[*Meta* XVI, 1-2]

Linguistic data banks and the technical translator

In the late 1950s and also in the early 1960s hopes were running high that translation by computer could be achieved. However, linguistic experts were not very much surprised when the first enthusiasm was followed by disillusionment. Many readers will be familiar with the report *Language and Machines* published by the Automatic Language Processing Advisory Committee (ALPAC) in 1966 and based on a series of inquiries into the state of the art. There is as yet no consensus of opinion among researchers whether or not the drastic criticism voiced by the Committee on the efforts to realize fully-automatic high-quality machine translation is justified; on the other hand, the Report makes recommendations which many consider to be guidelines and on which I should like to center this contribution. For an improvement of the quality of translation and the translation process, ALPAC submitted nine proposals on which future work in the field of linguistic data processing should concentrate. Of these nine recommendations six are of interest to the translator of technical or scientific texts :

1. [...]
2. Means for speeding up the human translation process ;
3. [...]
4. [...]
5. Study of delays in the over-all translation process, and means for eliminating them, both in journals and in individual items ;
6. Evaluation of the relative speed and cost of various sorts of machine-aided translation ;
7. Adaptation of existing mechanized editing and production processes in translation ;
8. The over-all translation process ; and
9. Production of adequate reference works for the translator, including the adaption of glossaries that now exist primarily for automatic dictionary look-up in machine translation.

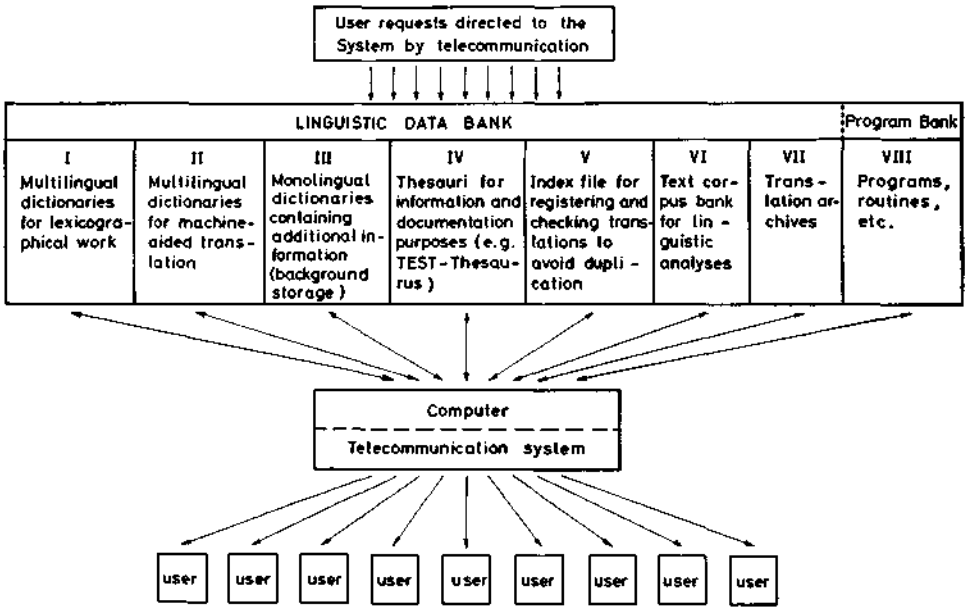
The above six items, in one way or another, bear on the problem of setting up a linguistic data bank for technical translators. Indeed, what seemed to be

revolutionary in 1960 has now become every-day practice for an ever increasing number of large translation services, namely the use of computers for streamlining and mechanizing linguistic processes, possibly with a view to establishing a linguistic data bank. I should like to limit myself only to those processes that are relevant to the production of technical translations including such fields as lexicography and text processing.

A MODEL OF A LINGUISTIC DATA BANK

In describing the following model of a linguistic data bank I am referring to systems which are either operational or possible. Other combinations or variations of data bank configurations, differing from those presented here, are of course imaginable and feasible. But in most cases they would concern data banks in which data for other linguistic purposes are stored, such as data for syntactic or semantic analyses, etc.

As we are interested only in a data bank with its branches or subbanks that serve as a central aid to the translator, the following model should meet our requirements. This model of a linguistic data bank for the translator consists of eight branches or subbanks. one of them being exclusively a program bank from which various subprograms, program blocks, elements or routines may be called up to operate the other data-subbanks or run linguistic processes according to specific needs.



Model of a linguistic data bank

The other subbanks are divided as follows : I. Multilingual dictionaries for lexicographical work ; II. Multilingual dictionaries for machine-aided translation ;

III Monolingual dictionaries whose entries contain additional information on Subbank I entries, particularly definitions (so-called background storage);

IV. Thesauri, particularly for information retrieval and documentation purposes ;

V. Index file for registering and checking all translations in order to avoid duplication ;

VI. Corpus bank for analyzing texts according to linguistic criteria ;

VII. Translation archives.

These subbanks should not be seen isolated from each other. On the contrary, there should be a steady flow of information between the subbanks as various combinations of data requests are envisaged. Thus, Subbank III is to serve simultaneously as background storage for Subbanks I and II. As everybody knows, the mere « word equation », *i.e.* source language term and target language term, as contained in a conventional bilingual dictionary, is often of little use to the lexicographer or translator, unless it offers additional information, normally definitions or contextual examples. Data stored in Subbanks I and II should in each case also provide the direct search address for Subbank III.

CONTENTS AND PURPOSE OF THE SUBBANKS

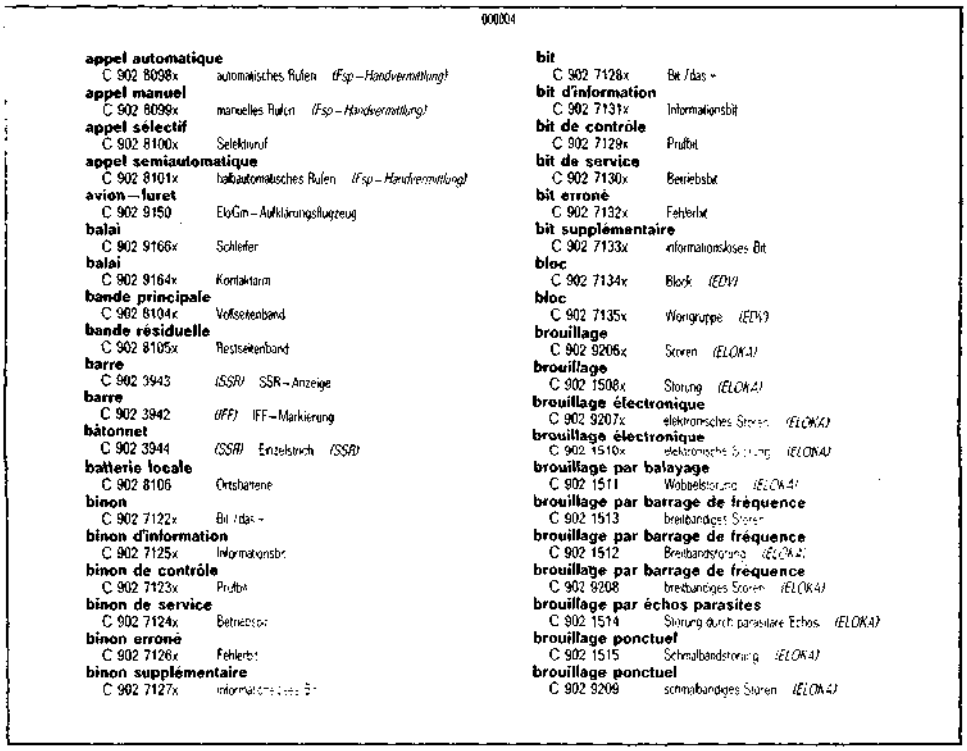
The various subbanks can be used for the following possible functions :

Subbank I

a) To take the example of the Bundessprachenamt (Federal German Bureau of Languages) : Among many other tasks in the field of linguistics, this organization is charged with the collection of technical terms, especially new technical and scientific nomenclature. To handle this ever increasing number of data by means of conventional card files would be hopeless as one can easily imagine. The contents of this first subbank could be referred to as a sort of superdictionary to which users direct their requests. The print-outs of these requests are sorted : 1. alphabetically ; 2. by subject field ; 3. by source, *i.e.* origin of the terms. Any combination of these criteria is possible.

b) The Bundessprachenamt is concerned with issuing dictionaries, glossaries and word lists for federal agencies, particularly the Federal Armed Forces who have the most varied needs for such material with respect to language, subject field, layout, etc. At times it happens that almost every day a request comes in to produce some kind of glossary. If, for example, the Federal Air Traffic Control Service urgently needs an English/German dictionary, then it must be possible to extract from the main dictionary, at short notice, all English/German word pairs covering the subject field of « air traffic control », supplemented probably by some related subject fields like meteorology, navigation aids, etc. This continuous handling of huge quantities of linguistic data would no longer be feasible without Subbank I.

c) For the purpose of coordinating technical terminology special Terminology Committees have been established whose central management is a responsibility of the Bundessprachenamt. These committees, of which there are eleven, have to see to it that all agencies and firms working for the Federal Government use



000804

appel automatique
C 902 8098x automatisches Rufen *(Fsp – Handvermittlung)*
appel manuel
C 902 8099x manuelles Rufen *(Fsp – Handvermittlung)*
appel sélectif
C 902 8100x Selektivruf
appel semiautomatique
C 902 8101x halbautomatisches Rufen *(Fsp – Handvermittlung)*
avion-furet
C 902 9150 ElyGin – Aufklärungsflugzeug
balai
C 902 9166x Schleifer
balai
C 902 9164x Kontaktarm
bande principale
C 902 8104x Volksetenband
bande résiduelle
C 902 8105x Restseitenband
barre
C 902 3943 *(SSR)* SSR – Anzeige
barre
C 902 3942 *(IFF)* IFF – Markierung
bâtonnet
C 902 3944 *(SSR)* Einzelstrich *(SSR)*
batterie locale
C 902 8106 Ortsbatterie
binon
C 902 7122x Bit /das –
binon d'information
C 902 7125x Informationsbit
binon de contrôle
C 902 7123x Prüfbit
binon de service
C 902 7124x Betriebsbit
binon erroné
C 902 7126x Fehlerbit
binon supplémentaire
C 902 7127x informationsloses Bit

bit
C 902 7128x Bit /das –
bit d'information
C 902 7131x Informationsbit
bit de contrôle
C 902 7129x Prüfbit
bit de service
C 902 7130x Betriebsbit
bit erroné
C 902 7132x Fehlerbit
bit supplémentaire
C 902 7133x informationsloses Bit
bloc
C 902 7134x Block *(EDV)*
bloc
C 902 7135x Wortgruppe *(EDV)*
brouillage
C 902 9206x Stören *(ELOKA)*
brouillage
C 902 1508x Störung *(ELOKA)*
brouillage électronique
C 902 9207x elektronisches Stören *(ELOKA)*
brouillage électronique
C 902 1510x elektronische Störung *(ELOKA)*
brouillage par balayage
C 902 1511 Wobbelstörung *(ELOKA)*
brouillage par barrage de fréquence
C 902 1513 breitbandiges Stören
brouillage par barrage de fréquence
C 902 1512 Breitbandstörung *(ELOKA)*
brouillage par barrage de fréquence
C 902 9208 breitbandiges Stören *(ELOKA)*
brouillage par échos parasites
C 902 1514 Störung durch parasitäre Echos *(ELOKA)*
brouillage ponctuel
C 902 1515 Schmalbandstörung *(ELOKA)*
brouillage ponctuel
C 902 9209 schmalbandiges Stören *(ELOKA)*

Sample page of a machine-produced dictionary French/German

a uniform terminology. This is particularly important when a number of firms share in the manufacture or maintenance of some major item or equipment; for only a standardized terminology can guarantee that a particular item or process will consistently be given the same name in all technical manuals of all industrial firms or users. The number of terms passed and standardized by these committees until now (November 1, 1970) amounts to about 70 000. If this terminological co-ordination between firms and users as well as between the various committees themselves had to be done manually, the number of personnel and the amount of time required would be extremely high.

d) In connection with Subbank I there is a special process which is worth mentioning. A special computer program adds printing instructions to linguistic data as they are being extracted from the main storage and transferred to an intermediate magnetic tape. This magnetic tape in turn controls an electronic photo compositioning machine (the so-called « Digiset » system, short for *digi*tal *set*ting*)*. The result is a concise print with a variety of types and sizes to choose from : medium, bold, italics, etc., which can bear any comparison with commercial type-setting.

Subbank II

a) Part of the linguistic material stored in the main dictionary of Subbank I is processed according to certain criteria to give a new special data file in Subbank II which serves as a direct machine aid to the translator. Subbank II is a concise extract

of Subbank I. In an article published in Number 5 of *Beiträge zur Linguistik und Informationsverarbeitung* we have shown that the majority of errors made by a technical translator concerns technical terminology and that he spends a great deal of his time searching for the correct term. Furthermore, we have demonstrated in experiments that a computer is extremely well suited to serve as a machine dictionary for the translator. Underlying this idea is the fact that, since a fully-automatic high-quality machine translation will not be feasible in the immediate future — in spite of the fact that primitive translations by machine are already possible — the computer can easily produce the terminological skeleton of a translation. This process is called « text-oriented glossary » which is widely-known among experts so that I need not dwell on it in detail. Under this process all technical terms are listed in the order in which they occur in the text to be translated, together with target-language equivalents and other pertinent information. In other words, the text-oriented glossary constitutes a special dictionary tailored to and valid only for the text to be translated and is discarded after completion of work. The word equations either follow the order of their occurrence in the source text, which is normally the case, but they may also be listed alphabetically, which is particularly convenient if a large translation project has to be split up among several translators. With the aid of this special alphabetical glossary it is possible to coordinate, among the various translators, the technical nomenclature right from the beginning of the translation throughout the entire translation process.

b) These text-oriented glossaries are not only an effective means of aiding the translator, they can also serve as a reading aid to the scientist or technician who has a working knowledge of the language and a full command of the subject matter involved, thus enabling him to read a foreign-language text in the original; as he is progressing with reading he can follow the vocabulary unknown to him which otherwise would handicap him in understanding the text.

c) The text-oriented glossaries have been in use for several years now and are a great success with the translator, once the expected psychological hurdle had been taken. One of the positive side effects of the procedure is that it forces the translator continuously to evaluate and check the terminology of the translated text, which results in a steady growth of the stored main dictionary (making up some 60 000 new dictionary entries at the Bundessprachenamt per year). Apart from coordination of terminology proper, also the amount of work, cost and time required for evaluation is reduced since each terminological problem is tackled and solved only once, the solution then being accessible to all users of the system.

d) At the moment, the procedure has the disadvantage that the requests for terminology have to be prepared manually. While the Bundessprachenamt has already developed a fully automatic requesting system, called « Autoquest » (short for *automatic request)* under which the technical vocabulary of a text is auto matically generated by the computer (the criterion for a sensible combination being the occurrence of the compound or word in the dictionary), this procedure can only be applied if the texts are recorded on a machine-readable medium, such as magnetic tape, paper tape, etc. As this is normally not the case, an optical

character reader for all current types of print would be required which unfortunately is not yet on the market.

Subbank III

This bank should contain that linguistic information which is required in addition to the word equation proper, such as definitions of contextual examples. The contents of this subbank, which might be called linguistic background storage. could as well be stored in Subbank I. But this would have the disadvantage that all the data of Subbank III would also have to be moved during most of the operations involving Subbank I, the main dictionary. Thereby, the required storage space and computing time would soar, while the speed of operation would be reduced. On the other hand, when requesting data from Subbank I it would be feasible to call up background information of Subbank III by a « Go-to-statement », producing the same effect.

Subbank IV

Subbank IV might contain a technical thesaurus mainly to be used for documentation purposes and information retrieval. As a rule, such a thesaurus is compiled on a monolingual basis. However, there is a project in hand in Germany which departs from this principle. This project comprises the translation into German of the Thesaurus for Engineering and Scientific Terms (TEST) of the U.S. Department of Defense which is performed under the direction of the German Institute of Documentation. The outcome of this work will be a bilingual thesaurus. As is the case with most thesauri, also the TEST-Thesaurus is structured hierarchically making possible the grouping of a larger number of terms under broader and narrower terms. Even though this thesaurus serves mainly for documentation purposes, its usefulness for the translation process will be considerable as the translator — via a descriptor — is led to a number of related terms.

In the thesaurus exact English-German cross-references can be provided by means of numerical addresses. Difficulties arise, however, due to the fact that in a bilingual thesaurus the terms or notions of the source and target languages are not always completely congruent. For instance, a term or notion in the source language may have a much wider scope of application than the corresponding target language term, so that a one-to-one relation would no longer be guaranteed. This is, of course, a universal problem.

Subbank V

The Bundessprachenamt — like other large translation agencies working on a national or supranational level — has to maintain a central index file of all translations — not only of in-house translations but also of translations produced by other services. In the case of the Bundessprachenamt this file is organized according to a variety of criteria, depending on whether the translated text is an article from a journal or some other publication, a document of an international organization, a technical manual, etc. The main purpose of this file is above all to avoid duplication of translation, and in the past years it has met this requirement in an excellent way. It is intended to computerize this file by using appropriate library programs as a model.

Subbank VI

For research work with a view to mechanizing further the translation process, but also for quantitative linguistics such as frequency counts, for compiling textbooks, etc., a subbank is needed in which homogeneous texts, selected on the basis of certain criteria and coded accordingly, are stored. It would go beyond the scope of this contribution to dwell on the linguistic possibilities such a bank offers. The Bundessprachenamt has used the capabilities of such a bank for linguistic quantitative research as such analyses are invaluable for compiling textbooks needed in the language courses which the Bundessprachenamt conducts for adults.

Subbank VII

As the seventh possibility of using a computer for the purposes of a linguistic data bank one can think of the transfer of a considerable part of translation archives to computer tape for temporary storage. This should pose no problem in those cases where, in the translation process, texts are recorded on magnetic tape to be printed later on, *e.g.* with the IBM composer-system.

Temporary storage of translated texts on computer tapes offers advantages in the following fields : a) Due to the fact that large numbers of typewritten pages can be stored on a computer tape, the space required for the translation archives would be extremely small; *b)* As compared to microfilms the system would have the advantage that, via descriptors or keywords, large batches of text could automatically be searched for particular passages and then be displayed on video screens as an aid to the translator ; *c)* For revised new editions of translations only the changed passages would have to be retyped. Insertion of changes and corrections into the old text would automatically be done by computer, as well as renumbering of pages ; *d)* In connection with Subbank IV, automatic indexing is feasible.

CONCLUSION

1. A linguistic data bank with its subbanks as described above, can be a valuable help to the technical translator, while, on the other hand, the translator himself constantly contributes to keeping it up to date. In addition to the trans lator, there are a number of linguists to whom this data bank renders valuable service. They are : a) Terminologists and lexicographers (Subbanks I and III, but also IV). *b)* Linguists and language teachers (particularly Subbank VI) as pointed out in discussing Subbank VI. Apart from linguistic quantitative analyses, future application of this subbank is seen particularly in the preparation of teaching material, for which the keyword-in-context method can be used. For this project also texts stored in Subbank VII may be processed unless they are too heterogeneous, *c)* Subbank IV is of particular interest to the documentalist, that is to both the indexer and the abstractor.

2. When establishing a data bank of this kind, due consideration should be given to the hardware which has to meet the particular requirements. Naturally it would not pay to install and operate a large computer just for a translation

service only. Thus the solution would lie in some time-sharing arrangement coupled with a teleprocessing system.

3. A linguistic data bank of this order is worth the expenditure only if a large « clientele » is using it. Similar to a bank, business with a linguistic data bank pays only if, figuratively speaking, a large enough number of accounts is maintained with extensive transfer of money to and from these accounts. Other wise, the very high expenditures for technical equipment and computing time would be out of proportion to the advantage that can definitely be derived from such an installation. However, if — as in the case of the Bundessprachenamt — hundreds and thousands of requests are directed to this data bank day after day, then the turnover is so large that it can no longer be handled manually ; thus, the use of a data bank is not only economical but has become a necessity.

Furthermore, it should not be overlooked that apart from the hardware, which can be rented, also the data stock should be considered which represents an extensive investment of brain work. This means that a data bank should only be established when the amount of data to be stored has become too large for manual handling. A few tens of thousands of terms which can be accommodated in a card file do not warrant the establishment of a data bank.

4. Several years ago when the European Community for Coal and Steel and the Bundessprachenamt started independently to store linguistic information in a computer, at first on a modest scale, later broadening the basis more and more, pioneer work was done, while today it has almost become a status symbol of any self-respecting major translation service to work with some sort of EDP-equipment.

What is now necessary, is a cooperation between these activities particularly for the benefit of those users who cannot afford an EDP of their own. It is the very idea of such data banks that they should be available to a large number of users with a variety of needs. The outlook for such projects is very promising. The Symposium on Terminology held in Germersheim in 1968 has given an impetus towards a cooperation between the various groups working in this field on a national and supranational basis.

FRIEDRICH KROLLMANN