COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

TERMSERVICE - AN AUTOMATED SYSTEM FOR TERMINOLOGY SERVICES

Bonka Nikolova
Irina Nenova

Laboratory of Mathematical Linguistics
Institute of Mathematics with Computer Centre
at the Bulgarian Academy of Sciences
Sofia
Bulgaria

The paper discusses the background, use environments and content of an automated system for multilingual terminology services, developed at the Laboratory of Mathematical Linguistics of the Institute of Mathematics with Computer Centre affiliated to the Bulgarian Academy of Sciences. Particular emphasis is given to terminology acquisition and facilities for automated lexicography.

INTRODUCTION

Automatic language processing makes a current use of multilingual databases as a component of the computer environment of machine translation projects or computer-aided translation, and as a means of coordinating terminological standardization across several languages. Translation database projects have been developed in North America and Europe and there is a great variability in their content, use and structure. What is common about them all is the attempt to speed up and simplify the translation process, the effort to support many languages, to increase the volume and reliability of the linguistic information contained, to assure flexible access to the database, to make it inexpensive and available to a large number of users.

BACKGROUND OF THE PROJECT TERMSERVICE

Nobody working in a scientific or technical field can do without some information published in other languages. This implies that the need for translation is enormous, which means that for each language efforts must be made to compile the correct terminology. The latter, in its turn, needs to be standardized.

As elsewhere, in our country translators and specialists in scientific or technical fields are asking to have the terminology relevant to their work made available in foreign languages. Moreover, conventional dictionaries, provided they exist, cause a lot of problems and often turn out to be an inefficient alternative. This is mainly due to the rather long publication terms which make a specialized dictionary inadequate when it is finally published, especially in rapidly developing fields where the growth of terminology is most pronounced. Besides, a conventional dictionary, as a rule, makes no provision for updating its files, mainly from feedback coming from users.

The only technique which offers enormous storage capacity and fast

retrieval of terminology is computerized data processing. A well set up terminology bank offers ample opportunities to solve the difficulties met by the translator or specialist in the field in dealing with scientific or technical texts.

To meet these needs, the Laboratory of Mathematical Linguistics at the Institute of Mathematics with Computer Centre affiliated to the Bulgarian Academy of Sciences has set itself a number of short, medium and long-term aims, designed to lead to the setting up of automated terminology services - the project TERMSERVICE.

FUNCTIONS AND USERS OF THE SYSTEM TERMSERVICE

The system TERMSERVICE is designed to be used in the following environments:
- in the computer-aided translation environment the terminological database can be used by human translators and specialists in scientific and technical fields as a computer-aided multilingual dictionary;
- as far as the terminological environment is concerned, the database provides sufficient linguistic information to conduct research on terminology and to standardize terms, abbreviations, acronyms in several languages.

As secondary resulting use environments we could mention:
- machine translation systems which can adapt and incorporate the terminological database to serve their aims in translating natural language documents;
- computer-aided instruction in foreign languages.

ACQUISITION OF TERMINOLOGY

Dictionary-making is a popular trade with established traditional procedures for compiling the lexical files to be included in a printed volume. People are also aware of the traditionally high cost of developing user-oriented specialized terminology. This task is quite labour-consuming and requires team efforts from a large number of translators, terminologists and lexicographers.

As far as the traditional methods of lexicography are concerned, the computer can help in alphabetizing and updating the files, which is a comparatively elementary level.

The literature notes the following ways of establishing linguistic material for the development of terminological databases:
- analysis of original documents in each language, the comparative study of these documents giving real equivalents of professional language usage;
- compilation of terminology by specialized institutions;
- inclusion of terminology contained in conventional dictionaries and other reference materials;
- interaction between currently existing databases and terminology exchanges;
- techniques other than by means of text, as, for instance, experiment, inquiry, introspection.

PARALLEL TEXTS ANALYSIS

We have chosen to extract terminology from previously translated material by the method of parallel texts analysis. The idea was taken up from presently existing automated systems for lexicographic services developed in the USSR by the All-Union Centre for Translation of Scientific and Technical Literature and Documentation, Moscow.

Preference is given to this method since it yields most reliable and genuine results about the state-of-the-art of terminology and allows to perform the lexicographical work mostly automatically.

An original text in one language and its translation in another are termed as parallel texts. The task is to separate the items of translation in the original text, to find their equivalents in the target-language translation, to establish a one-to-one correspondence between them and fix the latter in a correspondence file.

We have accepted the following definition for "an item of translation": an item of translation is the minimal language unit from the original text which is to be translated as a whole in the sense of no language units available in the translational text that reproduce the meaning of the components of the given item of translation, in case there are any.

Some authors point that an amount of 100-200 thousand wordforms have to be processed in order to reach a point of strong decrease in the number of new terms to be included in the dictionary.

THE PROGRAM PACKAGE AND THE RESULTS

The program package for automated compilation of terminology in English and Bulgarian contains 8 programs for processing of the parallel texts, and for compilation, maintenance and usage of a machine dictionary of English terms and their equivalents in Bulgarian. The programs are written in PL/1.

Original English texts from scientific papers and their translations in Bulgarian serve as initial linguistic corpus for lexicographic purposes. The aim is to process the texts so that the labour of the linguist in compiling the machine dictionary would be facilitated to a maximum degree, thus obtaining in short terms a dictionary covering to a high extent the terminology relevant to the chosen scientific field.

The output from the operation of the program package is, as follows:
- the texts of the original paper in English and its translation in Bulgarian in a form suitable for the coding of the translation equivalents;
- a dictionary-concordance containing wordforms from the text, pointed by the linguist together with the neighbouring context within boundaries specified by the linguist again. The contextual examples have the advantage of giving each term with a more precise meaning than if it were isolated and serve to remove polysemy.
- a dictionary of the wordforms from the original text and all of their equivalents that occur in the Bulgarian translation.

The package allows to introduce terms and their equivalents in the target-language in explicit form as well as to exclude separate

items from the dictionary.

Parallel texts analysis is our main source of terminology but, of course, it is not the only one. We are making inquiries among the mathematics professionals in the field about the volume and scope of the terminology they find relevant to their work. Besides, the terminology we have extracted from translated texts could be mapped to a mono or bilingual dictionary, or manual, or handbook specialized in the field. The facilities of an on-line mapping will additionally speed up and enlarge the possibilities of terminology acquisition.

The program package could be of help not only to the linguist and lexicographer but presents computer aids to terminologists as well. The automatic context look-up supplies easily usage samples, the text concordance facilities come at hand for building up text-oriented lists of terminology, and the merging of terminology lists from different sources offers opportunities to solve ambiguities or disagreements in translating a term.

DESIGN OF THE SYSTEM TERMSERVICE

The system TERMSERVICE is designed to support several languages, the main three being Bulgarian, English and Russian. We use direct linkage between pairs of meanings of a certain term in different languages.

The fields of specialization of the database are scientific and technical.

For each source-language term (a single word or a phrase) the following information is contained in the database:
- target-language equivalents,
- synonyms, if any;
- subject-field code;
- grammatical code, spelling variants, standard abbreviations;
- contextual examples of usage of the term;
- definitions are supplied for source-language terms having no equivalent in the target language;
- possible word combinations with the headword, if any;
- the source the term was extracted from;
- cross-references to other terms.

Since the system is not oriented only towards professional translators, general and specific terminology is complemented by common-use lexems, their selection being motivated by frequency of usage.

The logical access to the database can be accomplished by a lexical item, by synonym, subject-field code, grammatical code, source of the entry.

The modes of access to the database are batch or interactive query and computer output to microfilm or microfiche. Printed dictionaries will also be generated from the database.

Source-language terms are stored in conventional dictionary form. To assist the user's query, however, the project envisages a block of automatic reduction of inflected forms to standard form. Till that time the system will work with the help of human pre-editing.

The project also envisages automatic recording of unsatisfied look-up requests for update purposes.

CONCLUSION

The Laboratory of Mathematical Linguistics at the Institute of Mathematics with Computer Centre is the first in Bulgaria to develop a multilingual terminological database. We have started recently and the data capture function of our project is the most advanced. There is still a lot to be done in querying and updating, which is a matter of time and staff.
The current study of the user's needs and a highly efficient interaction and cooperation between presently existing terminology banks will provide for a successful computer aid to human translation.

REFERENCES

(1) Marchuk, Y.N. (ed.), Computational Lexicography (All-Union Centre for Translations of Scientific and Technical Literature and Documentation, Moscow, 1976) (in Russian).

(2) Presently existing Machine Translation systems and Automatic Dictionaries, Survey Information of the All-Union Centre for Translations of Scientific and Technical Literature and Documentation (Moscow, 1979) (in Russian).

(3) Ubin, I.I. (ed.), A Russian-English Frequency Dictionary in Electronics (All-Union Centre for Translations of Scientific and Technical Literature and Documentation, Moscow, 1977) (in Russian).