[From: *Overcoming the language barrier*, 3-6 May 1977, vol.1 (München: Verlag Dokumentation, 1977)]

A TERMINOLOGY DATA BANK FOR TRANSLATORS METHODS OF INTERROGATION IN THE TEAM SYSTEM

J. Schulz,

Language Services Department of Siemens AG

Abstract

The TEAM system was developed in the Language Services Department of Siemens in Munich. The core of the system, which is intended to provide machine aids to translation, is a multilingual terminology data bank, in which information of different nature, origin and quality is stored and processed. The aim is to make this information available in a suitable form to translators and other interested parties.

This paper discusses the stored terminology data as well as some aspects of their retrieval, in particular the following two interrogation methods: 1) interactive interrogation and access methods via terminals and 2) automatic interrogation, i.e. the processing of machine-readable texts, for instance with a view to the automatic production of glossaries.

1. TERMINOLOGY AND PHRASEOLOGY

One of the main functions of the TEAM system is the mechanized provision of translation aids, i.e. first and foremost the provision of the terminology required for the translation of specialized texts in the source and target languages. The system uses a multilingual terminology data bank in which a variety of terminology data are acquired, stored and processed centrally in order to satisfy the various demands made on the system. It aims:

1. to supply all interested parties on request with information on specialized concepts and their equivalent terms ('translations') in the most important (for the time being, European) languages. This information is to be compiled on the basis of comprehensive lexicographical/terminological scanning of the relevant specialist publications for terminology and wide-ranging cooperation with other terminology centres.

2. to ensure the use of uniform terminology by translators (in German and in the relevant foreign languages), particularly in view of the constant growth in volume and significance of technical documentation. This aim necessitates standardization of terminology in commercial, national, and international organizations, the results of which are stored in the data bank.

1.1 Concepts and terms

In accordance with the recommendations of Professor Wüster [1] and DIN 2330 (concepts and terms) [2] the terminology data bank is composed of information units, each comprising a concept which is generally defined. A single concept is rendered by different terms in the different languages.

However, it is not always possible to adhere rigidly to

this principle, given the constraints of everyday translation work, the result is therefore that monolingual or multilingual items of terminology as well as phraseology which have undergone varying degrees of lexicographical and terminological processing and clarification (ranging from single-word or context equivalents compiled more or less haphazardly by the users to standardized definitions with internationally approved equivalents in the various languages) will be stored and made available in the central data bank. However, - and here the underlying theoretical approach retains its validity in the practical context of a terminology data bank for translating purposes - the primary function of the system is to store equivalent terms for specialized concepts (technical and other). These equivalent terms come from an extremely wide range of sources. To a great extent, they are compiled by the translators themselves from specialised literature and the technical documentation on a wide variety of products, and cover not only the relevant subject areas in the narrow sense (e.g. a firm's range of products), but also the underlying and neighbouring subject areas as well as the areas of application.

These equivalent terms are sometimes to be found in the source documents with definitions, explanations or illustrations, but often the meaning of a particular concept can only be discovered from the context in which it appears and on occasion only after discussion with the author of the text (e.g. the designer). In the case of works parts lists or spare parts catalogues, the meaning of a concept is evident from the classification number consequent on a particular scheme of classification. All this has to be borne in mind by the translator and the terminologist who assists him when acquiring terms and settling on their equivalents in the other languages. (It is only on the basis of a comprehensive collection and documentation of terminoloqical/lexicographical information in a terminology data bank that work can proceed on the systematic standard-

ization of unambiguous and binding concepts and equivalent terms in particular subject areas.)

1.2 Multilingual terminology

At present, terminology is compiled in the most important European languages (including Russian), although this does not mean that every entry can and must have direct equivalents in all these languages. This underlines the need to adhere to a strict terminological procedure since it is only by clarifying the meaning of specialised expressions, i.e. the basic concepts underlying them, that it is possible to compare and select equivalent expressions in the various languages.

It is a fundamental feature of multilingual terminology acquisition that no distinction is made between the source language and the target language while storing items, - particularly in respect of scientific and technical concepts, which tend to become accepted in the various language areas at roughly the same time. It is therefore essential that the terms used in the various languages should refer to the same concept and be direct terminological equivalents.

In cases where, in a particular language, the given concept is not in current use, a kind of paraphrase may be used instead of a coined expression. When a paraphrase is used, the principle of reversibility of equivalent terminology (source and target language) is abandoned. The same applies of course in all cases where a particular concept in one language has no precise equivalent in another language, owing for instance to the fact that different geographical, historical and social factors have led to similar concepts being defined differently or that completely different concepts have to be indicated which are applicable and - significant only in a given linguistic (i.e. cultural) context (e.g. law, economics, historically influenced tech-

nologies etc.)

Depending on how the stored data will be used as an aid to translation - it will be necessary in each case to decide whether a terminological equivalent can be provided (if necessary with supplementary details and restrictions for the different languages), or whether the conceptual differences preclude this and necessitate the formulation and storage of different entries.

3 Multi-word terms

This approach to terminology leads naturally to the acquisition (and 'translation') of expressions consisting of more than one word. Multi-word expressions are widespread in technical language (although the ease with which German creates compound words makes this less obvious in that language than in others), since numerous concepts can only be described unambiguously by such word groups (e.g. adjective and noun in German, English compounds written as separate words etc.). The real linguistic units in a collection of terminological data are therefore not single words but often highly complex syntagms. The equivalents of such expressions in other languages cannot be obtained simply from 1 : 1 or word-for-word translations (e.g. E: "thumb screw", F: "vis à serrage à main" [3]). The same goes for terminology used in economic documents such as balance sheets and annual reports of international firms. These texts often contain complex syntagms which nevertheless have a fixed meaning in a given context e.g. "Dampf- und- Heisswasserverbrauch" as an item in a German accounting schedule or "Wagnisse wegen Schwankungen der Fremdwährungskurse" [4] in accounting, these represent terminological units to be rendered as similar units in other languages. Another important special case is that of the often lengthy titles of laws or agreements and of national or internatio-

nal institutions and organizations ('Conference on Security and Cooperation in Europe'). To ensure that these complicated titles (i.e. renderings of individual concepts) are used accurately and consistently in translations of official documents or other important texts, it is advisable to commit them to a translation terminology data bank together with their established equivalents in other languages. As is the case with complicated technical terms, these names and titles should also be retrievable by reference to simple parts of the title or synonyms and, in certain cases, unofficial contracted forms ('Helsinki Conference').

However, although the decisive criterion for the creation of terminological entries is the question of concept formulation, i.e. the definition of a concept and its place in a conceptual system and the related putting into terminological form of the linguistic expression [5]), two other points should be borne in mind - particularly in respect of terminological data banks to be used as a translation aid. In the first place, specialized expressions should be compiled and made available in the form in which they actually occur (in the day-to-day texts for translation), regardless of the degree of terminological clarification. Secondly, account must be taken of the linguistic form of the terms allocated to the stored concepts in order to facilitate their retrieval by the translator. One aspect of this question has already been dealt with in connection with the non-reversibility of paraphrases. In addition, complex (including compound) concepts - and the corresponding formulations - should be broken down into their component parts to the extent that these parts also have precise equivalents in all the other languages.

The possible linguistic forms of the terms are dependent largely on the language concerned (compound words in German, prepositional constructions in the Romance languages, compounds and genitive constructions in English.) To facilitate

retrieval, term entries should be as short as possible (consisting of a single word or a small number of words), and from the linguistic point of view limited to more or less fixed syntagms, although longer and variable syntagms cannot always be avoided (indeed, as mentioned above, paraphrases may be necessary.) These entries, however, require special treatment to ensure retrieval; this will be discussed in more detail later. From the linguistic (and processing) point of view there is a grey area here between term and phrase.

1.4 Phraseology

There can be no doubt that in technical language, and in particular in the translation of technical language, the important features are not only individual terms,consisting generally of nouns or compound noun expressions, but also complex phraseological expressions [6]). Verbal constructions, (i.e. combining nouns and verbs, which frequently only have a technical meaning in this combination, are of major importance. These constructions can be simple verb-object constructions ('to load a register', 'to link a programme', 'to set a switch') or whole sentences with finite verbs ('a program generates an error report').

These phraseological units can be stored together with the foreign-language equivalents in the same way as simple term entries and can be arranged and retrieved using the same principles and criteria.

As has already been mentioned, strict terminological theory requires the concept forming the basis of a term to be supplied in the form of a definition. It is impossible, however, to provide unambiguous and universally valid definitions for every single concept appearing in the texts to be translated (including specialized concepts). Instead, therefore, a precise and complete definition which may only be of

interest to the specialist, it is possible to provide a perfectly adequate aid to the understanding and translation of a particular term by a general explanatory note or even a fragment of text using the term in a typical, specific context.

The use of contexts is justified by the fact that many complicated (compound) technical expressions frequently appear in a modified or abbreviated form, either because the complete form has already been used previously in the text and is not repeated for reasons of style or economy of space, or because the context itself provides sufficient information to enable the abbreviated expression to be identified unambiguously (e.g. 'loop' instead of 'programme loop'). Moreover, terms - including technical terms - can often not be translated without a change in in the sequence or form of the expression: e.g. noun expressions become verbal expressions in the target language [7]. To overcome these translation problems, longer phraseological units - as opposed to single terminological equivalents - must be stored together with their equivalents in other languages.

As a general rule, therefore, in dealing with context entries it must be made clear whether the fragment of text is stored only as information supplementary to a given term, or whether the fragment itself, together with its equivalent translations in other languages is the actual item of information, retrievable via the terms ('keywords') it contains.

There are, for instance, texts in which the wording of whole passages is as important as the translation of individual technical terms. In the case of international laws and agreements, whole sentences or parts of sentences which are recorded in a fixed form and sequence together with their equivalents in other languages represent the

information units and parameters in a terminological system in the wide sense. [8]

2. SUPPLEMENTARY INFORMATION

Every terminological entry in the broad sense, i.e. single terms as well as phraseological units, must be supplemented by a series of data to ensure optimum identification by the user. This supplementary information can relate either to an entire entry or to information given in one particular language.

Mention has already been made of conceptual definitions. Of course, these cannot be and do not need to be complete - i.e. fully elucidated or even binding definitions - in a terminology data bank designed as an aid to translation. There is generally insufficient capacity or time, or indeed need for this. Frequently it is sufficient simply to give a brief explanation or comment, a note on usage, purpose or construction, examples or a short indication of the difference between the given term and similar terms and concepts.

2.1 Subject areas and system related factors

One of the most important items of data in any terminological entry to be stored in a comprehensive data bank is a subject reference specifying the area to which the concept belongs. Each concept can be accompanied by one or more of these references. Normally they are coded on the basis of a subject key index. (In the TEAM system, this index is based essentially on a polyhierarchical classification and can be broken down to a maximum of (at present) four decimal places.

The subject reference can also be given in an uncoded form, for example, in a mnemonic, abbreviated form. It is possible to convert those uncoded references into the coded

references by means of a computer program, and vice-versa, i.e. to produce 'labels' from the codes in any required form and language and to add them to the entry at the appropriate point. The translator can use the subject reference (when, for example, no definition is given in the full text) to decide on the suitability of the given term. A general advantage is that these references can be used to obtain a subject glossary from the body of information stored in the terminology data bank.

It is often necessary to provide a further note limiting the validity of an entry in addition to details of subject area. For instance, many technical or other terms are system-related, or relate to a particular item of equipment are only meaningful to certain users, or have a specific meaning only in a particular context (a firm's annual report, an international agreement etc.). All these and other systems data should follow a uniform pattern and be given in a form which every user can understand. They can then of course act as criteria for computerized selection.

Moreover, individual items of information retracting or explaining the use of particular terms can, if required, be added in brackets to any part of an entry - e.g. 'load (a register)'. This bracketed information accompanying a term is not, however, a criterion for the normal selection procedures, and is ignored by the sort or mechanized dictionary search routines. They are, however, communicated to the user when the answer to his question is either printed out or shown on a screen. If a particular specialized term, for example, has different regional forms, this is likewise either shown or printed out - e.g. '(US)' for American usage, '(GB)' for British English, etc.

2.2 Reliability and source indicators

A reliability indicator can be given for the whole diction-

ary entry (e.g. 'approved', 'provisional working concept'), and values can be attached to individual language entries by marking and coding source data. It is a basic rule that every term to be stored in the data bank should be complemented by details identifying the source (e.g. a standard periodical or other text) or the person who supplied the term. More detailed information on the sources, which are usually recorded in abbreviated form, are contained in a source file kept separately from the stored terminology.

Information on reliability can either be recorded in coded form together with the source data or in clear - e.g. in brackets after the relevant term (obsolete)'

Like the other bracketed information referred to above, however, the reliability indication cannot be used in this form in normal interrogation as a criterion for computerized sorting. For this to be possible, the relevant data must be recorded as independent and separate items of information in a special category in the terminological input. However, if need be, a character recognition routine ("concordance program") can be used to retrieve entries containing specified information in brackets.

2.3 Grammatical indicators

As in conventional specialized dictionaries, grammatical information is given to only a limited extent with the actual terms. By and large, the only details given with individual words (i.e. single-word) are the part of speech and, in the case of nouns, the gender. These data help the user to differentiate between homographs (Band m./n', 'close v./adj'). Each is of course recorded as a separate entry together with its relevant supplementary information. The grammatical information on individual words is stored via a separate category and may therefore be used as a

as a possible sort criterion. On retrieval this information is normally given immediately after the term but before any following information in brackets. In certain lexicographical output routines, grammatical indicators relating to the first word in a multi-word term may be entered immediately behind this first word. This applies in particular to the Romance languages, in which the noun usually comes before an adjectival or prepositional attribute: ('terminal m à ecran').

2.4 Synonyms

Together with the term, any number of synonyms can be stored as part of any entry in any language. A program can generate separate entries or appropriate cross - references for these synonyms in the source language. It should be noted, however, that these are synonyms in the strict terminological sense of the word, i.e. equivalents for a certain concept, in one particular language. The same definition and any systematic reference indicators must apply in all cases. The synonyms must also of course all be equivalent to the terms given in all the other languages in the entry. If this is not the case - where for example one of the terms has a slightly different nuance when used in a specific context - a new entry must be made. On the other hand, it is of course possible to store quasi-synonyms in a separate category in the entry as supplementary information. In the same way, other characteristics such as antinomy, hierarchy etc. can be included in entries to provide a basis for computer - generated references.

2.5 Data representation

In accordance with conventional dictionary practice all the terminological entries are stored in their basic forms: nouns generally in the nominative singular, verbs

in the infinitive etc. This principle is also applied to complex terminological units like multi-word terms and compound names. These and, in particular, the phraseological units mentioned above are also stored in their natural word order (e.g. - to take a very simple example - with the German attributive adjective preceding the noun: 'symbolische Adresse'.). Moreover, all entries are made using the correct orthography, i.e. using upper and lower case, as well as the accents and diacritics used in the European languages.

Languages which do not use the Latin alphabet are transliterated in such a way as to permit retransliteration into their own alphabets with the use of suitable output equipment (e.g. DIGISET - photocomposition units with the appropriate characters - e.g. Russian transliterated from the Cyrillic alphabet in accordance with ISO R 9). The terminological entries are subject to no (at least, no discernible) length limitation.

The same applies in principle to the component parts of the terminological entries, i.e. the information categories. All these categories are recorded, classified and stored on the basis of an extendable scheme enabling them to be automatically manipulated (sorted, verified, printed).

3. DATA RETRIEVAL

In general, there are two or three ways of retrieving stored terminological information:
1.The terminology data base is printed out in whole or in parts: either on high-speed printer lists, or using the typographically more sophisticated photocomposition system, or on microfilm or microfiches. This allows the required entries to be found by reference to the alphabetic sequence of terms or keywords. Generally, such lists will be produced by a particular sort routine, the existence and

content on the previously-mentioned information categories being the usual sort criterion.

2. An alternative means of access to the stored information is the retrieval of individual dictionary entries via a data terminal, e.g. a data display terminal. In this case, the primary access criterion will always be the term itself (in the source language), possibly with supplementary data (e.g. to exclude irrelevant subject areas). The idea is therefore not that the user should work through what may be a fairly comprehensive list, but rather that he should have direct access to a particular term.

3. A third means of interrogation which could possibly be placed somewhere between the other two mentioned above, is batch interrogation. By this method, a (limited) number of specific questions, derived from a particular text to be translated, are answered by the computer. The list of the required terms together with their equivalents in the foreign language(s) can be set out either alphabetically or in the input sequence, i.e. as the terms appear in the given text.

A similar result can be achieved by automatic interrogation (by way of contrast to the above-mentioned methods of interrogation by the user himself) i.e. automatic scanning of machine-readable texts.

All the above-mentioned methods of interrogation are based on the same data organization procedures and processing routines, and a number of basic requirements applying to all these methods and relating to the acquisition and storage of terminological entries have consequently to be observed.

3.1 Reversibility

One problem connected with the distinction between the source and target language has already been touched upon. If a concept is only one language and can only be expressed in other languages by means of paraphrasing (or an inexact translation), the given terminological 'equivalent' is not reversible as the paraphrases in the other languages can of course never be the subject of a search and can, moreover, never appear on a alphabetical list as possible search words in the source language. Paraphrase and similar auxiliary concepts should be specially identified in the acquisition and lexicographical processing of the material. Likewise, there is of course no point in reversing technical expressions which were coined in a particular country for a new development and which have as yet no generally accepted and current equivalents in foreign languages. In this context it is worth stressing that the general problem of the reversibility of dictionaries (particularly standard dictionaries) in the TEAM system is already limited by the more or less strictly applied terminological principles.

Vague word equivalents which are generally justified only in a specific context are not included in this electronic dictionary. Multi-lingual word or term equivalents must always be accompanied by sufficient additional information to allow them to be identified clearly, both semantically and syntactically, before they can be accepted as terminological equivalents. All ambiguities, including homonymy, polysemy and other semantic differences in technical words and expressions are analytically resolved and form the basis of separate lexicographical entries. (If required for more sophisticated lexicographical output forms, e.g. for the composition of dictionaries derived from the data base, these separate entries can be consolidated again in a single unit as a dictionary entry.)

3.2 Synonyms, inversions, abbreviations

Mention has already been made of the question of synonyms. If these are in fact exact synonyms in the sense defined they should not only appear as supplementary information to the required expression (in the source language or only in the target language): the stored translation equivalents should of course be accessible via every available synonym in the source language. For this purpose separate dictionary entries or merely corresponding references index entries are generated by program for the synonyms in the source language, and then appear at the correct alphbetical location. (In other words, they do not need to be written separately during data acquisition).

Another problem referred to earlier is that of recording multi-word expressions (word-group lexemes) in their natural word-order.

The search for such word-groups (fixed syntagms) is often supposed to ignore the first word, concentrating exclusively on a subsequent important part or section. This applies particularly to verbal or prepositional expressions ("ein Programm laden", "to link a module", "mit fester Länge") and other phraseological units. In cases such as these, the expression should be given an appropriate indicator at the acquisition stage, so that later (in the alphabetical lists or indexes) it does not appear in this form, i.e. ranged according to the first word. The original natural word order should, however, be retained when the information is being provided in the target language. Those parts of the word-group which can be used as search items must be indicated.

On the basis of this indication, a computer program can generate inversions (inverted entries), whereby the indicated part of the term (as required in the source

language) becomes the first word of the expression, with the rest of the expression following a comma. (In such cases, the part of the expression which has been moved to the front is replaced by a swing dash at its original place in the non-inverted word sequence: "..ein *Programm binden" : "Programm, ein ~binden"). These inversions can of course also be carried out in cases where the full expression is also to be found by reference to its first word, with the result that the entry can be found at different places in the dictionary by reference to more than one search word. Of course, complete inversion entries can be replaced by a suitable index reference (Länge, s. mit fester Länge").

Abbreviations are dealt with in a similar way. Whichever of the two forms is less common - abbreviation or full form - is placed in brackets behind the other. A programme can then produce the inverted entry or simply an appropriate reference.

3.3 Keywords

In the previous section, the general problem of the introduction of keywords in the TEAM system was touched upon. By reference to these keywords, complex terminological and phraseological units can be retrieved, either in dictionary articles composed, listed and printed by the computer or by interactive interrogation via a terminal. These search words or keywords within the syntagm are indicated during preparation and input of material - e.g. for the above-mentioned inversions - by simply placing an asterisk before the word concerned: ("mit fester * Länge").

It has been found to be more practical and economical to have the material prepared non-automatically than to produce automatically all the possible permutations (even

allowing for the elimination of certain insignificant words) for suitable dictionary or index entries.

The search words for complex terminological entries should appear in their basic grammatical form - the dictionary reference form - even though in the given syntagm they may appear in inflected form (e.g. "fest" as the keyword for "mit fester Länge"). The indication given to a keyword at the input stage should also respect this rule (by an additional indication of the inflectional ending), so that the inverted expression then reads: "fest, mit ~ er Länge" This marking is, however, only possible in the case of simple inflected forms where an ending is simply added to the basic form. Changes to the root of a word (e.g.vowel changes in German plurals) cannot be indicated in such a simple way in the original expression and cannot be handled automatically. (In such cases, the lemma is not replaced by a tilde in a multi-word expression even in conventional dictionaries, but is simply repeated in the modified form.) Instead, the keyword must be inputted and stored in the entry separately from the word-group. On the basis of this supplementary information, a program can then generate a suitable entry or reference ("duty, s.rate of duties").

3.4 Phraseology index

Phraseological units (mentioned earlier) can be handled in a very similar way, no distinction being drawn between terminological entries and those phraseological entries which are intended to be arranged alphabetically in a common list (glossary or dictionary) - either under their first word or under other keywords in the phrase can be indicated with a view to inversion and additional keywords can be used for reference purposes.

Phraseological entries, will however, be handled in a

different way if they contain lengthy fragments of text (with equivalent translations). There is little point in arranging these texts alphabetically. A more sensible solution is to bring them together in a file - in some kind of consecutively numbered form - and to add one or more keyword-based indexes corresponding to the number of languages contained in the entry. Sub-groups of the whole phrase can be inputted as index entries in addition to simple keywords, and these syntagms can be used as a means of retrieving file entries, regardless of whether or not they appear in identical form in the phrase. ('Dampfverbrauch' in addition to 'Dampf- und Heisswasser-verbrauch','Verpflichtungen übernehmen' in addition to 'die Verpflichtungen, die dieser Staat in Hinblick auf die Aufrechterhaltung des Friedens übernommen hat' [8]). These multi-part search or index expressions can of course themselves be indicated in such a way that they can be classified and retrieved in inverted form ('Wortlänge, feste' as well as 'feste Wortlänge' as a form of 'Rechner fester Wortlänger' or 'übernehmen, Verpflichtungen' as a form of the above-mentioned example).

4. INTERACTIVE INTERROGATION

In the following section, two methods of interrogation will be described in more detail, both of them requiring direct access to the stored terminological entries - in contrast to the sequential searching and processing of whole collections or parts of collections. The first method is interactive interrogation carried out by one (or more) human users [9]. The second is automatic interrogation, i.e. the determination and 'translation' of specialized expressions contained in a machine-readable text and requiring no human intervention [10].

By the first method, the translator (or other interested party) can carry out a conversational interrogation of the

terminology data base via a terminal in the same way as he would consult a dictionary. The terminals are teleprinters or visual display units with an entry keyboard linked to the computer. The keyboard can be used to feed questions into the computer. The answer appears immediately - as a rule a few fractions of a second later - on the same screen (or on the teleprinter).

4.1 Questions

The simplest form of a question is a single word to which the answer will be the corresponding dictionary entry. The question can, of course, consist of a compound expression, i.e. several individual words. Such a multi-word term is searched for as a whole, i.e. taking into account the given sequences of the individual parts of the term (words).

Of course, the interrogation program stored in the computer must be informed of the language in which questions are to be asked, i.e. in which section of the dictionary the answers are to be sought. As has already been mentioned, the stored dictionary entries are multilingual, any one of the available languages being in principle a possible source language (mention has already been made of limitations to this principle, even in respect of the widely technical languages).

If the original language is French, the question 'bande perforée' will receive an answer along the following lines:

D: Lochstreifen m.
E: punched paper tape, punched tape, paper tape
F: bande perforée

(The languages are stored in the dictionary in the above sequence. The relevant supplementary information has not been included in the examples). As long as the search language or source language remains the same, the language

instructions do not have to be repeated. A change in either can, however, be made any time by entering a new command.

As well as the source language, the computer can also be fed with details of the desired target languages. The result is that the answer given is then not the complete dictionary entry, which may comprise a large number of languages. If, for example, German were to be requested as the only target language for the (English) question 'paper tape' the answer would be:

E: punched paper tape, punched tape, paper tape
D: Lochstreifen m.

(In this case, the English language section of the entry is brought forward on the screen, the French entry being suppressed.)

4.2 Answer entries

In addition to the desired answer, i.e. the relevant language sections of the dictionary entries (source and target language and details of part of speech and source together with other supplementary information), certain other data appear on the screen. These include the previously-mentioned subject indicators.

In a new version of the interrogation program, the user himself will be able to determine the selection and allocation not only of the language sections, but also of the desired partial and supplementary items of information (entry categories). He will then also be able to decide whether or not the source data or any definitions, inter alia, should be displayed on the terminal.

Another important item is the consecutive number given to every stored entry and used in the administration

(including correction) of the data box. This number can be used to address and call up the entries onto the screen, a facility which is of more interest to terminologists and lexicographers responsible for the data box than to the translator.

It is, however, not always possible to retrieve a precise answer to every question. If the required expression has not yet been stored, no answer will be available ('Fehlt') It can also happen that several dictionary entries are available as answers to a single question. In this case, the first entry is offered, and the user is asked whether he wants a 'further answer'. These further answers can then be called up in sequence.

A comprehensive terminology data base (with several hundred thousand entries) will always be sub/divided on the basis of origin (i.e. subject area, source and system-orientation of the terminology) and processing and updating into numerous different sub-pools, which are bound to over-lap in places. For this reason alone there will always be cases where a number of answers are given, quite apart from the problem of homographs and polysemes. The interrogation program is being expanded to combine all these answer entries into a larger unit (allowing for the above-mentioned category selection) with the aim of giving the user a better overall view of the entries available as answers to any particular question.

By suppressing duplicate entries (i.e. genuinely redundant repetitions), the compressed answer text can be accommod-ated, for example, on a screen with the result that further answers do not have to be called up individually, thereby displacing the previous text.

As the difference between answer entries to a particular question sometimes lies only in the existence or absence

of equivalents in the various target languages, it makes sense for only those entries which actually contain the required target language to be given as answers. The requirement is therefore not only that all the irrelevant language sections in the answer be suppressed but also that the entry be ignored entirely if it contains no record in the required target language.

The system user can obtain a written record of the dictionary entries which he finds relevant simply by having the entry shown on the screen printed out on a teleprinter linked to the system (the two pieces of equipment need not be in the same room). He can either have written copies made of individual entries or of all the answers received to a particular question. For the purposes of servicing and processing the stored terminology data base, the individual entries can be printed out in the form in which they were originally inputted and stored (i.e. including all format data and codings) so that they can be used directly for correction purposes (the entries can of course also be displayed in the same form on the screen.)

4.3 'Browsing'

In addition to being able to call up several consecutive answers to a single question, the user can browse, so to speak, through the dictionary. He can have the next dictionary entry, i.e. the next one in the alphabetical sorting, displayed on the screen. If there is no room on the screen for the information contained in a dictionary entry, the user can, by giving an appropriate command, have the remainder of the information displayed on the screen. The entry last shown on the screen (or the first part if the entry is too large to be accommodated on the screen at one time) can be recalled to the screen at any time, e.g. after a series of further unsuccessful inquiries.

Apart from the above-mentioned forms of direct consultation/and browsing, there is one further rapid method of consulting the computer dictionary. Like many printed multilingual dictionaries or glossaries the dictionary stored in the TEAM system computer is divided into a main section containing the multilingual entries themselves, and a series of indexes, containing the search concepts (i.e. terms), set out by language in alphabetical order, and with a reference to or the address of the main entry. The user can browse in these indexes too, in an attempt to find his way gradually to the required information. This is done by asking a so-called "keyword" question the expected answer being not an entry with the required translations but rather a list of all stored expressions beginning with the given keyword (which may, incidentally, be made up of any character string). The computer recognises this kind of question from the fact that it is preceded by an asterisk.

The question '*real time', for example, would receive the following list as an answer:

1. real time data (...)
2. real time simulation (...)
3. real time working (...)

Depending on the type of equipment, a varying number of such index entries can be displayed on the screen. Should there be a large number of entries, the next section can be called up, and the first section of a lengthy list can also be recalled.

An appropriate list of index entries, can incidentally, also be requested whenever a question receives the answer 'Fehlt' (unavailable). In this way, the user can find out whether compounds containing the required word are available for consultation together with the equivalent expressions in the target language.

Each individual index entry is followed by the number of dictionary entries to which it refers. If, in addition, a particular target language is specified, details are also given of the number of entries containing terms in that target language. The complete dictionary entries can be obtained by feeding in the numbers (1, 2 etc.) preceding the index entry on the screen. (If there is more than one relevant dictionary entry, the system will then of course ask whether the subsequent entries are also required.)

4.4 Index

Inside the data processing equipment, the search for the terminological data always starts with the above-mentioned indexes for the various source languages (even if the user has not asked a keyword question). The full entry is only called on in a second stage, if required. As the structure of these indexes is of decisive importance for the working and performance of the interrogation system, a short description of them will be given. The index in each language is arranged in alphabetical order according to the rules applying to that language; in Spanish and Russian, for example, the alphabetical order is different from that used in German or English.

The index entries differ from the normal dictionary entries in terms of presentation and orthography, owing to a certain degree of condensing and standardizing of expressions. For example, all gaps or spaces between words are dispensed with, the same procedure being followed internally for all questions put by the user. By this means, questions can be answered successfully even in cases where, for example, multi-word expressions in an English question are written in a different form to that used in the dictionary (with or without hyphens, words separated or run together).

At present, the individual index lines have a maximum length of only 20 characters. This means that the 21st and subsequent letters of any term are not included in the index, the result being that any question consisting of more than 20 (condensed) characters will produce all the entries in the dictionary which begin with the same 20 letters. (If at some future time, it is found that this restriction generates too much noise in the answers, the maximum length of index entries - and hence also of dictionary comparisons - can be increased as required.

For questions put in French, all accents and cedillas are suppressed in the index. In the case of German, the ' ß ' is treated as 'ss' and umlauts are reduced to their basic letters (without diaresis). The difference between upper and lower case is eliminated as far as search and comparison are concerned, i.e. in the compilation of the index and the handling of questions. These standardization procedures lead to only a small number of ambiguities in specialized vocabularies and thus to only a small number of unwarranted answers. The dictionary entries themselves(as opposed to the indices) are, of course, shown in the correct orthography on the screen. As the terminals cannot show Cyrillic characters, the Russian terms are given in the transliterated form recommended by the ISO, i.e. using Latin letters with diacritic signs.

A language index, includes, not only the main terms from the dictionary entries, but also all the synonyms stored in the entries. The index also includes abbreviations as well as the original full forms - where both are contained in the dictionary entry - and hence both abbreviation and full form can be retrieved. In addition, the previously mentioned inversions of multi-word terms are also included in the index. (These inversions are always carried out automatically in cases where at least one of

the parts of a compound expression, although it is not at the beginning, is regarded as a possible search word and is given an appropriate indicator at the input stage. Hence the two-word expression 'bistabile *Kippschaltung' is also recorded in the inverted form 'Kippschaltung, bistabile'.)

This index can of course also be used to find phraseological entries provided that the appropriate keywords have been included in the index. These keywords (or whole sub-groups) can, as explained previously, either be provided with an asterisk inside phrases at the input stage to allow indexes to be compiled, or they can be recorded as supplementary information to a phraseology entry.

It is also possible either to set up a separate data pool for this kind of phraseological data, capable of being interrogated by the same program, or to mix phraseology and terminology in the same pool. In the latter case, it will be possible to interrogate only one or other type of entry using appropriate selection commands, or to interrogate both together subject to priority, i.e. a certain terminal output sequence.

4.5 Search strategy

The search methods discussed so far have all been based on alphabetical (but not sequential) access to information, i.e. the search criterion is always a word, a word-group or a part of a word; in other words, a sequence of letters. This method is therefore similar to that used by translators when working with conventional dictionaries. The keyword question (with the possibility of abbreviating the end of a search word at will) enables a successful search to be carried out despite the difficulties presented by certain grammatical and morphological

aspects (inflection and word derivations) and by attributive and other word-groups.

A selection on the basis of subject area (or similar criteria) can only be carried out when several stored entries have been found using the above method. A general selection and search facility using other descriptors and their logical combinations (as in other information systems, particularly in document retrieval) is therefore not provided, nor do we believe it to be necessary for the purpose of direct dictionary or terminology interrogation. (Of course, appropriate batch routines are available, e.g. for the production of terminology lists. Using a time-sharing system, these can be called up via terminals.)

On the other hand, it may be advisable to give the user of the terminology data bank the opportunity to make inquiries and further searches in the stored terminology collection by means of content-based (i.e. conceptual) relations. So far, only the synonymy relations have been fixed, so that terminology entries can be found via all the recorded synonyms as well as via first or preferred terms. In addition, it is intended that hierarchical relations (such as generic or partitive relations) as well as relations concerned with definition or particular associations should be recorded and evaluated, so that a whole conceptual structure can be given as to the answer to a question.

5. AUTOMATIC INTERROGATION

Finally, a few points on the automatic interrogation of a terminology data bank. By contrast to the interactive method described above and the text-related batch interrogation which was briefly discussed earlier, automatic interrogation does not depend upon the questions being put by a human user who extracts the expressions to be translated from a text and prepares them for the interro-

gation procedure. In this case, a program compares the expressions to be translated in a machine-readable text with the stored terminology (without human intervention) and "translates" them.

Two main difficulties have to be overcome in the processing of technical texts: the individual technical words can occur in the text in inflected form, i.e. in a different form from that in which they have been stored in the machine, and they can form a part of a word-group (a multi-word term), the meaning of the word depending on its function within this word-group.

The first of these problems creates no difficulty in non-automatic interrogation: it is usually a simple mental process to convert inflected word forms to their basic forms for the purpose of a dictionary search. The second problem does, however, also involve the translator in some work, e.g. repeated consultation (at various points) of one or more dictionaries.

Both of these problems are more difficult for the computer to solve, as - unless special steps are taken - it has no facility to compare with the human translator's knowledge of morphology and syntax, and his more extensive semantic skills and technical expertise.

Automatic interrogation is nevertheless based to a considerable extent on the conventional dictionary consultation of the translator, viz:

1. Words which, in the given context, do not rank as insignificant, are sought in the alphabetical dictionary.
2. In the search, i.e. the comparison of the question word with the dictionary entries, inflexional endings are eliminated.

3. The question words must be checked to see whether they form part, in the given context, or a word-group recorded in that form.

Before the actual search is made in the dictionary, it is advisable to eliminate all insignificant words. Words which should rank as insignificant for the particular search intended, are contained in a list based on frequency surveys of appropriate technical texts. This list therefore represents a first, relatively simple, addition to the information available to the computer for analysis purpose.

Making a search in an automatic technical dictionary is a matter of comparing the question word with the stored entries, as in the case with conventional dictionaries. A computer can, however, make this comparison only by strict reference to typographical characters. Even very minor differences can therefore result in the interrogation yielding a negative result. The question of orthographical variants and the available characters and the question of the establishment of a standard orthography were dealt with in the section on interactive interrogation.

5.1 Reduction to basic form

Before a dictionary search can be made, the question words must first be converted to the basic form used in that dictionary. There is, however, no simple logical method of deciding (without comprehensive grammatical analysis) whether a particular question word is in an inflected form, and how this should, if necessary, be reduced to the appropriate basic form.

Since the system of automatic interrogation should work as far as possible with only the information contained

in the data base, it is advisable to use an artificial reference form obtained by eliminating all possible inflections from the question words. This is not therefore a matter of reducing a word to its basic root in the strict sense, but rather of simply eliminating the letters at the end of a word if they could possibly represent an inflexional ending (regardless of whether the ending actually is an inflexion in the given case.)

The reduced word can then be found directly in the dictionary if the ending is a genuine one which is simply added to the end of the word in its dictionary form (e.g. Zugriffszeit/en), or if the succession of letters eliminated as being a possible ending is indeed part of the basic form, but was likewise eliminated in the dictionary entry, i.e. in the appropriate search word of the dictionary entry (Prüfzeich/en).

This method of treatment requires that the list of possible endings be also extended to cover sequences of letters arising from possible combinations of endings. In practice, therefore, the items to be eliminated are not only genuine endings, but also sequences of letters preceding these endings which are identical to a possible ending and which are for this reason suppressed in the reference form (Prüfzeich/en-s). If parts of the word root are eliminated in the reduction process, this may result in reference forms that are not plausible to the human user. (e.g. "Zin/s" as opposed to "Gestein/s"); this is, however, of no significance as far as automatic processing is concerned. This system is primarily designed, of course, to deal with noun expressions which form the bulk, of technical terminology. (Variable verb compounds or vowel changes in strong verbs are not recognizable by this method.)

A list of actual endings will therefore form *a* set of

rules to act as supplementary analysis information for the computer and with the help of which question words can be reduced to the machine-oriented reference form. This set of rules can be easily modified and adapted to special lexicographical needs. Exceptions or limitations to the rules can be incorporated into the list, and should be listed before the relevant rule. In certain cases it may be advisable to incorporate additional data on substitute endings in the set of rules, i.e. series of letters which can in certain circumstances replace another ending. For instance, in English the ending 'ies' could be replaced by 'y', unless it proves more practical to replace all 'y' 's at the ends of words (apart from cases where the y follows a, e, or o) by 'i' and to eliminate the ending 'es'.

5.2 Recognition of multi-word terms

As has already been stated, it is not sufficient simply to look up individual words in the dictionary and to give their translations. Multi-word terms, i.e. expressions consisting of more than one word, should be treated as lexical units. There are, however, certain difficulties involved in trying to retrieve them as a whole in the dictionary, not only because individual parts of the multi-word term can appear in the text in inflected form (... eines bistabil/en Multivibrator/s), but also because it has to be decided before the question is put which word group in the text (of the many possible combinations, i.e. which is the first word and how many successive words does the group contain?) should be regarded as the expression to be look for in the dictionary. This problem should be solved by the use of a procedure similar to that used to look up words in a conventional dictionary. The search is based on the individual words which are the first words of a multi-word term and which are stored as such - with a reference to the complete term - in the

dictionary (i.e. the terminological data base). The terminology collection required for the 'translation' of the terms is therefore used at the same time to formulate the questions. Apart from single-word terms it has been found useful to limit the search initially to fixed syntagms represented by the multi-word noun terms which retain their natural sequence in the text. Once it has been established which is to be regarded as the first word of *a* group, the text must be scanned to find the longest sequence of words beginning with this word which is likewise stored in the dictionary (with the same initial word). If the above-mentioned automatic interrogation procedure is followed, it is possible to work through a given text word for word. If one or more answer entries are found in the dictionary for a single question word, a distinction should be drawn between the following cases (disregarding the possibility of ambiguities caused by the reduction of endings):

1. Single words, used as technical terms in the given text.
2. The question word as the first word in one or more multi-word terms in the dictionary. Whether, and if so which, multi-word expression is present in the text must be checked by examining the subsequent words; the longest sequence of words appearing as a whole in the dictionary - starting from the longest stored sequence - is taken as the appropriate specialized expression (e.g. 'object' for 'object code').
3. A combination of cases 1. and 2. In this case single words are only given as answers if they do not form part of a larger word group in the text. If a direct concordance is recognized for an individual word, the initial words with their relevant entries do not need to be given as possible alternative answers

5.3 Setting up the terminology dictionary.

It is clear from what has been said so far that the interrogation procedure must be based on a suitably organized dictionary. The terminological data base used as the specialized dictionary consists of multilingual entries (including synonyms) with - as was explained earlier - any of the languages being in principle a possible source, (i.e. sort) language. The result is, as a rule, that there must be several search concepts for each terminology entry. There are therefore in practice two stages involved in gaining access to an entry: i.e. the automatic dictionary is subdivided into a main section and an alphabetical index to this main section (as in the case with the interactive system). The index, on which the first interrogation stage, i.e. the search itself, is based, is in turn sub-divided into the individual language sections, which refer exclusively to the entries in the main section containing the complete dictionary entries.

The index section of a dictionary intended for automatic interrogation contains the search words in their standardized and reduced reference forms, as described in a previous chapter. This index is therefore produced by a system of automatic processing similar to that applied to the text to be analysed, and can therefore be built up automatically to form a data base (i.e. by means of a program).

Search words are, in their most basic form, all single-word terms reduced to their appropriate basic forms. In the case of multi-word terms (but not simple paraphrases), the complete terms as well as the initial words must be included in the dictionary to enable the appropriate word group in the text to be identified more readily. Every part of a multi-word term is subjected to the same reduction process as individual words to produce the dictionary

reference expression for the word group. In addition, it is probably advisable to suppress certain auxiliary words in these terms to produce a suitable search expression (Zyklusseit des Speichers, clearing of a call, terminal à écran) .

5.4 Translation aids

One of the subjects of automatic interrogation may be the production of a glossary based on a lengthy machine-readable text. Such a glossary would list the specialized expressions found in the text in alphabetical order, together with their foreign language equivalents, or possibly a list of terms in the order where it is not intended to subject the text to be translated to full automatic processing in this way, but rather to leave the translator himself some means of intervention, this method can still be useful as an extension to existing interrogation systems. This system can, for example, relieve a qualified translator of much preparatory work since the reduction of question words to their basic dictionary form prior to interrogation becomes superfluous and a fragment of text underlined by the translator can be fed into the computer as a question without further preparation.

It goes without saying that a particular method of interrogation cannot by itself fulfil all the various needs and requirements of the translation of technical texts. The various methods must be used together and in conjunction. A comprehensive collection of basic terminology can, for example, be transferred to microfiches on the basis of selected (e.g. subject field-oriented) lists and updated at regular intervals.

For larger scale translation projects, text related interrogation or automatic interrogation can be used

and supplemented by the use of the direct interactive method. It is also advisable - with the aim of reducing the burden on the translation services - to make these aids available not only to the professional translators but also to as many people as possible in a firm, department or association who come into contact in the course of their work with foreign language literature.

Bibliography

1. Wüster E.: Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik, Bonn 1966

Wüster E.: Die Struktur der sprachlichen Begriffswelt und ihre Darstellung in Wörterbüchern Studium Generale 12/1959, pp. 615 - 627

2. DIN 2330 initial standard. Begriffe and Benennungen. Allgemeine Grundsätze. Nov. 1974

3. Wüster, E.: The Machine Tool. An Interlingual Dictionary of Basic Concepts. 1968

4. Fachwörter des Rechnungswesens (Accounting Dictionary), Edition of 6.76 German-English. Siemens AG

5. Drozd, L and Seibicke, W.: Deutsche Fach- und Wissenschaftssprache (1973) pp. 118 et seq.

6. Warner, A.: Internationale Angleichung fachsprachlicher Wendungen der Elektrotechnik. ETZ Publication, No.4,1966.

7. Jumpelt, R.W.: Die Übersetzung naturwissenschaftlicher und technischer Literatur. 1961., pp 166 et seq and 86 et seq.

8. Concordance of the Treaty establishing the European Commission of the European Communities. (Extracts used for demonstration of the TEAK system at the Brussels Terminology Office of the EC Commission, 1974).

9. Schulz.J. and Göricke.H.: The Dictionary in the Computer. (scheduled to appear in "babel" early 1977)

10. Schulz, J.: Automatische Abfrage einer Terminologie Datenbank. Nachrichten für Dokumentation 27 (1976)pp.62-66.