Large Lexical European Projects and the Multilingual Aspect

Nicoletta Calzolari, Antonio Zampolli
Istituto di Linguistica Computazionale
Dipartimento di Linguistica, Universita' di Pisa
Via della Faggiola, 32 - Pisa 56100 - Italy
e-mail: glottolo@icnucevm.cnucc.cnr.it
Tel: +39/50/560 481
Fax: +39/50/589 055

(Topic: The paper will deal with many -carefully selected- topics that are listed in the call for participation, focusing attention on how they are treated in the major European cooperative projects aiming at building multilingual lexicons.)

We aim at providing an overview and a comparison of multilingual and Machine Translation (MT) related issues that emerge and that are handled within the major European projects in the lexical area (ET-7, Acquilex I and II, Multilex, Genelex, ET-10 Semantic Analysis of Cobuild, ET-10 Collocations, ET-10 Statistical Text-corpora based complements for Eurotra, Delis, Eurolang and Eagles), most of them with industrial involvement.

Most of these projects focus on building rather large multilingual lexicons or substantial lexical fragments for NLP applications. Some of them are more theoretically motivated, while others are more oriented towards applications. Concepts such as reusability and standardization, with different levels of emphasis, are shared by all of them. How these are reflected and realized within each lexical model, in particular with respect to their bilingual/multilingual components, deserves careful examination. Such an examination may lead to the highlighting of those areas (e.g. the choice of a representation formalism) where there is a trend towards convergence among the different approaches, or even a consensus, and contrasting these with those areas where divergences are clear (e.g. the acquisition methodologies) and where project specific choices are made.

We will concentrate on a number of exemplary/representative topics, examining their treatment in the different systems, touching issues concerning, e.g. lexical knowledge acquisition methodologies, integration of different sources of information, the global lexical model, interdependencies between different descriptive levels, representation formalisms, implementation, model of the lexical entry and the linguistic "standard", the theoretical background (if any), etc.

In particular, we want to concentrate on those aspects which are specifically relevant in a multilingual dimension as opposed to those which MT lexicons have in common with monolingual ones. From this perspective, emphasis must be put on topics such as the acquisition and formalization of semantic and world knowledge or the treatment of collocational data in a multilingual lexical environment.

Such an overview should be of use i) to encourage a definition of a minimal level of consensus on theoretical insights and technical descriptions as a basis for sharing basic tools and data, ii) to design - if possible - common priority tasks, and iii) to progress more quickly from this commonly agreed development level towards as yet unsolved areas, according to the evolving needs of the R&D community. This may be seen as the

setting-up of a reference framework for the evaluation of the state-of-the-art for large European cooperative actions, which may be compared, at the Workshop, with similar overviews for the US and Japan.