A. Zampolli

Computational Linguistics and MT in Italy

During the first of the three periods into which the history of machine translation (MT) is usually subdivided, Italy was one of the more active countries. Prof. S. Ceccato and his research group in Milan University proposed an original approach to translation problems, principally based on a conceptual associative network.

The activities carried out at the Ispra EURATOM Center by research workers from various countries have been highly seminal for the development of MT in Europe and in North America, including the Georgetown University project and its derived systems.

The 1961 NATO Summer School on machine translation, organised in Venice, influenced the development of both linguistics and computational linguistics in Europe and, in particular, the spread of studies on formal grammars.

During the "Dark Ages" of MT following the ALPAC report, practically none of the projects survived in Italy; efforts were concentrated on basic research in linguistics and computational linguistics and, in particular, on establishing and increasing cooperation between these two fields.

An example is the series of Summer Schools which were organised in Pisa during that period. In particular, in 1974, the programme was conceived with the goal of establishing interaction among linguists, computational linguists and Artificial Intelligence experts. At the 1977 School the pathway to functional lexical grammar was created and several of the teachers form the group working at present in the Bay-area.

In Italy, as in several European countries, at the beginning of the present-day third period, research in MT has been revived within the framework of the official Italian participation in EUROTRA. Some other minor activities in private domains are being developed by multinational enterprises, mainly for the development of an Italian component to be added to an existing system.

It may be interesting to briefly examine the specific way in which MT activities fit into the framework of research promoted in Italy in public sectors, particularly at universities and at the CNR (National Research Council), which may be summarised into two main trends and which, in a certain sense, may be considered complementary to each other:

- The first trend focusses, above all, on the study of theoretical models, Research, developed principally in the computer science departments and integrating the contribution of artificial intelligence, aims mainly, by means of the construction of 'toy-systems' with an extremely limited linguistic coverage and no build-up on each other, at the exploration of particular computational properties. Their relationship to the development of MT is mainly in the creation of a basic technical know-how;

- The second trend is characterised in particular by projects aimed at building multifunctional methods and tools which will constitute the background of a language technology oriented towards the development of language industries, The keyword is "reusability" and has two complementary meanings:

- to construct a repository of linguistic knowledge in such a way as to be able to extract information useful in various natural language processing components, to be inserted, through appropriate interfaces, into computational systems embedding different linguistic theories and aiming at different types of application;

- to design methods and tools capable of extracting relevant linguistic information treasured in collections constructed in the past both for human and computer usage.

At present, the majority of efforts concentrate on:

- large written and, where possible, spoken textual corpora, representative of various textual types associated with methodologies for the identification of the qualitative and quantitative characteristics of specific domains and sublanguages, and with robust parsers for analysis of real-world texts;

- large monolingual Italian lexical knowledge bases; studies on the feasibility of a polytheoretical representation of the linguistic information; methods for extracting from traditional dictionaries, available in machine readable form, implicit information, in particular that of a semantic and conceptual nature; tools to inter-connect lexica and corpora into a specialised linguistic work-station.

These activities are strictly connected to the progress of MT, not only in a generic way due to their relevance to the development of NLP components, but also in a specific manner, because of the particular care and priority given to the multilingual aspects both at scientific and at organisational levels, As examples of this we may refer to:

- the participation in a project, sponsored by the Council of Europe and coordinated by an Italian group, aiming at the creation for various European languages of compatible corpora and of methodologies for contrastive multilingual analysis;
- coordination between the European project and the North American efforts, in the framework of the Data Collection Initiative, promoted by the ACL to create large reusable corpora for free use;
- the participation in a project, co-sponsored by ACL, ACH and ALLC and co-financed by the NEH and the CEE through an Italian university, for the creation of guidelines and standards for encoding dictionaries and texts and the representation of their linguistic analysis;
- the coordination of an ESPRIT project, for the acquisition of lexical knowledge from existing dictionaries of five European languages, and for the connection of this knowledge in a multilingual conceptual structure;
- the promotion of an international group in which representatives of various linguistic schools aim at the definition of the feasibility of a polytheoretical representation of the lexical information required by the various grammatical theories.