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

THE USE OF EDP IN TERMINOLOGICAL WORK

G. Beling
Forschungsinstitut für Funk und Mathematik
Abteilung Rechner und Führungssysteme, Meckenheim

This paper begins by explaining how lexicography, terminology, and thesauri are related to one another and what common linguistic problems have to be solved in these fields. On the basis of this explanation the paper presents premises for multilingual terminological work in network systems. The important factor for EDP utilization is the differentiation between acquisition format, data bank format and structure, and interchange format. The importance of data exchange will be demonstrated and the proposed interchange format for terminological data, MATER, explained. Examples will be used to illustrate specific problems of multilingual terminological work and EDP utilization.

I . Terminological work - more than applied linguistics

In information and documentation (I & D) language plays a very special role, whether the task is one of processing verbally written documents or of using linguistic means to describe data not formulated in words. However it is not language which is the object of study here (that can be safely left to the linguists) but rather the use of language to formulate and communicate ideas and opinions. This means that all linguistic research must proceed from pragmatic considerations (in the semiotic sense) of communicative use.

Our field of interest can be even more narrowly defined. Our examination concerns not just any languages but technical languages, i.e. it is not the formulations used by Goethe or Shakespeare that are of interest, but current technical language in its most varied manifestations. For I & D purposes therefore linguistic considerations should usually be on a synchronic basis; a diachronic approach will rarely be required to explain conceptual relations (cf. Section 2).

Disregarding two aspects of linguistic usage which are not as yet regarded as the central problem of I & D, viz the improvement (or standardization) of the composition of texts and the reproduction of texts in various natural languages, our interest is concentrated on the investigation of concept systems and their representation by verbal means This may be interpreted as terminological work (in the broader sense):

> "Terminological work (in the broader sense) is the analysis of concept systems and their mapping onto the terminology of technical language in one or more natural language environments"

In accordance with (1) terminology is thus interpreted as:

> "the systematically represented set of terms which are defined in a technical language at variance to their use in common language, or which are not used in common language".

This definition makes it clear that, unlike research into technical languages which examines not only the vocabularies but also the syntaxes of natural technical languages used for human communication, the aim of terminological work is to determine and specify the artificial language core of a technical language.

Within terminological work thus defined two ways of considering language may be distinguished:

- retrospective approach

  Here the given concept system is described on the basis of contemporary usage of language. This is the procedure in lexicography and the standardization of terminology. Lexicography is not interpreted in this case as "the art of dictionary-making" (2) but rather as "the important task of continuously recording usage which, even in the rase of a finite terminology body, is subject to constant change through blurring of precisely defined meanings and through polysemization resulting from the use of a living language"* (3). Terminology standardization is the prescription of unambiguous terms for widely-used concepts and the allocation of suitable definitions to these terms; its basis is lexicography as described above.

* For instance the use in mathematics of the frequently (synonymous) terms: "order", "partial order", "quasi-order", "full order" etc.

- prospective approach

  Here, complete concept systems arc analysed or built up and mapped onto systems of terms. The aim is to prescribe separate definitions for terms* used as synonyms for various concepts and to suggest new terms to fill the "linguistic gaps" within the concept systems. This may be called terminological work(in the narrower sense)(cf.1). To prevent such a "future-oriented" linguistic treatment from becoming unprofitable "art for art's sake" it must of course go hand in hand with "modern" lexicography in tracing the use and criticism of the proposed terms and concepts and where necessary, in adopting its results to the changing interpretation.

Within the special field of I & D itself, we likewise encounter these two approaches to language: on the one hand in the attempt to achieve a uniform terminology in one's own subject (e.g. 4, 5, 6), and, on the other, in the processing of many documentary languages. On account of the lack of system lexicography to date, terminological work (in the narrower sense) is predominantly forward-looking and endeavours to introduce order into the terminologies "infiltrating" from the various related specialized fields, The creation of documentary languages, on the other hand, combines both approaches to linguistic treatment, for of course allowance must be made for prevailing usage; it is, however, inevitable that contradictory uses of terms, concepts and concept systems will have to be eliminated on the basis of pragmatic experience.

* For instance for the terms "data base" and "data set" (cf 4)

2. What is "multilingualism"?

For all the success that language processing as outlined above may have in terms of better understanding between specialists, one purpose of language is unchanging: it is used for communication between people. But the latter will always have a non-identical conceptual background. Consequently it will never be possible even for a specific point in time to attain an intersubjective unambiguous one-to-one correspondence between concepts and terms. At best one can map the vocabulary of a language surjectively into a concept system.

Thus "multilingualism" can never be avoided. A distinction may be made between the following cases:

- multilingualism within a subject held

  This is reflected by the fact that complete unanimity about concepts and terms does not exist even within the same subject field and the same natural language. Three separate influences on the "language" of a subject field may be observed.

  a) The "harder" the science is, i.e. the better it can be described by means of mathematical models and methods, the more uniform this terminology will be (e.g. in theoretical informatics).

  b) The slower the "rate of scientific advance", i.e. the longer scientific works are read, quoted and converted the more the parallel existence of diachronic and synchronic linguistic consideration will impede the evolution of uniform terminology (e.g. in philosophy).

  c) Different schools of thought in the same subject field will result in various concept systems being represented by the same terms (e.g. in linguistics).

Though in many specialized fields terminologists regard this complex of problems as terminological work (in the narrower sense), there is an almost complete absence of basic research into the observed use of language (lexicographical gap) .

- multilingualism between several subject fields

It can frequently be observed that within one and the same natural language some concepts and terms (though often only terms in the sense of "word shells") of one specialized field penetrate into other fields. This more often applies to the more recently established subjects (cf. I & D itself), but this phenomenon also occurs in long-standing subject fields (e.g. mathematics). If one wants to disassociate oneself from "scientific charlatanry" here (7), then this is a fitting task for the conscientious terminologist who very meticulously and consciously propagates the use of new terms.

- socio-cultural multilingualism

This does not refer to the multiplicity of dialects and sociolects within a standard language, however important and interesting this question may be for linguistics, but to the peculiarities of similar standard languages in different countries although the language is colloquially called "English" (e.g. UK or USA), "French" (Canada, France, Belgium) or "German" (FRG, GDR, Switzerland, Austria). Usually this is regarded as merely a problem of everyday language (e.g. the differences between France: "quatre-vingt-dix", Belgium, Switzerland: "nonante", Germany: "Fahrradschieben", Switzerland: "Velostoßen"). However, this phenomenon is also becoming more widespread in technical languages (e.g. US/USA: "special" vs. "technical" for German: "Fach..."; FRG/GDR: "Socialism, Democracy"). If lexicography were to concern itself with this problem if would help translators.

- international multilingualism

  This is the problem generally interpreted as multilingualism posed by different natural languages. This is the problem confronting the translator; it is usually approached using lexicographical aids. This may suffice for the "old" technical languages and for translating one Indo-European language into another; for other, less technical natural languages which do not yet possess the appropriate technical vocabulary at all (e.g. Arabic, Hindi) the problems involved are quite insoluble without a large amount of prior terminological work (in the narrower sense).

In practice the above problems of "multilingualism" will hardly be encountered in pure form. However, it may be inferred that in terminological work (in technical languaage) different procedures have to be observed depending on which of the above cases is concerned. If for instance a German-English dictionary of linguistics is to be compiled allowance has to be made for the fact that the vocabulary of each of the natural languages concerned is governed by the respective schools of thought, i.e. the dictionary has to be subdivided into a series of overlapping subdictionaries. Due also to the "national" peculiarities of the schools, a considerable amount of terminological work (in the narrower sense) would be required to compile new terms and definitions.

3 The use of EDP

Leaving aside the use of data processing for purposes of linguistic research as such which may be viewed as the main task of computational linguistics (e.g. generative grammars) the use of EDP can be divided into a number of stages:

- data processing as a universal aid

  The work of terminologists entails a whole scries of routine steps that can be performed more quickly and reliably by a programmed EDP system. Furthermore many operations can only be performed with the aid of EDP. This includes data acquisition and input and format-checking, the various cases where it is necessary to check that the data are complete and consistent and the output of data via various media, including high-speed printer, COM and phototype-setting, and all the possibilities for compiling indexes.

  The final product of this use of EDP, which can considerably reduce the burden on the terminologist and the lexicographer, may be dictionaries of all kinds and thesauri.

- terminological data banks

  If data sets requiring frequent change are to be quickly and selectively searched in accordance with various criteria, it may be practical to set up terminological data banks. In the past these have been developed and used primarily as a aid to translation services. Needless to say the emphasis in such cases lies on facilitating the work of the specific service in question, with that neither the structures used nor the data stored are comparable. Thus, even if technology continues to advance a direct link between information systems of this kind can be established only with difficulty and with a considerable intellectual effort, on account of the incompatibility of the data and data-descriptions (computer link resp. load link).

- data exchange

  In order to achieve a division of labour in the manifold tasks in data acquisition it seems much more expedient to develop a uniform format for the interchange of data on external data carriers than to attempt to standardise internal data structures within EDP systems. A proposal to this effect will be dealt with in further detail below.

- generation of multilingual texts

  A frequent problem facing internationally cooperating information systems is that descriptions of documents are required in several natural languages. The alternative to employing hordes of translators is to generate summaries in several natural languages from a series of monolingually extracted words and contexts using a multilingual dictionary, (language-dependent) sentence patterns, and appropriate EDP programs (e.g. the TITUS system). This procedure will be valid in strictly defined subject fields with a finite vocabulary body provided an informative abstract is sufficient and highly elegant style is not required. However it is unlikely that such a system will be used to produce complete texts by an author automatically and simultaneously in several languages.

- automatic translation

  The discussion on whether it will be feasible and expedient to translate complete texts from one natural language into another automatically and without human and (human translators' share of the work less than 2o%) is something which I should prefer leave to more prominent experts. In my opinion, however, we are at present further than ever from an acceptable solution.

  Nevertheless, for the requirements of I & D a compromise seems to be feasible and desirable. Before complete texts are translated (and then discarded as useless by the recipient), a summary should be made automatically in the target language. This applies above all to languages less frequently understood (e.g. Russian in Western Europe) and to specialized fields with an established terminology. In such cases even an incomplete automatic translation of an abstract may yield enough information for the decision as to whether or not the whole text should be translated by a human translator.

This means that irrespective of the usual requirements made of translations for the purposes and demands of institutions and firms a high-speed summary is produced in advance, especially in the case of technical articles in journals, before the translator goes into action at the user's request.

The above list does not claim to be exhaustive. Rather it is meant to show in theoretical terms what is already feasible today with the assistance of EDP and how important a uniform interchange format is for multilingual terminological work.

4. MATER = magnetic tape interchange format for terminological/lexicographical data

4.1 General considerations

There is no denying that more and more meaningful use is being made of EDP for the purposes of terminological work. However, the result is that independently of one another a series of terminological data banks have come or are coming into being. Each of these data banks contains different languages and data, corresponding to the requirements of the users. This and the wider range of EDP systems and programming languages used mean that there is hardly any compatibility regarding content and data structure. Nor will standardization be able to change this situation very much in the near future, since no user can be told how to design his data bank optimally for his own purposes.

In this connection allow me to include by way of digression a few words on data formats for different purposes.

The first stage in the compilation of any data bank is data acquisition. This can be exploited for EDP use only if an appropriate

data acquisition format

is established. This will always be tailored to the user's data and his data acquisition equipment. Consequently there is hardly any scope for standardization here. This acquisiition format must be distinguished from the format in which the collected data are stored in the EDP system, which, according to (8), may be designated as the

implementation format.

The latter will be governed not only by the nature of the data to be stored and which of its parts can be used as search criteria but also by the type of EDP system used, the type and number of external storage devices, the access algorithms and the programming language used. Standardization is even less likely here than in the case of the data-acquisition format.

Data-acquisition format and implementation format are, however, very closely connected owing to the fact that it must be possible to map the data fields used for the former into the set of those used for the implementation format (data records utilizing all available categories will be mapped onto the implementation format).

It is therefore not yet possible (and this situation will hardly change for some considerable time) to build up a linked network of terminological data banks via teleprocessing, unless one regards the possibility of interrogating various data banks from one terminal as already constituting such a network. But here, too, considerable problems arise due to the use of various data structures and inquiry languages and to frequent switching.

As a result of this lack of coordination there is a tremendous amount of duplication in data acquisition, quite apart from the fact that findings made by one individual user have to be made "anew" by other users if these are to keep their data banks more or less complete. Then there are the data acquired for specific problems only to disappear after a time into some drawer system of archives (one need only think of the vast numbers of machine readable full texts, the existence of each of which is known only to a negligible number of insiders!). This enormous expenditure in time and money would be considerably reduced if it were possible to interchange data between the various data banks. For the time being, only interchange on external data carriers is feasible. An interchange format of this kind for data exchange on magnetic tape offers users at least the following advantages:

- uniformly defined categories make it easier to structure data so as to be compatible
- the transmitting agency needs only one program to provide its data (or sections of its data) for interchange purposes
- the receiving agency need "only" modify its input program to be able to process magnetic tapes from various transmitting agencies, i.e. to scan the transmitted data records and to incorporate data of the desired categories into its own data bank.

Such a general interchange format must be suitable for use in all activities of terminological work. It should therefore make it possible to interchange not only dictionaries of all kinds (e.g. encyclopedic lexica, monolingual defining dictionaries, multilingual lexica for translation),but also thesauri, classification systems, frequency dictionaries, concept standards etc.

A format of the above kind with the name "MATER" was proposed at the plenary session of ISO/TC37 "Terminology" in 1974. In the meantime, a draft of a standard has been prepared for MATER by the competent Technical Standards Committee for Terminology (FNT) at the German Institute for Standardization (DIN) and submitted to the technical world for discussion. A draft of an international standard has since been prepared.

An interchange format of this kind can be used effectively only if it is employed as a standard internationally and not just recognized nationally.

4.2 File structure in MATER

The set of all data forming one entity (e.g. a lexicon entry with definitions or a thesaurus entry with all its references etc.) is referred to in the following as an

interchange unit.

these interchange units are stored in the master file of interchange units. In general one interchange unit as a data record will occupy one block on the magnetic tape. It is, however, permissible to combine several short records into one block. On the other hand, records may exceed block limits only when they exceed the maximum permissible block length. Consequently records may not be segmented arbitrarily.

The first record *in* the master file, which at the same time must be a block in itself does not contain an interchange unit but holds information about the automatic processing of the file itself, such as date of compilation, details on the originator, tables for the transliteration of the character set used into extended ISO 7-bit code (EBCDIC) etc.

If the data bank consists of information from various sources, a complete bibliographical description for each interchange unit would take up considerable storage space. In addition, one interchange unit may contain several references, or, conversely, the same reference may apply to several interchange units. Provision has therefore been made for storing these comprehensive bibliographical details in a special file (on the same magnetic tape or the same set of tapes) and for including a reference in the interchange units themselves indicating the number of the corresponding record in the bibliographical file, or alternatively for indicating sources in concise form only (e.g. standard number: IS DIS 2709). This offers the additional advantage that the category scheme for this bibliographical information does not have to be developed in MATER itself but can be derived from a documentary approach (MADOK). In the same way further information for storing digitalized images, EDP programs etc. can be stored in another auxiliary file.

In addition to these files on tape a standardized accompanying formular is envisaged for interchange operations; this will contain information which is required for processing but which itself cannot be stored in machine-readable form - e.g. coding of accents by protypes, number of files, density, subject field codes etc.

4.3 Structure of interchange units

If one considers the physical structure of the data records which make up the interchange units, three parts may be distinguished :

a) The fixed fields

These consist of a series of data that can be used for quick processing or selection of interchange units. In addition to some length indicators relating to the constant size of the fields in the directory, these include the following:

- record length (decimal);
- record status. This indicates whether the entry is a new input, a change in the interchange unit or a deletion. In the case of changes the whole revised interchange unit, not just the changed fields, must follow;
- starting location of data: address of the first data field in the whole data record;
- identification number: consecutive count of the data records ;
- date of compilation of interchange unit (not of the data record);
- subject codes of interchange unit;
- language(s) of interchange unit;
- reference to the number of the corresponding data record in the bibliographical file (or other files) .

It is these data fields that make meaningful interchange possible; after all the transmitting agency does not produce for specific recipients, the latter must be in a position to filter out from the file the interchange units or those parts thereof that he wants to use for his own purposes.

b) Directory

Although the interchange units and the data fields contained therein are stored sequentially on the magnetic tape, the directory makes selective access to individual fields possible. For this purpose the directory contains a field for each data field in the same sequence as the data fields. These directory fields contain the following information:

- a three-figure code to identify the data field. This code is identical to that used in the category catalogue (and in the data field);
- length of data field (including field identification and field end character;
- starting location of the data field, in relation to that of the first data field;
- an indicator, which can serve both to identify the specific language of the data field and to provide a consecutive count in the event of repetition of the same data fields (with the same field identification).

c) Data fields

Here the terminological data itself is stored. Each field contains information belonging to one category and is of variable length. The field identification and indicator are repeated at the beginning of each data field. A fixed special character marks the end of a field.

When the directory is consulted this information is in fact redundant; however, it allows users who want to work through the fields sequentially to do so without recourse to the directory.

The data fields must be in the same order as in the directory. They must therefore not be arranged in the order of their field identifications. This makes it possible to repeat several identical fields (e.g. separate fields for multiple synonyms of the same descriptor!) or to group together within one interchange unit closely related data fields (e.g. morphological description of each synonym). Fields belonging to the same data group are signalized via the indicator.

The end of each data record is marked by a special characirr (redundant for safety reasons).

All the above information describes the form and structure in which an interchange unit is stored on the magnetic tape.

4.4 Category catalogue

If interchange is to be practicable, not only the physical structure of the data but also the categories used in interchange must be standardised. Consequently a category catalogue intended for use in interchange accounts for a substantial part of the draft MATER standard. If this catalogue is to be used for the extremely varied interchange purposes described above, it will, of course, take on vast proportions.

The category catalogue will therefore contain the following groups of categories:

- concise description and nature of source
- copyright information
- main entry
  This will generally be a lexeme. This will include all possible additional forms such as inflected forms, abbreviations, index forms etc.
- definitions and texts accompanying the main entry

- morphological-linguistic description (as yet only related to the German language)
- synonyms and homonyms of the main entry
- relations to the main entry
- additional information

It will be clear even from this short list that no user will be able to satisfy all categories at once. The catalogue should rather be regarded as a range of offers (like the catalogue of a mail order firm) with the aid of which interchange units can be structured meaningfully and compatibly. With this in mind the intention is for the transmitting agency to use in each interchange unit only those categories and the corresponding identifications for which it can also supply data. It is then imperative that the categories and their identifications be used exactly as described in the catalogue.

4.5 Application

The format described above has already been tried out in practice by transferring a number of machine-readable data files to interchange tapes. This does not mean that interchange has already commenced but it has been demonstrated that the programming effort necessary remains within bounds.

5. The problems of terminological work with the aid of EDP

In conclusion I should like to touch upon a number of problems of monolingual and multilingual terminological work .

When preparing a systematically classified compilation of definitions in a specialized field, it is inevitable that the texts of the definitions will contain technical terms that have already been defined elsewhere in the collection. It is relatively easy to make allowance for this when deciding on new definitions. But as soon as an inventory consists of a few hundred definitions it becomes almost impossible, unless EDP is used to establish whether a newly defined term has not already been used and marked in existing definitions. Only with the aid of EDP can this network of mutual context relations be checked. All this becomes even more problematic when translating. Quite apart from the fact that it is hardly possible to translate a "forward-looking" terminology from one language into another, how is the poor translator to check the uniformity of the translation in the target language if he has not access to the identification of the defined terms in the source language? What does he do when terms in the source language are newly coined or when assigned definitions differ from those given in other sources?

Let me give an example:
In German it is customary in data processing to adopt internationally standardized definitions where this is at all possible. Here a phraseological dictionary can be of great assistance to the translator. However, this fails when in German these same definitions are made the subject of terminological research aimed at finding inconsistencies with regard to related specialized fields. In this process an attempt is made to uncover precisely the (concealed) homography and polysemy which do not necessarily exist

in other languages. This involves differentiating between definitions by means of qualifying expressions which are necessary in one language only and perhaps cannot be translated.

This may be illustrated as follows:

a) In data processing, the English "character" is translated as "Zeichen", while semiotically "Zeichen" is to be translated as "sign".

b) How is one to translate an (artificial) definition such as "Zeigzeichen sind Sinnzeichen, denen keine sprachliche Lautung zugeordnet ist" (9) at all without creating corresponding neologisms in the target language?

This should be regarded merely as a contribution to the discussion on the limitations of multilingualism in order to show that, though EDP is a great help, it cannot solve everything. The intellectual work involved in overcoming language barriers cannot be avoided.

6. Literature

(1) Wersig, G.: Probleme und Verfahren der Terminologiearbeit: Der Sprachmittler Nr. 2, 1973, S. 58-71

(2) Das Große Duden-Lexikon Bd. 5, Mannheim 1966

(3) Beling, G.: Terminologie, Thesaurus und Klassifikation - Zusammenhänge und Unterschiede in: Kschenka, W., T. Seeger, G. Wersig (Hrsg): Information und Dokumentation im Aufbruch, Festschrift für Prof. Dr. H.-W. Schober, Pullach bei München, 1975

(4) Beling, G., G. Wersig: Zur Typologie von Daten und Informationssystemen, Pullach bei München, 1973

(5) Komitee für Terminologie und Sprachfragen der DGD (Hrsg): Terminologie von Information und Dokumentation, Redaktion: U. Neveling, G. Wersig in: DGD - Schriftenreihe Band 4, Pullach bei München, 1975

(6) Wersig, G., U. Neveling (Compiler): Terminology of Documentation, The UNESCO Press, Paris 1976

(7) Vajda, E.: Die Notwendigkeit der Terminologiearbeit in der Information und Dokumentation in: Deutscher Dokumentartag 1974, Bonn-Bad Godesberg vom 7.-11.1o.74, Band 2: Probleme der Terminologiearbeit, München 1975

(8) Bernhardt, R.: Formate in der nichtnumerischen Datenverarbeitung in: vgl. (7)

(9) DIN 2338: Das Begriffssystem Zeichen Entwurf 1971 Berlin