Terminological data banks and grammatical information

Svatava Machová

The rapid development of science and technology now taking place requires international cooperation (cf. Stellbrink, in this volume). Such cooperation is closely connected with the continuous growth of science terminology, which is crucial for the process of effective communication between specialists. The answer to the question whether there exists a term introduced for a certain concept in a given language and what kind of term it is, has become a serious problem for terminologists working in newly developing fields.

The development of computer technology and its availability, along with the development of computer text-processing systems have created the prerequisites for the solution of this problem with the aid of computers and also with the aid of so-called terminological data banks.

By terminological data bank (TDB) we mean, in accordance with a number of other authors, terminological material stored in computer memory media, in a form of records structured in fields constituted by terminological elements. As the basic function of TDB's we regard accumulating, storing, processing and disseminating terminology.

It is well known that designing the structure of an entry in TDB's is a rather unrewarding task. It is expected to be accomplished within the initial stages of the project, although not more than the assumed goals and the assumed users of the TDB is known at the stage.

In the present paper, attention is focused only on one part of the TDB-entry, namely its grammatical field. It is our objective to consider how much the content and the extent of the grammatical field of a TDB-entry are linked (i) with the objective of the TDB and (ii) its users, (iii) with the type of the natural language in question and (iv) with the theoretical linguistic knowledge of the staff supplying data. We discuss the way these factors influence the final make-up of the grammatical field in a TDB-entry and we try to find out whether it is possible, and wise, to design any universal make-up of the grammatical field independent of these factors.

TDB's may have various goals. They may be oriented to assist translators of scientific literature or to assist terminological standardization; they may be oriented to facilitate the work of lexicographers or foreign language teachers and students. They may finally be oriented to computer text-processing. Naturally, there are TDB's having more than one objective. Projects aiming to give a detailed description of the entire general and specified vocabulary of a language go beyond the scope of this paper.

If a TDB has one of the first three goals cited here, that is, to serve translators as a constantly up-dated on-line dictionary (whether the latter is a component of a machine-aided translation system or not), or if it is to be used for automatic generating of printed bilingual or multilingual terminological dictionaries and glossaries, the designer assumes a user capable of studying and translating texts of the language in question, hence he assumes a user conversant with the grammatical laws of that language. Therefore, in these cases, we consider it sufficient to store only basic information in the grammatical field of these TDB's (viz. the specification of parts of speech).

If the TDB designer is exclusively oriented towards assisting foreign language teaching and learning, he presupposes a user who has not yet mastered the language adequately. A designer of this kind of TDB's should expand the grammatical field by information on the valency properties of verbs and nouns, the conjugation characteristics of verbs and the declension characteristics of nouns, all of which is not necessary for TDB's oriented towards translators. Thus, the content and extent of the grammatical field has to be designed in accordance with the presupposed TDB users' competence in the natural language in question.

In all the above cases, the front-end user was a human being. But if the TDB is to be incorporated into a computer text-processing system, that means, if its front-end user is a computer, the TDB designer is faced with one of the most complicated situations conceivable, as far as solving the question of the content and extent of the grammatical field. At present, three types of computer text-processing systems can be considered in connection with TDB's: computer proof-reading, computer indexing and machine translation.

If the TDB is to be incorporated into a computer proof-reading system operating in a publishing house for scientific literature, it will have to involve some lemmatization process. The complexity of lemmatization processes and the amount of grammatical information required for their successful operation are strongly dependent upon the type of language being tackled. The information concerning the properties of a TDB-entry on a morphological level, that is to say, information required for a lemmatization procedure, appears to be much richer for inflected languages than is the case for analytic ones: For analytic languages (such as English), the process of lemmatization is relatively simple, for strongly inflected languages (such as Czech, Russian and other Slavonic languages) it is rather complicated. If, in the TDB which is going to be used in some computer text correction system, an entry is being designed for an analytic language, the content and extent of the grammatical field can be settled by the TDB designer independently of the concrete strategy of any lemmatization process. All that a designer has to know is the morphological structure of the natural language in question. If, however, a TDB-entry is designed for an inflected language, the extent and content of the grammatical field cannot be decided independently of the strategy of the lemmatization process chosen for the system. For each lemmatization process involves theoretical linguistic issues of its own.

In keeping with them, various kinds of morphological information are needed for its successful operation, and different classifications of the same morphological phenomena may be required.

If the TDB is to constitute part of the lexical component of a computer indexing system or of a system for machine translation, the content and extent of the grammatical fields cannot be decided before the linguistic strategy of the system chosen is known. This is because each system of computer indexing and machine translation is based on a distant linguistic theory and, in the process of text analysis and synthesis, in accordance with that theory, it is necessary to retrieve and focus attention on different kinds of properties possessed by lexical units. We do not think that it would be possible to find a kind of representation for such information that would be independent of and neutral vis-a-vis a particular linguistic theory chosen as the theoretical basis of a computer text-processing system.

Each TDB requires a continuous supply of new data, otherwise it loses its raison d'être. Recruiting a uniformly trained team of workers collecting these new data is a difficult problem in the implementation of TDB's.

A multidisciplinary and multilanguage TDB cannot do without numerous staff supplying data, composed predominantly of specialists in different disciplines. Their linguistic education may vary. Moreover, it is well known that the classification of language phenomena according to the principles of one linguistic school may be unambiguous in some cases, questionable in others and, in some cases, we are even unable to find any guideline for the classification of a given language phenomenon within the framework of the linguistic school in question. The higher the demands of TDB designers are as to the extent of information given in the grammatical field, the more serious such factors appear to be.

Therefore, the TDB designer, when designing the content and extent of a grammatical field, has to consider what level of linguistic knowledge he is to expect from the staff supplying data and whether the latter is in possession of it. The TDB designer is expected to work out exact evaluation criteria for the language phenomena concerned. No decision-making can be left to the linguistic intuition of the staff supplying data.

It follows that the decision as to the extent and content of the grammatical field in a TDB-entry is not merely a matter of linguistic theory. A significant part is played by pragmatic factors including the human terminologist and the human user of TDB's with their knowledge, habits and interests, as well as the context – in the wide sense of this word – in which the TDB is to operate. Sticking rigidly to implementation of certain linguistic principles when designing the extent and content of the grammatical field, regardless of these pragmatic factors, cannot but constitute one of the causes of failure of such projects.

The above fact leads us to the conclusion that it is not wise to try to settle the extent and content of the grammatical field generally, because each TDB exists in a specific context. This view is also supported by the fact that the hardware on which TDB's are operated at present, is now different from what it was previously. The big computers led to designing big national TDB's. The micro- and minicomputers of the present enhance the prospects for designing smaller TDB's specialized in particular profiles.

In Czechoslovakia, as a part of the State Research Plan, work has been started on a Czech TDB at our State Library. The project of the Czech TDB aims at devoloping a six-language TDB (Czech, Slovak, Russian, English, French, German), the main orientation at present being to facilitate lexicographic work. The Czech TDB is term-oriented and is going to be tested experimentally on the terminology of computer technology and electronics. The staff supplying data is going to be large. This specific context is reflected in the content and extent of the grammatical fields of the TDB-entries. They contain only the information concerning parts of speech and grammatical gender in nouns. Although the Slavonic languages handled in this TDB are inflected ones, no information is given as to the type of declension or conjugation, because the users of dictionaries created on the basis of this TDB are assumed to be translators of scientific literature.

The important features required of the TDB of the future are: simplicity, high quality and good service to users. The grammatical field, being a component of TDB-entries, should possess these qualities too.

References

Bahr, J. (1978), Reflections on the project of a lexical data bank, in: *Cahiers de lexicologie* 32, 55–64.

Calzolari, Nicoletta/Ceccotti, Maria L. (1980), A project for an exhaustive lexical data base system, in: *Segunda Conferencia International sobre bases de datos en humanidades y ciencias sociales,* Madrid, 47–49.

Fillmore, Charles J. (1969), Types of lexical information, in: Kiefer, Ferenc (ed.), *Studies in Syntax and Semantics,* Dordrecht, 109–137.

Goetschalckx, J./Rolling, L. (eds.) (1982), *Lexicography in the Electronic Age,* Proc. of a Symp., Luxemburg.

Snell, Barbara (ed.) (1983), *Term Banks for Tomorrow's World,* London.