

# Data processing at the translator's service

Friedrich Krollmann

## Data processing at the translator's service

Friedrich Krollmann

Our theme this afternoon is one which reflects the influence of modern technology — or rather of various modern technologies — on an activity which is of interest to all of us — translation. The obvious question arising is: in which areas and to what extent can data processing be employed to make the work of the translating and allied professions easier? This question, however, requires immediate qualification. When I refer to electronic data processing I mean, of course, not the constantly expanding fields of application in which this new branch of technology is establishing itself but the limited field of (computational linguistics which, though a sub-branch of non-numeric data processing, is not necessarily to be taken as identical to the latter. The bipolar relationship between computational linguistics and translation science involves two components, "data processing for translation" and "translation science (including other branches of applied linguistics, such as terminology) for data processing". The short time available limits me, for all practical purposes, to saying a few words on the former of these two components.

Before the Congress commenced, I was informed that many of those present today come from the ranks of literary translators, a fact which does not make my task of elucidating on the above considerations any the easier. Data processing is a specialised techno-scientific field based on strict logical-mathematical procedures and on methods of representation which can be structured and explicitly formulated. This can by no means be said to apply to literary translation, and we can be thankful for this small mercy, for to apply the processes of data processing to literary translation would be to pave the way for the triumphant revival of prescriptive perfectionism and the classical formalism associated therewith. However, I am sure that not everybody present in this hall is a literary translator; indeed, I feel safe in assuming that a World Congress is per se

representative of all aspects of translation. Of course, the extent to which what I have to say on the use of computers will be of interest over such a wide range of translating activity will vary considerably according to the type of text to be translated. And when categorising types of text, the categories must correspond to activities ranging from, on the one hand, interlingual re-fabrication as, for example, in the case of lyric poetry to the direct transfer of spare-part and stock catalogues on the other, whereby the transition from the one extreme to the other is a continuous process through infinite intermediate shades. For translation is all of this.

One can also categorise texts according to whether the difficulties involved are difficulties of formulation — the extreme case being that of esoteric or highly emotional texts — or difficulties presented by large numbers of specialised terms. Of course, there will be borderline cases of texts to which both these criteria can be applied. To deal with such cases the translator will require not only a well-developed skill for formulation but also an extensive specialised vocabulary. But that wide sector of translation work in which the translator's freedom of formulation is severely limited covers not only the translation of catalogues but also the translation of technical and scientific texts. The further we move in the direction of specialised vocabulary texts, the more help we can expect from the computer *in the actual translation processes*, for the time being at any rate; conversely, the practical applicability of the computer declines the more formulation problems a text poses. This applies, as I have already indicated, only to the practical process of translation as such.

When we consider the theoretical investigation of the translation process from a linguistic point of view we are faced with an entirely different situation. However, practical limitations to the scope of my talk

prevent me from delving into the enormous potential and numerous methods afforded by computational linguistics which could be used to the benefit of translation science as such or to promote further research in this field. To mention but a few examples, such research could include investigation into the relationship between the surface structure and deep structure of languages, the representation of elementary relationships on an abstract level using the principles of transformational grammar, the linguistic analysis of isolated language phenomena (reflexive verbs, prepositions etc.). (In this connection it is worth mentioning that concordances, a useful and important tool in terminology work, can be produced by computer without difficulty). As last year's "International Conference of Computational Linguistics" held in Pisa once more demonstrated, research is tending unequivocally towards themes such as these.

From our point of view as translators, both translation and computational linguistics are components of applied linguistics. The German Gesellschaft für Angewandte Linguistik (Applied Linguistics Society) has a translation science section and an entirely separate computational linguistics section, thereby manifesting that it considers both branches to be component parts of applied linguistics.

Whether computational linguistics can be considered as a scientific subdiscipline in its own right remains, however, an open question: to the best of my knowledge there is no Chair of Computational Linguistics in existence, not in Germany at any rate. In the course of time, data processing as such has attained scientific autonomy under the alias of computer science, but does this mean that the use of computer purely to solve linguistic problems has already been accepted as a science in its own right? In other words, if we consider the computer to be no more than an aid or a tool, to be used as a dictionary is used when translating, then computational linguistics in the strict sense of the word does not exist. The computer and its various uses may be an asset to the researcher in linguistics but remain no more

and no less important than is electronics in the field of medicine (although the term "medical electronics" is gaining ground here).

We are not concerned here with giving a definitive evaluation of the importance to science either of translation or of computational linguistics. Suffice it to note that the linguist has found in the computer an instrument which enables him at least to cast some light on linguistic problems by applying mathematical (or, more precisely, Boolean) methods of approach.

The use of computers enables the disciplines of linguistics and computer science to be of mutual benefit:

1. thus the field of linguistics draws benefit from the vast and as yet unexplored potential of computer science. For example, computer science can give new insights into the synchronic and diachronic aspects of linguistics. Also of great advantage to linguistics are the increased speed of work, the vast data storage capacity, and the ability to re-run processes at will, all of which are afforded by the computer;

2. conversely, computer science profits from results achieved in the field of applied linguistics — and translation science as a branch thereof — in that it can use these results with a view to extending the applications of the computer in the field of non-numerical data processing. For example, experience gained from the use of computers in the linguistic field can pave the way for improvements in "user compatibility" (i.e. dialogue flexibility) of interrogation systems generally.

Having thus established that these two fields of knowledge, though apparently so dissimilar in their nature, can nevertheless be singularly complementary, we must proceed to examine the present state of the relationship between the two and how this position was attained.

Computational linguistics is not much more than a decade old and is therefore still in its initial stage of development. However, even this short history is characterised by setbacks and disappointments. Why should this have been so?

### *In the Beginning — Automatic Translation*

In computational linguistics, as in several other "modern" branches of science, many people gave way to the initial temptation to reach for the stars, only to be cast down to the solid ground of facts. Today's science of chemistry is a product of the medieval alchemy which strove to convert base metals into gold. Though we have still not succeeded in obtaining gold from lead, modern chemistry has provided us with a great number of products which — to my mind at least — are of more value to mankind than is gold.

Computational linguistics has undergone an analogous development. The pioneers, in a classic example of "jumping in at the deep end", attempted to reproduce human linguistic-intellectual thought processes by programming machines to produce fully automatic translations. The idea was conceived more than 25 years ago when Warren Weaver wrote in a letter to Norbert Wiener, the father of cybernetics, that he considered the translation process to be essentially one of decoding. The source text could be looked upon as an encoded version of the text in the target language and, as such, could be deciphered as a cryptogram.

The fundamental error inherent in this theory lay, of course, in the assumption that the surface structures of language are sufficiently clear and universal as to allow all linguistic phenomena to be defined explicitly and unequivocally. The fact that language is in reality a far more complex structure was completely ignored at that time. It is important, too, to remember that the first people to investigate the possibilities of automatic translation and who, in many cases, were of the conviction that the problem could essentially be solved using technical methods were not linguists but natural scientists. They maintained that it was possible to obtain empirically a system of linguistic rules, i.e. algorithms for the simulation models in the computer: far more important was the fact that its high speed of operation would enable the computer to cope with even the most extensive and intricate set of grammatical rules.

On the basis of this hypothesis and with a happy-go-lucky carefreeness a number of researchers and establishments, particularly in the United States set out in the late fifties and early sixties on the adventure of automatic language translation. Minor initial successes on the basis of simple sentence patterns enkindled sensational reports in the press, thus giving the public the erroneous impression that automatic translation was just around the corner.

These initial successes, however, were relatively easily achieved. For example, it is possible to write a machine grammar for a large number of sentences of simple structure and containing no syntactic or semantic ambiguity. This grammar can then be used as a basis for the translation of such sentences into another language. Unfortunately, this rather restricted machine grammar cannot be applied to sentences involving structural phenomena other than those already comprised. The original attempt to continue extending the compilation of rules on an empiric basis eventually floundered in a labyrinth of exceptions and exceptions to exceptions.

The surprising thing is that one of these pioneer programme systems, that devised by the research group founded by Leon Dostert at Georgetown University, is still in use, though no longer in Georgetown. Of the three automatic translation processes at present in use in the West, two are derived from the process developed at Georgetown University.

Soaring hopes soon gave way to sober reality. When tangible results failed to materialise, the Automatic Language Processing Advisory Committee (ALPAC) was founded in the USA on the initiative of the National Academy of Sciences and the National Research Council. In 1966 ALPAC produced a report entitled "Language and Machines" dealing with the practicability of automatic translation. This report drew the conclusion that, for the time being, automatic translation should be dropped as a field of research as it was impossible to obtain high-quality translations using an automatic process and the cost of research was disproportionate.

tionate both to the advantages to be gained and to the cost of man-made translations. This meant that research in this field came to a rather abrupt end in the United States, due primarily to the discontinuation of state support. Only the air force continued to sponsor work in the USA and today this service operates an automatic translation process in Dayton, Ohio.

Research in the field of automatic translation has still not recovered from the blow delivered by the ALPAC report. The most recent international congress on computational linguistics was held in Pisa in 1973. At this congress there were hardly any projects in the field of fully automatic translation to stand out against the background of general research into semantics, pragmatics, syntax and morphology. The few "new" projects introduced were subjected to sharp criticism in the discussion which followed and in some cases were torn apart. This proves that, for the moment at least, automatic translation is, by comparison with other computational linguistics projects, of no importance as a scientific topic.

The conclusion which has been drawn from the various abortive attempts at automatic translation is that it is not the technical but the linguistic aspect of the problem which presents and will for some time to come continue to present the insurmountable difficulties. The prerequisite for automatic translation is a grammar which not only unequivocally categorises the surface structure of a language (lexicon, morphology and syntax) but also exposes the deep structure, makes semantic ambiguities explicit and takes into account contextual and pragmatic associations, i.e. the circumstances surrounding transmitter and receiver — for these are all aspects which the human translator incorporates in his work. Since any automatic translation can only be as good as the grammar on which it is based, there is no chance of developing a high-quality, fully-automatic translation process in 1974 — or, for that matter, in the foreseeable future.

One might maintain that there are at least three automatic translation processes (Ispra [Italy], Oak Ridge [Tennessee] and Dayton

[Ohio]) at present in operation. But all three give results which are suitable for informatory use only and which are, for the most part, difficult to understand. To gain detailed knowledge of the subject matter, the texts which they have "translated" have to be completely re-translated by a human translator.

A further objection might be that the quality of a translation is a relative concept — this has recently been the subject of some considerable discussion — taking form only in the mind of the person who commissions or reads the translation. This is admittedly so. The quality criteria applied to the translation of an international treaty must perforce be other than those by which the quality of a technical article translated for information purposes only are judged. However, even in the case of a rough translation for information purposes there is a certain lower tolerance level beneath which the quality of the translation must not be allowed to sink, as, for example, as a result of misleading or erroneous information or of comprehension difficulties so great as to require excessive effort on the part of the reader.

Attempts have been made to compensate for linguistic deficiencies by pre-editing the text to be translated and post-editing the text in translation. There are several objections to this process (known as semi-automatic translation). Pre-editing requires of the editor that he be intimately acquainted with all the rules which the computer uses in its analysis of the original text. The question therefore arises: why not have the pre-editor produce the translation in the first place? Post-editing the product of the automatic translation process — this constituting an aid to the computer in synthesising the target language text — also appears to be uneconomical. Every one of the experts in this hall today is well aware that it is easier for the revisor to translate a text anew than to edit the inadequate output of a mediocre translator. And since the best the computer can achieve is the level of the mediocre human translator, revising the computer product is even more difficult, time-consuming and uneconomical.

### *Operational Procedures*

Having discussed what cannot be achieved by the electronic computer at the present time, let us proceed to consider what is already attainable.

Here we must call to mind the main purpose of data processing equipment. This is — broadly speaking—to process stored data and to make stored data available to the user as quickly as possible and in compliance with the individual wishes of the user. To establish the connection between the computer and the translator, we may say that the quality of any translation depends not only on the translator's aptitude and education but also on the amount of information available to him. It goes without saying that a translator who is equipped with copious reference works, dictionaries and archive copies can produce a translation more easily and more quickly than his colleague who [possesses only a minimum of sources of information. The electronic computer is a tool which enables a practically infinite volume of information to be stored, kept up to date and scanned in a matter of seconds. Everybody having access to the computer therefore enjoys the privilege of being able to extract specific data at will from a vast number of linguistic information units. I shall return to the question of unlimited access to such a source information — a very controversial point.

As we have seen, the fully automatic high quality translation is at present unattainable and semi-automatic translation — with pre- and post-editing — is uneconomical, but how can we use the computer as a supplier of information for the translation process? We can first of all establish that there are a number of possible applications which have already been perfected to the point of practicability and, which are indeed being exploited. In this context I could mention some examples of our own work at the Federal Language Bureau of the Federal Republic of Germany. Here I must also touch upon adjacent areas. Foremost in my mind are EDP work in the fields of terminology and lexicography and the linguistic aspects of modern information retrieval processes.

Let us first consider the potential afforded by EDP in the field of lexicography:

### *Dictionary Work*

The dictionary aspect plays an important part in all automatic translation research projects. For two reasons: firstly, a translation process which does not use a dictionary is inconceivable and secondly, the practical compilation and processing of a dictionary poses no great problems, especially when the work concerned is a whole-word dictionary without reduction to word stems. The main difficulties associated with the storage of a large multi-lingual dictionary are rather those of time, personnel and expense. However, once a voluminous (at least 100,000 entries) electronically stored dictionary has been established the emphasis of work lies on inventory maintenance, i. e. complementing and amending the data store to keep pace with new developments and acquisitions in the field of terminology. The European Coal and Steel Community and the Federal Language Bureau were pioneers in the compilation of substantial electronic terminological-lexicographical reference works. Compiling and maintaining an electronic dictionary can be greatly simplified if their scope is limited to simple and composite terms taken from specific technical fields, omitting colloquial language, phraseology, metaphors, figures of speech etc. Such electronic technical dictionaries may be entirely conventional in structure, differing hardly at all from printed works, except perhaps in that the intervals between reprints are shorter.

Of course, there is no reason to limit electronic dictionaries to being mere collections of words in the source and target languages. Examples of use in phrases and contexts, definitions, source references etc. for the words compiled in the dictionary can be stored in a background memory. The entry may incorporate a reference to the availability of such background information. This information can then be scanned and allocated as desired. In addition to purely grammatical qualifications, as for gender or part of speech, dictionary entries may also contain identification tags, e.g. reference to

subject field, or a hierarchical statement, e.g. "generic term", "subordinate concept" or "associated concept".

In my experience, the storage and maintenance of large volumes of terms — possibly in a number of languages — depends on and must be preceded by the establishment of central terminological data banks at language institutes large enough to cope efficiently with the work involved. The high cost of running such terminology banks — or even sections thereof — would prove prohibitive for smaller individual translation services. There have been a number of approaches to establishing such terminology banks, some of which have already been implemented.

If the terminological data bank is to cater optimally for the diverse requirements of its users, particular attention must be paid to the output side of the computer. There are a number of forms of output which can be used:

a) The computer may produce the vocabulary list required by the user (normally an extract from the complete inventory, e.g. one or more subject fields or sources) in the form of a normal printout. Such technical glossaries can be made to order within the space of one hour. However, though the quickest output medium, the direct printout is not suitable for mass distribution because of its limited reproduction capacity (up to four copies only), poor legibility and unmanageable format.

b) Computer-controlled phototypesetting offers a more elegant alternative. Using this process, any selection of words from the computer memory can be made available to the user in the form of a conventional dictionary (i.e. printed in major and minor case letters). When produced in sufficient numbers, these dictionaries have the advantage of being relatively inexpensive. Their prime asset is, however, that they can be kept up to date.

c) A third possibility is to output the entire inventory on microfiches. These are extremely cheap to manufacture and reproduce but to be able to use them the subscriber, who will receive revised sets at regu-

lar intervals from the data bank, will require a reader. However, microfiche readers will in any case soon be standard office equipment. Since new sets of microfiches can be distributed at short intervals, the subscriber has permanent access to an extensive and completely up-to-date terminology collection.

d) A further method of showing the inventory stored in the computer memory at any time involves the use of visual display units directly connected to the computer. However, high transmission fees in Europe (in contrast to the USA) make teleprocessing extremely expensive, with the result that this method is economical only in the case of major subscribers with a permanent demand for data or — to exaggerate a little — for subscribers with premises in the immediate vicinity (say a 600 meter radius) of the central computer. Under these circumstances, this method must be thought of as local rather than remote processing. From the point of view of economy, visual display units should be made available to terminologists and lexicographers rather than to translators, for whom the considerably cheaper microfiche output will normally be fully adequate.

#### *Text-related Glossaries (Computer-assisted Translation)*

The processes I have just mentioned are designed basically to produce complete dictionaries and alphabetical technical glossaries, i.e. the translator's general tools, with the aid of the computer. However, electronic terminology banks can also be used as direct aids to translation.

As already mentioned, the storage capacity of a computerized dictionary can be immense. Therefore, if the translator always had to work with this enormous volume of data, no matter which form of output presentation were chosen, he would be faced with a large proportion of redundancy, which is not necessarily desirable. The computer makes it possible to generate another type of word list in a minimum of time. These lists, which we might term "disposable" glossaries, include only such terms as appear in a certain text and for which the

corresponding target-language terms are sought. We have called these lists "text-related glossaries" and for a number of years we at the Federal Language Bureau have now been working with this medium with considerable success. The lists may be considered to represent a selected, relevant sub-inventory.

The practical application of the text-related glossary procedure is as follows: the translator underlines all the difficult or unfamiliar terms (generally specialised terminology) which he encounters during his initial reading of the text to be translated and compiles them into a typewritten list. These terms are fed in as search questions via a data medium. The computer then locates the terms in the data bank and outputs them, together with their target-language equivalents, via a high-speed printer. The retrieval process may be limited as desired by the introduction of qualifying criteria, e. g. by specifying a certain subject field (in practice, that of the translation).

The final product is a list in which the source and target-language-terms appear side by side *in the order in which they occur in the text*. Simplifying the procedure a little, we can say that all the translator has to do is to compare his original text and his text-related glossary as he proceeds with the translation. Of course, the glossary cannot do more than assist the translator in making decisions, for in many cases it offers a number of possible target-language equivalents, as do other dictionaries also. Nevertheless, we have found from our on-the-job studies and many years of translating experience that text-related glossaries can be a decisive aid towards improving translation work both quantitatively and qualitatively. The usefulness of these glossaries increases in proportion to the frequency of technical terms in the text to be translated; conversely, they are of less assistance when the difficulties posed by the text are principally problems of formulation. The procedure in its present form is of little value for literary translation work, though this is not to deny it some interest as a model. It is recommended primarily for use by large-scale translation services with a technical-scientific bias,

where a number of translators are often engaged in work on a single major project and the co-ordination of terminology presents special difficulties. In such cases it is also possible to combine the text-related glossaries produced for the individual translators and produce a single alphabetical list from which each translator, and in particular the revisor, is able to see which terms appear in the texts of the other members of the team. The alphabetical list can then be used as a basis for discussion in a vocabulary conference with a view to establishing a uniform target-language terminology for use in the project.

At present, this procedure still involves one time-consuming manual phase: the translator must himself extract the vocabulary I required from the original text, and the list so obtained must be transferred into the I system by an operator either directly via I visual display terminals or indirectly via data media. Various institutions are at present engaged in attempts to develop automatic optical character scanners for full texts. As soon as such instruments have reached a satisfactory state of perfection, it will be possible to read texts into the computer automatically without human intervention, thus eliminating the detour via manual extraction and input. Extracting the specialised vocabulary from the texts would present no problem as automatic generating processes have already been developed. Broadly speaking, these processes are based on the stopword principle, i.e. the computer "suspects" all words and groups of words between certain stopwords — articles, prepositions, conjunctions, modal verbs etc. — as being simple or composite technical terms. Punctuation marks — with exception of the hyphen — are also considered to be stopwords. The series of words found between the stopwords are then reduced by the computer according to a fixed set of rules and compared on the "longest match" principle with the entries already in the computer. This procedure ensures that the terms finally extracted are meaningful words or composita.

One such process has already been tested in a number of experimental runs which gave

a surprisingly high recall ratio. We consider these results so encouraging as to justify our keeping this procedure under observation with a view to adopting it should an optical character scanner for all types of print and script be perfected.

#### *Text Banks*

There is yet another field in which the computer can be used for linguistic purposes, a field in which the computer is of benefit primarily to linguists, dictionary compilers and grammarians but is also of interest to the literary translator. The immense possibilities of this field of application have not yet been exhaustively sounded. I am referring to the establishment of text banks in which data are stored in a computer memory in the form of complete texts. There are a number of text collections of this type already in existence, though these were established for a variety of purposes. Examples are the Brown corpus (one million running words) established in the USA and containing American English texts and a new major corpus for British English texts, recently established at Lancaster. The international society for "Literary and Linguistic Computing", established in 1973, is also worthy of mention in this context. In Germany, large text collections are maintained at the Institut für Deutsche Sprache (Contemporary German research project), at the University of Saarbrücken (under Prof. Eggers), at the University of Marburg (German Linguistic Atlas) and by the LIMAS research group in Bonn; but these are only a few of numerous examples.

The greatest advantage is to be obtained from the computer when it is used as a means of storing, sorting, rearranging and counting data, i.e. for procedures for which it is particularly suitable and in which its use does not give rise to uneconomical expenditure of time and energy but rather helps to save time and effort. The feasibility of many processes is directly dependent upon the use of the computer, since the effort they involve represents a generation of human work. This is particularly true for linguistic-statistical work, a field in which it is possible to distribute the work-load

rationally between man and machine: the machine uses the theory of combination to produce automatically large volume of data, from which the language researcher then extracts the relevant information units and combinations. Such linguistic-statistical work is not limited to word level; in Australia, Beebe has recently been researching in the field of the statistical collection and analysis of English syntagms. Such research is of value not only to linguistic theory but also to applied linguistics, i.e. to the translator and to translation training, since it supplies verifiable information on the vocabulary and stylistic resources used by various authors.

One of the most obvious cases in which the computer can be used to store complete texts presents itself when standard texts are to be translated. The principle examples of standard texts are catalogues, operating instructions and descriptions of equipment which have to be brought up to date at regular or irregular intervals. Such texts, requiring frequent retranslation but based on a previous edition, may be filed on magnetic tapes and stored in the computer. Since many large translation offices already use magnetic tapes to control a typewriter or typesetter (e.g. a composer system) when producing target-language documentation, the same tape can subsequently be re-recorded onto an archive tape. The advantage of such an EDP tape archive over the otherwise cheap and versatile microfilm archive is that all subsequent alterations to the text can be entered on the tape by computer, though the difficulty of selecting a specific sentence or word has not yet been completely overcome. The microfilm is a lifeless piece of filed information; the magnetic tape is dynamic — each new version could be used to control the typesetter as it appeared. This process relieves large translation services which have a high proportion of alteration work of such repetitive and time-consuming tasks as producing new fair copies, typefaces and paging. This work can be left to the computer.

#### *Conclusion*

As we have just heard, the computer can be of considerable assistance to the translator.

It can relieve him of routine work but cannot replace him completely. The translation process remains entirely in the hands of the human translator, so there can be no question of "de-humanising" this intellectual activity. Rumors to the effect that the electronic computer will make the human translator superfluous are disproved by the fact that all attempts to develop a fully automatic translation system have failed. The translator has no reason to fear the computer.

In conclusion I should like to mention a few aspects which limit the computer's suitability for use in translation work:

— only large-scale translation services in industry and the public sector can afford to store terminological data and translation aids for their translators;

— conversely, individual translators and smaller translation bureaux can, for the most part, not afford a direct connection to a linguistic data bank because the fees for such data links are prohibitively high, at least in Germany;

— to mail enquiry lists to a central data bank costs time; the majority of translators working in the private sector of the economy have to work to deadlines;

— a microfiche subscriber service can only be provided by the proprietors of a large, efficient data bank.

Since the expense in terms of finance, time and personnel involved in setting up and running a central data bank is very high — and we have gained considerable experience to this effect — such a data bank can only operate on a supra-regional basis.

This centralised position gives the data bank a certain monopoly over the translators, not only from the economic point of view but also by virtue of its linguistic privileged position which enables it to take on prescriptive aspects. The economic monopoly may be averted by running the data bank as a public utility on an individual cost-sharing basis. On the other hand, the data bank will inevitably exercise a certain standardising influence on terminology which, though desirable in the field of technology, would hardly be opportune in other fields such as journalism and literature. Unfortunately, these deliberations will not prevent the computer from insinuating itself further into our private lives and into our professional lives as translators. However, since we are still in the initial stages of development we are in a position to fashion this development and to manipulate it along propitious lines. I consider this to be a task in which the translating profession should participate actively, for otherwise we forfeit the right to complain if, one day, the spectre of dehumanisation which looms over our professional activity should actually materialise.