Meta
Journal des traducteurs
Translators' Journal

META

Computer Aids to Translation

Friedrich Krollmann

Volume 26, Number 1, mars 1981

L'informatique au service de la traduction
Machine Aids to Translation

URI: https://id.erudit.org/iderudit/003353ar
DOI: https://doi.org/10.7202/003353ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Krollmann, F. (1981). Computer Aids to Translation. *Meta*, *26*(1), 85–94. https://doi.org/10.7202/003353ar

Computer Aids to Translation

FRIEDRICH KROLLMANN

1. THE STARTING POINT

The Federal Office of Languages of the Federal Republic of Germany (Bundessprachenamt, BSprA) is (when taking into account its subsidiary agencies) one of the world's largest language services. Its Department of Applied Linguistics has to produce, on a continuous basis, large numbers of translations into and from a multitude of languages. Just a few languages, however, constitute the lion's share, and they are: English with about 60%, French with just under 20%, and Russian with almost 10%. The BSprA deals primarily with technical texts of a high degree of difficulty, which must be translated with the highest possible level of accuracy. Only about 5% of all these translations are so-called "information transfers" which can be translated in a cursory manner.

Leaving aside the special circumstances in Canada, the translation problem for Germany is particularly acute. Depending on the subject field, approximately 60% of the world's technical literature is written in English, whereas only 5 to 10% appears in German. Furthermore, in practically all of the international organizations that are important for us, English is an official language; in most of them (apart from the European Communities), however, this does not apply to German (UNO, NATO, etc.) which again causes a huge quantitative translation problem.

Over 50% of our translations are regulations — mostly of a technical nature — for the Federal Armed Forces; they must therefore be produced in a manner that is ready for publication, such that they can be widely circulated and are accepted as legally binding. A good 80% of the remaining translations are bi- or multilateral documents which, as contractual documents or aids to political decision-making, must also meet the highest standards.

The volume of the texts to be translated induced the BSprA to study at an early stage possible ways of automating the translation process. The reasons for doing this were of a diverse nature:

— First and foremost, it was a matter of improving the quality of the translations;

— the translation output was to be increased, whereby

— the costs were to be reduced;

— the translation process had to be speeded up in order to ensure the topicality of the information contained in the translations;

— and finally, the use of uniform terminology was required within the linguistic service.

After the research and experiments in developing a system of fully automatic machine translation — initially conducted in some parts of the world in a mood of positive euphoria — came to an abrupt halt in the Sixties[1], a computer-aided translation (CAT) system was developed at the Federal Armed Forces Translation Agency (a predecessor of the BSprA), which has been operating successfully now for 15 years. Based on a terminology data bank, it has undergone continuous improvement and development.

This system is based on the empirically tested finding that searching for and using the correct technical term is the most time-consuming factor and the area in which the most mistakes are made, when technical texts are being translated[2].

It can therefore be regarded as a real and genuinely feasible alternative to the fully automatic high quality translation, which has not as yet reached fruition. Although machine translation processes have been and are continuing to be developed in the meantime which serve the purpose of speedy information transfer within the scope of current awareness systems, the process adopted by the BSprA is today still (and will probably continue to be for some time) the only path which can realistically be taken in the pursuit of high quality translations[3].

2. STRUCTURE OF THE DATA BANK

Along with the EURODICAUTOM system adopted by the Commission of the European Communities (or its predecessor developed by Bachrach and Hirschberg), the lexicographical information system (LEXIS) used by the BSprA was the first system to use a terminology data bank as a machine aid to translation. Basically, the following stages are either already in use or in the development phase[4]:

a) Compilation of dictionaries either in printed form or as computer output on on microfilm (COM).

b) Production of subject-field glossaries or technical word lists in alphabetical order, also by means of computer typesetting.

---

1. See Language Machines, Computers in Translation and Linguistics. A report by the Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences-National Council, Washington, D.C. (1966)
2. Krollmann/Schuck/Winkler: Die Herstellung textbezogener Fachwortlisten mit einem Digitalrechner — Ein Verfahren der automatischen Übersetzungshilfe. In: Beiträge zur Sprachkunde und Informations-verarbeitung, Verlag R. Oldenbourg, München, 5/1965.
3. Reviews dealing with this question are to be found in:
   — Bruderer, H.: Handbook of machine translation and machine aided translation, North-Holland Publishing Co., Amsterdam, 1977, and
   — Hutchins, W. J.: Machine translation and machine aided translation, in: Journal of Documentation 34, 1978, 129-159.
4. Krollmann, F.: An Interactive Terminology Bank, Background and Status, in: Computer Support to Translation, Proceedings of FBIS Seminar, Washington, D.C., 1978.

c) Direct aid to the translator in the form of text-related technical glossaries.

d) Means of retrieving secondary terminological information (according to bibliographical reference, definitions, contextual examples, etc.).

e) Reference to literature in the case of specialized terminological questions.

f) Record of translated texts.

The greater part of the vocabulary originates as feedback from translations. On the 1st of July 1980 there were over 1,4 million entries in the data base, of which

English/German accounted for	887 000
French/German accounted for	212 000
Russian/German accounted for	300 000
other languages (Italian, Dutch and Portuguese) accounted for	18 800

The same figures apply to entries in German/English, German/French, etc. More languages are to be added.

3. OPERATING THE DATA BANK

Various processing modes are available at all times for operating the data bank:

a) Changes to the data base

(i) Data acquisition. This is the direct input of new entries and deletions to the data base via display units. The input is keyed in directly onto a clearly formatted screen, and subsequently printed in a protocol. The entries are then ready for correction and release for inclusion in the data files.

In the case of deletions only a consecutive number in the data base with which every entry is marked, is input as the identification address.

Before new entries are included in the data base, a doublet search program eliminates entries with identical character strings, in order to avoid double entries. These entries are printed in a special list so that they can, if necessary, be newly processed by human intellect.

These data are then finally adopted in the overall data base at regular fortnightly intervals. All new entries and deletions are recorded in a separate list, over a period of six months, until a complete list appears again. The date is indicated in the case of every change to the data base.

(ii) File maintenance (on-line). This user mode allows direct changes to the data base. In this way changes can admittedly be made with a minimum of fuss, but access to this mode must be limited to a handful of persons, in order to prevent unauthorized persons from making alterations to the vocabulary at will.

(iii) Blanket changes to the data base (e.g. of certain subject fields, sources or even specific though recurring terms).

b) Direct access by the user

(i) Daily queries (batch processing)
Translators' query lists can be output in text-related or alphabetical order or both. At the same time, delimitations as regards language, subject field and source, as well as filing for the purpose of later use, and linkage with other queries are possible. The inventory is compressed (suppression of special characters including blanks), such that spelling variations can be dealt with (e.g. it makes no difference if the query is spelled
micro image
micro-image or
microimage).

(ii) Interactive mode (on-line). This convenient user mode allows pin-point queries or short lists to be answered directly. The contents of the screen can be printed out in the form of a list at any time and paging backwards and forwards, as with a dictionary, is also possible, as are all forms of delimitations.

Because of cost factors, however, display units for the on-line interactive mode are available only at a few central points in large linguistic services, otherwise they would not be used to their full capacity. Only when they can be combined with automatic word processing (editing) equipment, which can be connected to information systems for the translator it will make sense for the translator to have direct access to the interactive mode.

c) Further possible types of output

In addition to the possible types of output by means of the display screen and high-speed printer, there are other types of reproduction available :

(i) Computer typesetting, in particular phototypesetting. This relatively costly form of output will of necessity be chosen only in cases such as the printing of a dictionary for extensive circulation;

(ii) COM (Computer Output on Microfilm). This particularly handy output format has a special advantage, namely that the total vocabulary in the data base can be looked up anywhere without the use of the computer. Virtually the only costs are incurred by the first-and-final expenditure for a microfilm reader, whereas the microfilms themselves can be produced at an extremely cheap rate, and can be replaced by revised editions at relatively short intervals. This particular form of output nevertheless entails disadvantages as well, and these will be outlined later on.

d) Further possible forms of output for lexicographical purposes include :

— lexical concordances (according to character sequences in the data record). These are of particular interest if the overall data base is searched for a certain term which can also be a component of compounds;

— list based on a particular entry date.

4. ADVANTAGES AND DISADVANTAGES OF EACH USER MODE

The user modes described above serve first and foremost the translators working in the Federal Armed Forces Linguistic Service, but they are also available to other users, e.g. language teachers, and personnel with foreign language proficiency etc. Industrial language services, which in some cases do a considerable amount of translation work for the Federal Armed Forces, also participate in the use of this data bank.

For practical purposes, user modes 2 a) and b) above (dictionaries and technical glossaries) can be considered to be one and the same. Dictionaries and glossaries which are destined for extensive circulation and which must therefore be durable, are printed by means of phototypesetting. This is the case, for example, with dictionaries of the Federal Armed Forces; they are issued at intervals of several years and reflect (as far as possible) the vocabulary which is laid down as binding for the troops.

Technical word lists and glossaries which are only distributed amongst a small circle of users, or which are produced with a specific purpose in mind, are issued as high-speed print-outs and then duplicated by means of a lithopress. These specialized glossaries, for instance for a certain subject area, or according to a specific source, are relatively inexpensive and can be speedily produced, and virtually constitute expendable material.

The user mode which is most in demand in the BSprA itself, is the direct translation aid taking the form of text-related glossaries. In such cases, only those terminological queries which occur in a particular translation job are keyed in. These queries need not be composed exclusively of single words; on the contrary, relatively long compound words are perfectly acceptable. In answer to these queries the system produces the so-called text-related glossary, the terms being arranged in the order in which they occur in the text. In the first request stage only basic units of information are printed out. Further means of retrieval are available and they are gradually being put into practice. With regard to the output, the person making the query can stipulate additional restrictions, such as delimitation of the possible answers to certain subject areas, or deletion of all missing words. In these instances, the user receives only those answers which are relevant for the subject area with which he is currently dealing. However, the system does at the same time inform the user that, although there may be no answer available in the subject area he requires, there are nevertheless solutions which can be suggested for other subject areas. He can therefore question the system further.

The reason for restricting the user in this first request stage to purely lexicographical information lies in the fact that the overwhelming majority of the translators is already perfectly satisfied with this information. He is only searching for the "mot juste"; the meaning of the term is already clear from the source language. If, in reply to the first request, the user were to be provided with a great deal of additional information (definitions, contextual examples, references, for instance), then he would be overwhelmed with an abundance of

superfluous information at this initial stage of his work, whereby only a lot of printed paper and no new knowledge would be imparted to him. Access for the purpose of seeking additional information must, however, be possible. The presence of further information is indicated at this first stage.

Text-related technical word lists are processed in batch and output on paper by highspeed printers. This gives extremely good value for money and has the further advantage of constituing a medium with which the translator is perfectly familiar and on which he can make additional notes and comments, etc.

The most important advantage of the hardcopy method lies in the fact that it renders possible a considerable amount of feedback into the system. It is quite obvious that the computer will not know the answer to every single question. For a certain percentage of his queries the user (translator) will repeatedly come across the so-called "missing message", in other words a gap in the dictionary. Since no gaps are allowed, however, in the translation, the translator must run down the term he needs in the conventional way (research using specialized literature or conventional dictionaries, questioning of experts, etc.). This enables the gap to be closed immediately. The necessary terminological research only needs to be embarked on once, since the end-product of the research is immediately input into the computer and is thus available to all other translators from that moment. This induces a considerable rationalization process of the lexicographical aspect of translating.

In addition to this, a standardization of the terminology is also being achieved and this is a particularly desirable effect where extensive series of regulations are concerned. Since all translators participating in the translation of such series of regulations use the same data base, the uniformity of the terminology can be significantly better guaranteed. A further favorable factor in this context is the ability to produce a consolidated list in alphabetical order on the basis of the text-related technical word lists requested by several translators working on the same translation project, such that the translators can coordinate the target-language terminology they are to use by means of this very overall alphabetical list. A great advantage of this terminological coordination lies in the fact that new entries (e.g. from section 1 of a manual) are available (e.g. when work on section 2 is begun) due to the very fast updating techniques.

No matter how great the merits of the COM system may be, these advantages mentioned above are lost with the COM process. The human tendency to do only the minimum results in the fact that when the translator comes across gaps in the COM display, he will, at the most, supply a few terms, but never to the same extent as in the case of the hardcopy system, which simply seems to encourage feedback. The disadvantages of the hardcopy batch processing system lie in the fact that the necessary terminal equipment must be available where the translator is working, and that he must wait a few hours for the result of his query, because the queries are temporarily stored until a number of them can be processed together. The latter disadvantage can be overcome

without any difficulty by means of a flexible organization of working hours or, in individual cases, by making use of the interactive mode for urgent queries.

5. BACKGROUND STORAGE

It was stated above that, in most cases, the specialized translator, in his search for a target-language equivalent, is satisfied with the dictionary entry supplied by the system. However, he will sometimes require additional information, which must be delivered to him if he inputs further questions into the system. This second request stage in search of additional information results in information which is relevant to his problem and which is extracted from the background storage. The following information is primarily involved at this stage (of use not only to the work of the translator but above all to that of the terminologist):

a) Definitions
These become necessary when the translator wishes to delimit precisely the connotation of a word or compound word, or to compare two alternative expressions. It is only through the definitions that the meaning of such dictionary entries are clarified.

b) Contextual examples
These are particularly important for ascertaining the correct phraseology. In this way the translator learns, for example, that "eine Spannung anlegen" must be translated as "to apply a voltage". The meanings of words in their particular context are clarified by contextual examples.

c) Synonyms and antonyms, on which the translator often depends, if only for the purpose of achieving certain stylistic effects.

d) Cross-references to broader, related and narrower terms. These references are of special significance, not only to the translator, but also to the terminologist, because they enable him to work in the semantic field. They also represent an excellent means by which a conceptually-structured dictionary can be compiled.

e) References (for further reading)
Such information is considered by a great many technical translators to be among the most important which the system can deliver. A translator can only translate that which he understands. If the dictionary entry alone does not mean anything to him and the abstract definition does not help him to understand the operation or function, etc., of something the translator must be able to seek further information on the subject in primary literature (technical books, reference books, etc.). It must therefore be possible to recall references to such literature from the background storage.

All of this above-mentioned information available in the background storage invariably refers to one word or compound word in *one* language. In contrast to the foreground storage (LEXIS I), which is based on bilingual

dictionary entries for the various language pairs, such a background storage (LEXIS II) is only built up on a monolingual basis. All the above-mentioned units of information also invariably refer, therefore, to this one language. For the purpose of ascertaining the degree of congruence between two concepts, for instance, it is necessary to examine the dictionary entry, ascertained from the main storage, by comparing the definitions of the source- and target-language term respectively.

6. TRANSLATED TEXTS

Translations completed by the translation services or language services themselves represent an important source of material which can be consulted. Since it is to be expected that identical or at least similar types of texts and content are to be translated, where the clientele remains the same, it is important that the translator have recourse to previous translations.

For the purpose of administering the entry and exit of translation requests, a central administration is a necessity. Where a language service has a large number of subsidiary offices — as is the case with the BSprA — this administration has the additional task of keeping a check on the duplication of translations in order to prevent the same text being translated several times at the various subsidiary offices. This check on the duplication of translations is necessary, not only for reasons of cost, but also to ensure that several official translations of one and the same document are not produced analoguously.

The organization of such a central administration is, of necessity, based on the usual bibliographical references, which are to be found in virtually every archive or documentation system. But if translated texts are also to serve as material which can be consulted as an aid to later translation, then, with the various intended purposes in mind, this can take the following forms :

a) Former translations can be used for information purposes so that — as is the case with other elementary literature — the translator can familiarize himself with the technical content.

b) It should be ensured that the terminology is consistent with previous translations. At this stage, the terminology bank and the previous translations supplement each other since, under certain circumstances, the terminology bank offers several target-language versions for a source-language term and these are then clarified by the previous translations.

c) In many instances, whole passages of previous translations can, or even must, be adopted word for word in a new translation. This applies, for example, to quotes taken from treaties, etc.

Such an information system must, therefore, not depend exclusively on traditional search criteria as specified by the bibliographical categories, but should rather consist of additional categories of queries (e.g. short title, acronyms, references to primary documents, etc.).

Finally, aids for identifying documents by extracting the basic technical content, using the key-word system, are of particular significance. These include system concepts, project and equipment designations, names of contracts, etc. This system must be designed in such a way that translations can be relatively easily traced, even without bibliographical references.

7. THE IMMEDIATE FUTURE[5]

What developments can we now expect along the path from human translation to the utopian goal of fully automatic high quality machine translation?

The first stage to this end, that is the utilization of machine aids by the human translator (computer assisted human translation), can today be regarded as an accepted fact in most of the operational procedures in action today. This computer aid can take the form of a variety of types of application; here are the applications available now:

— Terminology data banks both on-line and off-line, as described in the preceding paragraphs.

— Utilization of automatic word processing systems for the input, storage, filing and manipulation of data.

A second step involves the interactive linking of word processing with lexical and other types of information systems for the benefit of the translator. I am of the opinion that during the immediate future development should be specifically concentrated on this field, since the possibility of combining word processing equipment and information systems stored for the translator — from the terminology bank to complete texts — will probably constitute the most effective aid to retaining the high level of quality of human translations.

The third step into machine translation process is that where we are dealing with the so-called restricted language situations, i.e. those cases where the text in its source language is subject to specific restrictions with regard to both syntax and vocabulary. Examples of application already virtually realized are, firstly, the Canadian project METEO where relatively stereotype phrases are translated from English to French (with the possibility of a relatively cheap subsequent improvement by human effort) and, secondly, the TITUS project of the Institut de Textile Français where abstracts from specialized journals are written with restricted syntax and a limited vocabulary in various European languages. However, other types of texts covering narrowly defined subjects which were admittedly intended as free text but which, in view of their narrowly restricted subject field and relatively impoverished language, can be equated with texts deliberately written in controlled language, are also candidates for machine translation. Operating instructions written in a relatively stereotype manner are an example of such texts (see TAUM AVIATION).

5. For information concerning the efforts of the Commission of the European Communities in this direction see esp.: Sager, J. C., The Computer and Multilingualism at the European Commission, Lebende Sprachen, Heft 3/79.

I personally do not think much of pre-editing normal free texts to convert them into controlled-language texts because, in my opinion, the expense incurred in this pre-editing stage would probably turn out to be so high that the costs involved in pre-editing, machine translation and subsequent post-editing would, at the very least, not undercut the costs of having the text translated by a human in the first place.

The last step leading to the fully automatic high quality translation is, of necessity, machine translation followed by human post-editing. In contrast to those cases where controlled language texts are involved, I consider this objective, bearing in mind the present stage of development of these systems, to be as yet unachieved in the case of genuinely free texts. When I say "genuinely free texts" I mean, incidentally, specialized texts and not texts of an artistic and literary nature which I would never entrust to the computer.

With regard to fully automatic translations requiring post-editing, it may be said that human translations need revising as well. Not even the work of good translators is perfect and in virtually all large linguistic services it is the practice for translations done by the translators to be revised by a revisor. This type of machine translation with human post-editing aiming at high quality translations will, of course, also have its chance one day. It is just a question of where the break-even point is. The output we can obtain today still necessitates such a high degree of effort on the part of the revisor that the advantages of the computerized stage are offset. The break-even point for the cost effectiveness of a process with a higher correction quota lies admittedly higher than for human translation (because a machine is, of course, much cheaper than a human translator); but as long as the revisor of a machine translation can more speedily do the translation himself than edit the machine output, then we still have some way to go before we reach the breakeven point.

Another crucial point is input. Is the text to be translated already on tape in readable form? Is the font of the text readable by an optical character scanner? Or — worse: is it necessary to key in manually hundreds of pages with an average error rate of 4-5%?

Such translations (though without or with minimal human editing) could, however, be used for current awareness systems where the output is only scanned for information purposes. In languages with other than the Latin alphabet this is most important because the normal educated European is unable to read even a bibliography written with Chinese or Arabic characters.

Achevé d'imprimé à Montmagny sur les presses des ateliers Marquis Ltée le 10 mars 1981