Machine-Aiding of Translation Processes by User-Controlled Dictionary Look-up

GUNTER NEUBERT

Institute for Applied Linguistics, Dresden Technical University, Dresden, German Democratic Republic

Reviews automatic dictionaries of various types. The usefulness of terminology data banks in different spheres of human activity is illustrated. The operation and various applications of the Dresden electronic specialized dictionary are described.

It is very well known that machine translation being the long-term offer of linguistic engineers for the full-scale automation of the translation process cannot easily and in a short time be implemented in all fields and in all languages. One of the potential by-passes, simultaneously constituting a necessary step towards machine translation, consists in automating the operation of dictionary look-up.

1. Comprehending the special terms of the text to be translated and selecting the adequate term from sets of equivalents listed in the dictionary entry, or constructing a likely equivalent by processing the different terminological informations collected, present most of the problems of the translation process and demand the greatest proportion of expended time, sometimes up to 80 per cent of it. For that reason, the preparation of a data file containing terminological information and utilisable with programs is obviously most promising in regard to machine-aided translation. In addition to the consequent concentration of lexicographical power, such a terminological data file, or data bank as it is more frequently called, allows various other data handling operations to be automatically performed, in this way overcoming some of the shortages of traditional dictionaries.

For example, it is feasible to derive special lists of terms, or vocabularies, containing all the stored terms with specific features, such as being used in a special field or in the context of a special product; and/or being a member of a special category of concepts like instruments, tools, machines, procedures, etc.; and/or pertaining to a given character string like English *pump,* German *Pumpe* or Russian *nasos.* Interpreters may use such vocabularies for preparing themselves for a congress or discussion, translators may complete their terminological knowledge, e. g. in case they must translate a text that deals with a scientific or technical subject not well known by them.

Furthermore, computerisation of terminological information retrieval leads to, and has as its indispensable base, computerisation of the entire terminological work including compiling terms, preparing entries of the bank, editing dictionaries or other human-readable vocabularies, and updating the bank as well as the dictionaries derived from it.

Selection programs, which group the terms of the bank according to information elements previously added. provide a substantial aid to linguists interested in special purpose languages. Not only fundamental investigations will be supported, and in some cases, made possible for the first time, but also work aimed at the improvement of the retrieval process itself. This aspect will be detailed further below.

Language teachers and students may make use of the same facilities as interpreters and translators. Specialists responsible for arranging the concepts of a given field in the form of an information retrieval language, or thesaurus, may at least find help for the most exhausting routine parts of their work, particularly for the representation of the hierarchically or otherwise ordered set of concepts in the form of different types of indices required for comfortable use of the thesaurus.

2. Language data handling is not cheap. For this reason, and in order to make the once collected terminological treasures available to all users concerned, our team at the Dresden Technical University, which consists of specialists in special purpose languages and computer scientists, decided to provide the terminological data bank we had to develop with a maximum degree of multi-usability. In order to meet the resulting demand for the data bank not to be orientated at any of the particular tasks listed above, the so-called EWF (from German *Elektronisches Wörterbuch der Fachsprachen* — Electronical Dictionary of SPL; this designation has since become obsolete) was given a general-purpose lexicographical and data processing structure and set of programs. The lexicographical work performed to fill the data bank's stock is concept-oriented. That means that all the terms of identical or nearly identical meaning (synonyms and equivalents) are united within one entry, together with accompanying information elements. Those information elements, which are added by the lexicographer, concern the represented concept, as special field, category of the concept, definition or context, if available, etc., as well as the individual terms, such as part of speech, gender (if applicable), terminological quality (e. g. deprecated, to be preferred, standardised, etc.), source, responsible terminologist/lexicographer or editor, and so on. Internally, spelling is orthodox, though the characters carrying diacritic signs cannot be printed by the peripheral devices at present available, on output, and must be written in the form of parochial character combinations, on input. The computer used up to the present has been the Soviet BESM-6. Just now, work is being devoted to transferring the data bank and prograins to the computers of the Unified System of Computing Devices of the CMEA countries, which especially offer a higher degree of compatibility; hence, more facili-

ties for international data interchange and a higher degree of accessibility for all kinds of users will be available.

The programs perform data input and updating operations, both of general and of specialised types, data retrieval, selecting operations, and editing procedures. Besides, linguists are given, facilities to analyse German word compounds for their morphematic and lexical constituents, and to morphematically decompose Russian words.

3. These latter routines are considered to be most important for enhancing the data bank output in the case of machine-aided translation. It is well known that a considerable amount of the questions put to a dictionary will not be answered when based on a character-by-character comparison of the Searched Term (ST) with the set of Compiled Terms (CT). As some of our investigations clearly show, the amount of directly answered terminological questions by even a high-level special dictionary does not exceed some 30 per cent. In addition to this fact, which will hardly be overcome by strengthening terminological efforts, as it results from the fast and progressively growing development of science and technology, translators sitting in front of a common displaying screen cannot survey the alphabetical or semantic environment of their ST. They are dependent upon their experience in using dictionaries and, besides, have to type question after question on the terminal keyboard, until they finally find sufficient auxiliary information to form a likely term in the target language.

Various techniques of matching ST and CT are known, and are being provided by producing terminological data banks. As the most natural way, searching difficulties can be reduced by eliminating such character differences as resulting from capitalisation, diacritic signs or the German 'ß', and word differences due to the somewhat arbitrary hyphenisation of the English language. These difficulties can be reduced further by cutting off one, two or more of the last characters of the ST after having compared them with the characters or character strings respectively, given a list of possible endings in the language concerned. In this way, differences between word forms like *hydrocarbons* as ST vs. *hydrocarbon* as CT, or reversed, or *Verdampfen* vs. *Verdampfung,* or *vsasyvanie* vs. *vsasyvat'* are ignored. One of the most widespread techniques is based on the principle of longest-possible match, i. e. to base the answer on the CT that is the longest continuous character string identical with the initial characters of ST.

All these and some other related techniques not mentioned here have the disadvantage of not enabling users to control the generation of the term substituted for their original ST in the case of non-existence of a corresponding CT in the data bank's store. For that reason, the retrieval program of the EWF offers the facility of prescribing the length of comparison in each particular case of an ST not found in the data bank. Users may command the execution of any of the following procedures:

(a) matching the first *n* characters of ST;
(b) matching the last *n* characters of ST;
(c) matching the first *n* characters of an index word of the ST in the case of multiple word compounds; or
(d) matching the last *n* characters of that index word.

In general then procedure (a) provides users with the equivalents of the determining constituents of ST, while procedure (b) gives the equivalents of the determined constituents, at least in the case of word compounds with the corresponding word order. The reader can easily imagine modifications for other cases of word order.

In order to avoid a time-consuming count of ST characters, users may also prescribe the value of *n* as the number of residual characters of ST to be excluded from the comparison, by adding a negative sign preceding *n.*

Without having repeated his ST the user receives a list of CT corresponding to the enlarged retrieval conditions printed out on the display, and decides which of the entries is or are to be put out completely. Thus, he can collect terminological information step after step, if necessary, to form the desired term.

The above mentioned procedures enable users to specify searching policies dependent on the actual ST. However, it is evident that in many cases the programs by itself, or more exactly the linguistic algorithm being the base of the program, can generate auxiliary terms that may be considered to be probable synonyms or constituents of the original ST. In all languages concerned, groups of synonyms or term variants exist like the German *Wiederholungsstart* and *Wiederholstart,* the English *seat valve* and *seating valve,* or the Russian *sglaživajuščij kondensator* and *kondensator dlja sglaživanija,* which can be mutually transformed following language-bound rules which, however, have not been sufficiently defined up to now. Just as well, programs can be made to dismember word compounds, which are especially relevant in German, into their constituents. For those transforming or dismembering routines, morphematically analysing algorithms are the necessary fundamentals. At present, they are in the stage of draft for the German and Russian languages.

In the case of non-direct man-machine dialogue, users can be given the possibility to pre-call the earlier mentioned match enlarging procedures to be executed during the processing of a list of ST only in the case of non-existence of the original ST in the stores of the data bank. However, use of this principle is justified only in lexicographically well-compiled fields to avoid a considerable 'swelling' of the output lists.

In any case, each step towards solving the problem of an effective retrieval of terminological information will also be a step towards solving the problem of continuously filling the terminological stores, for satisfied users will readily undertake the additional task of contributing their terminological by-products from translating to an easily accessible file.

REFERENCES

1. Baumann, E.; Neubert, G. Das Elektronische Wörterbuch der Fachsprachen — Aufgaben und lexikographische Struktur. *Wiss. Z. Techn. Univers. Dresden,* 1974, 23, No. 3/4.
2. Neubert, G.; Kukuczka, H.; Meyer, E.; Gründler, H.-V. Der Elektronische Wörterbuch der Fachsprachen — rechentechnische Struktur, Eingabe und Ausgabe. *Wiss. Z. Techn. Univers. Dresden,* 1974, 23, No. 3A.
3. Baumann, E.; Neubert, G.; Meyer, E.; Reinhardt, W. Das Elektronische Wörterbuch der Fachsprachen Russisch/Englisch/Deutsch — EWF. *Wiss. Z. Techn. Univers. Dresden,* 1975, 24, No. 6.
4. Kukuczka, H.; Meyer E.; Neubert, G. Eine weitere Ausbaustufe der Dresdner Terminologiedatei EWF. *Wiss. Z. Techn. Univers. Dresden,* 1979, 28, No. 1,