration between the two Commissions.

The items chosen were dictionary creation and the utilization of dictionaries in machine-readable form.

Professor Zampolli was asked to act as consultant of the Commission for lexicography. Following this meeting, collaboration and exchanges between Prof. Zampolli and Prof. Ilsson have already started. A similar type of cooperation is envisaged with the Commissions for Terminology and for Translation.