[*Translating and the computer 13: The Theory and Practice of Machine Translation – A Marriage of Convenience?*
Papers presented at a conference … 28-29 November 1991 (London: Aslib, 1992)]

THE EUROTERMBANK

Paul Nekeman

Euroterm

EUROTERM

Euroterm Maastricht was founded in 1987 to carry out research in the fields of terminology and computer-assisted translation. The organization was set up at the initiative of the Dutch State School of Translation and Interpreting and has received considerable financial support from the provincial authorities of Limburg.

In addition to doing research work, the organization has been very successful in providing language services, including translation and interpreting. The latter are now the domain of a Limited Company, Euroterm Translations BV, established by the original foundation, which is the major shareholder. The organization is also a partner in MINT, a European Economic Interest Grouping, set up in combination with other translation companies in a number of European countries for the specific purpose of handling multilingual translation assignments. Revenues from Euroterm's commercial services help provide a solid financial basis for its research, while the relationship between research, education and practice provides opportunities for cross-fertilization.

EUROTERMBANK

Since late 1988, Euroterm Maastricht, together with the Dutch State School of Translation and Interpreting, has been engaged in setting up a large multi-lingual terminology database. The intention of the original project was to investigate the possibilities and problems involved in large-scale storage and retrieval of terminological data. The term bank was also to serve an educational purpose, in the first instance for students of translation in Maastricht. The subsequent involvement of Elsevier Science Publishers and the success of the conversion of ESP dictionary material have led to the decision to open the term bank to public users.

Hardware and software

The term bank uses a PRIME 4050 mini-computer with Adlib (ADaptive LIBrary management system) software, a package originally designed for library cataloguing applications. Euroterm itself developed the terminology database application and the OPAC (On-line Public ACcess) user interface for on-line access.

Structure and languages

The present term bank consists of three separate databases: the classification system, comprising a UDC thesaurus, a source bank (containing full bibliographic references to the sources of terms, definitions and contexts), and the term bank proper.

The term bank caters for up to 14 different languages, each having equal status. At the moment, the available languages are:

Danish	Dutch	English	French	German
Italian	Latin	Portuguese	Spanish	Swedish
Japanese (in transcription)				

Euroterm is currently investigating the possibility of including five other languages, in their original scripts. This requires the development of screen drivers which can accommodate the various scripts.

Arabic	Chinese	Greek	Japanese	Russian

Each language can be used as the source or target language and all fields appear in each language record.

Each record comprises the following fields:

Field name	Length	Type
publisher	40	alph
term	64	alph
source term/page	9+10	num/alph
part of speech	6	alph
country code	3	alph
definition	64	alph
source definition/p	9+10	num/alph
context	64	alph
source context/page	9+10	num/alph
author	30	alph
date	8	num
reliability code	1	num
extra linguist.info	64	alph
source extrl./page	9+10	num/alph
broader term	64	alph
narrower term	64	alph
classification	64	alph
UDC	40	num

The names of the parts of speech follow Wüster's categorization, while the country code complies with ISO 3166. The reliability code and the UDC code are not visible on the screen.

With the exception of the fields "reliability code", "source", "page", and "publisher", all fields are so-called "repeated fields", which means that either the field repeats itself within the same record or that the text continues on the next line within the field in question. In the first case, repeated fields can be used to record synonyms, spelling variants and abbreviations in

the "same" field as the term. Since all repeated "term" fields are indexed, a query will always result in the entire record being displayed, meaning that cross-references to other records are not required. In the second case the repeated fields serve to accommodate definitions and contexts, with lines of text following one another in word-wrap fashion.

The source fields are numerical and refer to the corresponding numbers in the source bank, while the page numbers can be in Arabic or Roman numerals. The source on screen takes the form of a text extracted from the source bank.

Access to the EuroTermBank

The EuroTermBank is accessible from anywhere in the world for registered users who have at their disposal a PC, a modem and communication software. They may access the term bank either via the public telephone network or via their national data network linked to DATANET 1, the Dutch national data network. The latter accessing procedure provides a more reliable, quicker and cheaper service. Searching the database is menu-driven and user-friendly; context-sensitive HELP functions are available at every stage.

Users can search for monolingual, bilingual or multilingual information. The first two modes present terms with definitions and contexts in one or two languages (if available), while the multilingual mode provides equivalents in all languages present.

Content

Much of the content of the EuroTermBank consists of material that is also available in printed form. Elsevier Science Publishers have undertaken to make available all their dictionary material for incorporation in the term bank. While their most recent publications are available in machine-readable form, more than half of their dictionaries are not, and the problem of efficient conversion of printed matter into electronic files has yet to be solved. Optical character recognition techniques do not at present yield results of sufficient quality to make them preferable to manual retyping. Other publishers and organizations are also being invited to contribute their material.

Procedures

The re-use of terminological material available in a different form presents certain problems. Some are easily solved, being basically of a technical or organizational nature. Others are more difficult because they originate in the necessity of applying the theoretical principles of terminology.

The simplest problem (but not always the easiest to detect) is that of damaged data. Magnetic carriers and recording equipment are not 100% error-free and inevitably some of the data arrives in mutilated form.

A general problem concerns the specified limits of the database and the record formats. Every database is necessarily a compromise: the number of fields is fixed, as is field length, and even the number of records is pre-determined (in our case by the physical capacity of the hard disk in combination with certain operations carried out by the software). Terminological or dictionary material conforms very often, but not invariably, to the limits set by our database. The exceptions can usually be traced to special circumstances, such as the use of a paraphrase in the absence of the proper term, with the maximum field length for terms thus being exceeded, or the existence of 53 synonymous terms for one concept.

Much of the material which has so far been converted into the EuroTermBank database format was supplied in the form of typesetting tapes. These are magnetic carriers of data and basically serve as a transport medium between the data entry equipment and the printing press. The information on these tapes is coded in such a way that the data appears on the printed page in the required form. The internal coherence of the coding system and even the consistency of data recording are of minor importance if the final printed product has the desired appearance.

For a database, which has a different field for each type of information, coherence and consistency are vital. Both the coding system (i.e. the codes used for the various types of information) and the linguistic codes (such as names of parts of speech and geographical codes) must be standardized and used consistently to arrive at a coherent information system. Within the latter two categories of codes, the number of deviant forms is limited and a more or less exhaustive list can be built up, allowing an automatic entry procedure to guide the input both of data that conforms to the standard and of data that deviates from it.

More difficult to deal with is the incorrect use of existing codes. Data entry and coding for typesetting multilingual dictionary material is often performed by personnel who have an insufficient command of all the languages involved and who do not understand the data. The exact nature of the information categories which they specify may be unknown to them. The final result, in printed form, may appear perfect both to the proofreader and to the user, but the underlying electronic data recording format may not be perfect at all.

The authors of multilingual dictionaries usually supply their data either as camera-ready copy or in a database or word processor format. Automatic spelling checks can only be performed efficiently on monolingual material, and this requires the separation of the various languages and the recording of mistakes for later manual correction of the data within the database. The use of automatic spelling checkers is essential in order to obtain consistency throughout the collection. Such consistency within the term bank as a whole would require all the material incorporated to conform to a standard applying to the entire term bank, and consequently the conversion of deviant data into the standardized form.

Most dictionaries on the market today have been compiled with a printed product as the objective rather than a terminological database. The finished product as conceived by the compiler influences the recording strategy and organization of the material. An example of such an approach is the addition of annotations (between brackets) directly after the term, to indicate the specific usage or circumstances in which the term should (or should not) be used. In a database such notes can be placed in a special annotation field. The use of brackets to indicate variant spellings of parts of a term is also problematical, since they affect the indexing system and make it impossible to search for all variants. Instead, all variants should be included in full in the record in one of the repeated fields of the type TERM.

Another example, which also illustrates a more fundamental problem, is the use of numbers to separate different usages or definitions (i.e. meanings). If these are put in one field and consequently also appear on the typesetting tapes as one piece of information, the user of the term bank has no way of selecting the required meaning or usage except by reading the entire record, which includes both relevant and (for him) irrelevant information.

The more fundamental problem underlying this phenomenon is of course the diametrically opposite approaches of most dictionary authors (even in the case of technical dictionaries) on the one hand and of terminologists on the other. The EuroTermBank is basically a concept-oriented terminological database. Feeding such a database with dictionary material is bound to bring to light the fundamental differences between the lexicological and the terminological

approaches to special languages (or sublanguages). This indeed happened in the case of the EuroTermBank when it began to incorporate Elsevier's technical dictionaries.

In order to maintain the principle of "one-concept one-terra" (possibly with synonyms), any material which takes as its starting-point a physical manifestation (i.e. a string of characters) should be analyzed and split up if found to consist of a variety of matter. A term which can be used to denominate more than one concept should be given more than one entry in the database, each occurrence being separated from the others by a distinguishing feature, which may be a subject area classification code, a grammatical category code, or a different definition or context.

FUTURE DEVELOPMENTS

At present, the EuroTermBank is still in its infancy. The number of concepts stands at 200,000, the number of terms at 1.2 million. Although these figures will no doubt continue to increase, the quantitative changes will not be the most important ones. Various qualitative changes are necessary to raise the term bank above the status of an electronic dictionary library to a multi-purpose, multilingual information system.

To make the term bank a source of terminological data that can be used by both human beings and by NLP systems requires definitions and contexts that conform to a certain standard, and the addition of subject area codes and codes for morphological, semantic, syntactic and collocational features enabling NLP systems to draw certain conclusions, such as those necessary to disambiguate phrases and/or sentences.

At the present moment, the re-usability of information contained in term banks presents a problem. According to a recent Eurotra study, large terminology collections are "of little use beyond the character string of the term itself and its grammatical category" (J. McNaught & Blaise Nkwenti-Azeh, W. Martin & E. ten Pas, Feasibility of Standards for Terminological Description of Lexical Items, EUROTRA-7). Those which have been set up for machine-translation systems are too system- and application-specific to have wide applicability, while term banks set up for direct use by humans lack information vital to NLP systems.

In order to keep the term bank up to date and to augment the information in it with standardized definitions and contexts as well as grammatical data of a more elaborate nature than the basic parts of speech, it is necessary to develop tools which automatically or interactively extract new terminology from texts.

Other tools will make possible the comparison of duplicate data contained in the term bank, and also the comparison of such data with data from other sources, with the aim of improving the internal coherence of the collection and avoiding redundancy.

There are already tools available for the conversion of terminological material from different sources into the term bank format and for the conversion of term bank material into other formats, such as electronic dictionary software packages for use on PC (notably the Mercury/Termex software).

Perhaps more important than anything else, however, is the need for cooperation between the various parties involved in the creation, processing and management of terminological data. In other words, there must be cooperation between linguistic experts, computer experts and subject area specialists. Only by coordinating efforts can one avoid duplicating work and wasting valuable time and financial resources.

By encouraging such cooperation, and by developing the necessary terminological tools, Euroterm aims to contribute to the enrichment and augmentation both of its own term bank

and those of other organizations. This assists in the creation of the necessary infrastructure and resources for a European language research and engineering industry that will benefit not only the Member States of the European Community but also the entire world language community.

AUTHOR

Paul Nekeman, Director, Euroterm, P O Box 8, NL-6200 AA Maastricht, The Netherlands