MACHINE-READABLE DICTIONARIES, LEXICAL DATA BASES AND THE LEXICAL SYSTEM

Nicoletta Calzolari
Dipartimento di Linguistica, Universita di Pisa, Pisa, ITALY
Istituto di Linguistica Computazionale del CNR, Pisa, ITALY

I should like to raise some issues concerning the conversion from a traditional Machine-Readable Dictionary (MRD) on tape to a Lexical Data Base (LDB), in order to highlight some important consequences for computational linguistics which can follow from this transition. The enormous potentialities of the information implicitly stored in a standard printed dictionary or a MRD can only be evidenced and made explicit when the same data are given a new logical structure in a data base model, and exploited by appropriate software.

A suitable use of DB methodology is a good starting point to discover several kinds of lexical, morphological, syntactic, and semantic relationships between lexical entries which would otherwise have remained unexploited. Moreover, the transformation of a "very large-scale" MRD into a LDB provides the means of operating throughout the lexicon in a really extensive manner. I think in fact that an "almost exhaustive" approach to lexical facts is essential both for reliable investigations of a lexical system, and for many kinds of linguistic applications which cannot be restricted to a particular domain of discourse.

The possibility of abstracting significant regularities from recurrent patterns of natural language definitions by means of suitable computational methods, and of reaching a formalization of a number of important structuring relations within the lexicon will be discussed. An overview of the "associative links" already produced in the Italian LDB, and of other allowable interconnections will be given.

In a "relational" organization of a computerized dictionary with complex interlinked structures, each word acquires its meaning as a result of its position in some of the partitionings created by the formalized relations. When an entry is activated, all of its relations with other entries can be activated, too. Conversely, when a relation is activated, all of its linked concepts are made immediately available. Conceptual and linguistic information at many levels is thus interactively retrievable from the LDB following the appropriate pointers.

I shall especially take into consideration those types of relations which can be of relevance not only for "Computational Lexicology" research, but also in a more general Computational Linguistics framework. An example is provided by derivational relationships which, when formalized, give rise to families of semantically and syntactically connected entries, linked to the same base-word node, and substitutable in different syntactic formulations of the same conceptual meanings. Another example concerns case or argument relations, both (a) between lexical items, and (b) governed by lexical items. From (a) I expect to achieve, from the natural language definitions, useful information on the different lexicalizatons of case-slot fillers in the case-frames of typical actions. In contrast, with (b) I can establish an encoding with each entry--and often with each word sense--of information on its surface and deep case-argument structure. The utility of the extensive inclusion of similar information in a LDB which should be the input for a lexically driven parser, for machine translation, etc., is obvious.

As a conclusion, it should be pointed out how a LDB must be considered at the crossroad between texts and system, and in this perspective some essential properties of a LDB must be stressed. A first property is "multifunctionalism"; it is connected to the role of interfaces to the LDB. We must tend towards creating 'a single' integrated system which, through many different interfaces, can be adopted for all the range of possible applications, and by all the possible users, where user means both a human user and a computer program. Another important property is that of being "multi-perspective." This property of multiple access can create something like a constellation of sublexicons, which altogether capture the many possible structures which can be observed in the lexical system, along many dimensions of relatedness. The mediating function of a LDB between system and texts can thus be considered as the mapping of lexical structures, of many kinds, on linear unstructured texts.