Erik Sundström
TNC, Stockholm, Sweden

Introducing TNC and the TERMDOK System

Sundström, E.: **Introducing TNC and the TERMDOK system.**
In: Intern. Classificat. 5 (1978) No. 2, p. 86–90

The Swedish Centre of Technical Terminology, TNC, is a non-profit organization engaged in elaboration and dissemination of terminology. Work at TNC is centered on a procedure which is intended to safe-guard the influence of interested parties on the shaping of the terminology. The TERMDOK system has been conceived of as an information handling aid in this work. It contains programs for updating and search in a term bank, making possible interactive editing of terminological data. A shift in emphasis is discernible in the elaboration of terminology, from systematic work on more or less clearly delimited technical fields to a form more adapted to the need for rapid decisions on individual terminological questions. Computerization is one of several factors influencing this tendency. (Author)

The processing of terminological information is full of recurring procedures on various levels. Searches yield data to aid in further searches. Changes or additions at one point in a system of definitions are likely to produce chains of adjustments involving most if not all of the system. The elaboration of new sets of definitions necessitates frequent comparisons with already established terminology.

An institutionalized activity whose very existence rests on its ability to survey, process and keep up-to-date an extensive base of terminological data will depend heavily on an efficient information system. These are the conditions governing operation since 1941 at Tekniska Nomenklaturcentralen (The Swedish Centre of Technical Terminology), abbreviated TNC. These, also, are the reasons why much work has been directed, since 1968, towards the design of a system for computer-assisted processing of terminological information at TNC, the TERMDOK system.

1. Activities at TNC

TNC performs normative work on technical language. A series of glossaries has been published, now comprising 70 volumes. An average glossary contains some 700 technical terms with definitions in Swedish, occasionally also in English. Corresponding terms are given in a number of languages, mostly English, French and German but not infrequently also in Danish, Finnish and Norwegian. In some cases, Russian and other languages have been included.

The most important property of a TNC glossary is its concept structure, manifest in the wording and mutual correspondence of the definitions. This structure invariably rests on a conceptual model of the technical field under consideration, and is also expressed by means of a classification, which in most cases is not shown explicitly in the printed version of the glossary. The conceptual model and the classification are necessary instruments in the elaboration of the terminology. However, in the experience of TNC, they are seldom suitable as the main search instrument, at least not so to the general user. Lately, a concession has been made to international usage in so far as the classification scheme has been reproduced as a separate part of the glossaries.

1.1 Adjustment of views

The description above only refers to the formal side of the elaboration of terminologies for different fields. The gist of the work, and the single most important justification of TNC's existence, is the bringing together of a representative selection of interested parties in a procedure leading to the ultimate establishment of an approved terminology.

Four out of a staff of seven are chiefly engaged in this work, the other three having mainly over-head tasks: administration, economy and documentation. Committee meetings are the typical setting for the procedure mentioned. A tendency is clearly discernible, however, towards decision-making in a form more adapted to the need for rapid and yet tenable recommendations on individual term questions. This tendency is both evoked and accelerated by the bringing into action of computerized methods for terminological work.

An adjustment of the almost invariably divergent views on salient points is necessary to achieve a satisfying result. It implies specific demands on the staff involved. The interested parties are represented by experts on the field under consideration. Judgements passed by these experts on factual circumstances can seldom be questioned by TNC. However, TNC's representative can act as an arbitrator when the experts display conflicting views. Such conflicts serve to bring to the surface questions of major terminological relevance. When the experts agree, the need for terminological efforts can justly be questioned.

This part in the procedure, the arbitration, is greatly alleviated by a more than cursory insight in the specific technical field and above all by a broad technical knowledge. Such knowledge will help by providing analogues in fields other than the one under consideration.

Hand in hand with this admittedly delicate task, the TNC representative has to provide an inflow of linguistic discrimination into the procedure. Again, experience of similar problems in earlier terminological work, together with insight into the fundaments of linguistic science, constitute a necessary prerequisite for successful achievements by TNC in the elaboration of terminology.

1.2 Other activities

Authoritative rules for the use of technical language are formulated and published by TNC as separate volumes, enjoying a wide distribution among technical writers at all levels. Two news bulletins are issued regularly, one on

general topics and one on the rationalization of terminological work.

By agreement, all draft national standards, and several international standard documents in terminology, are submitted to TNC for checking. Consultation is given to several private and government organizations. A free telephone service is offered on terminological questions reaching beyond the fields of science and technology. Articles on terminology are published in a number of technical magazines. Frequent appearances are made by TNC staff at courses and meetings dealing with technical language.

Essentially a private institution built on a membership system, TNC relies heavily on government support to all these activities. Responsible for this support since 1968 is the Swedish Council for Scientific Information and Documentation, SINFDOK.

1.3 Co-operation schemes

TNC is a partner in NORDTERM, a co-operation scheme set up by terminology agencies in Denmark, Finland, Norway and Sweden. NORDTERM aims at elimination of duplication of labour and better co-ordination in form and content of cognate terms in the languages used by the partners. NORDTERM meetings have been convened two times, in Stockhohn (Sweden) 1976 and in Bergen (Norway) 1978. Two working groups have been set down, one to take care of a course on terminology to be held in Copenhagen at the end of June 1978, and one to realize on an experimental scale communication between term banks in the participating countries, see section 2.7 below.

Originally a Danish initiative, the NORDTERM cooperation scheme has proved to be of great value to TNC. The benefit derives in no small measure from the fact that the partners, though working in a similar environment and having the same general needs to fulfil, manifest entirely different attitudes, experiences and modes of operation. It is probable that NORDTERM will bring the partners closer to each others in this respect. However, at the time being the differences serve to stimulate discussion and thus to augment the quality of methods used and results reached.

From its very beginning in 1971, work at Infoterm has been followed with close attention by TNC. The need for efficient collection, analysis and dissemination of terminological information of all kinds and at all levels on a global scale is sorely felt by TNC. The plans put forward by Infoterm on the development of a network for terminology, TermNet, have been emphatically endorsed by TNC.

2. The TERMDOK System

2.1 Scope of the system

A description of the system must start with an attempt to define its boundaries and general functions. The designation TERMDOK should not primarily be taken to stand for a specific set of computer programs, although some of the most unique facilities do exist in that form. Rather, TERMDOK is the collective name for a number of rationalization measures, currently investigated and brought into action at TNC. The various activities may seem disparate at times, but converge towards a common goal, the ultimate integration of all information handling functions at TNC. TERMDOK includes methods used in the collection, storing, processing and dissemination of terminological information, to give a frequently quoted formulation of the scope of the system.

The following list of sub-systems, though not exhaustive, will give an out-line of areas of particular interest in the work.

- pre-TERMDOK, since 1968, defined a term record format and put down methods for computer-assisted photo-composing of glossaries
- TERMDOK 1, from 1973, formulated rules for the establishment and growth of a term bank
- TERMDOK 2, from 1975, carried into effect direct access to the term bank, making possible interactive processing of terminological information
- TERMDOK 3, from 1977, aims at augmented functions in the system by means of linguistic techniques
- TERMDOK 4, from 1978, is directed towards telecommunication between different term bank systems and is intended to yield, as a spin-off, methods to make terminological information publicly available in a data network.

The items in the list represent the historical development of the system, or more exactly its financial history. Most of the work has been done in concentrated efforts and under strict time schedules. It has been funded by SINFDOK.

Though admittedly of an ephemeral character, the partitions given in the list above may serve as a point of departure in a cursory description of the TERMDOK system.

2.2 Record format considerations

Building a term bank system invariably begins with the breaking down into content categories of the logical entity in terminological work, the term record. A term record might be envisaged as the sum of all information relating to one and the same concept.

It was early recognized in TERMDOK that a distinction must be made between the following functions of a term record format.

- a term record format as a set of formal rules for the structuring of terminological information (acquisition format)
- a selection of standard content categories as a means to communicate and receive terminological information (exchange format)
- a term sheet layout as one of several possible applications of the formal rules and as the chief input and output aid for the term bank (processing format)
- a term entry, printed in a glossary or displayed on a CRT screen, as one of several possible arrangements of the information contained in the term bank (presentation format)

On the term record level, a fundamental difference exists between linguistic and extra-linguistic categories. Of the latter, 100 are permitted by the system. Some ten of these are in frequent use, including project designation, term number, classification codes, date of establishment, identification of editor and references to illustrations.

The record format can accomodate linguistic information in 100 content categories for each of 26 different languages. So far, ten languages have been represented in the input: Swedish, English, French, Spanish, German, Danish, Norwegian, Finnish, Russian and Japanese. In frequent use are some twenty content categories: entry term, synonym, near synonym, antonym, pronounciation, different aspects of grammar, abbreviation, definition, remarks (on usage, special meanings etc.), comments (pros and cons of the term and its definition), cross-references on different levels etc.

2.3 Computer-assisted composing of glossaries

Probably the most important rationalization measure in an organization like TNC, prior to the establishment of a term bank, is to achieve computer-assisted photo-composing of glossaries. Savings in manual work is connected mainly to two factors: simplified input and output operations during the early stages of elaboration of the glossary and greater convenience in the generation of indices of various kinds.

The blessings of computer-assisted composing are best put to use with the existence of local input. This is achieved at TNC by means of a cassette tape encoder with visual display, disc storage facility, low speed printer, and programming capability. While greatly relieved by time-saving codes for recurring information, the editor has to cope with a major problem in the representation of special characters (Greek letters, superscripts etc.) on different graphical levels. Editing, to be flexible, must give cheap print-outs with a high degree of correspondence to the final photo-composed text. TNC has access to fairly flexible methods in this respect.

Of great potential importance is the possibility to call any special character by name. By writing '§beta' on the input tape, the cheap print-out gives the string 'beta', while photo-composing will provide the desired result of the Greek letter beta. As a generalization of this procedure, the editor can write Russian words in the recommended ISO transliteration and leave to the system to convert them into Cyrillic script.

2.4 Establishment and growth of the term bank

To safe-guard against negative suboptimizing, broad studies were carried through at the beginning of the work. It became evident that provisions must be made from the very out-set for a term bank. This must have the following characteristics:

- a completely flexible term record
- neutrality as to language and classification system
- freedom from built-in data structures
- no restrictions in input or output media.

By definition, a term bank contains terminological information which can be processed and accessed subject to criteria which are not necessarily predictable at the time of establishment of the term bank. One important point is that, as a rule, access is not meaningful if the term bank does not contain information with a certain degree of completeness. The term bank must, in other words, have reached full size on some level before it can be put to real use. This fact is in clear conflict with the following guiding principle, layed down by TNC at an early stage in the work as a concession to financial restrictions:

- computerization of terminological work should proceed in well-defined steps, each complete in itself and each carrying its own costs.

Planning had to circumvent this conflict.

It was early recognized that the single most expensive step towards establishment of a term bank, made up of all term records published since 1941 by TNC, was the actual writing on magnetic medium. This lead to a decision to avoid massive input at the first stages of the work. Given a term record format, TNC had to choose objects of rationalization which did not presuppose access to a term bank in the first place, and which did not rely on a high degree of integration with other terminological activities in the second place.

As a by-product in all activities complying to these restrictions, terminological information was converted into machine-readable form. By mid-1978, the term bank contains some 50000 term records. The rate of growth is well over 1000 records per month, representing the elaboration of new glossaries at TNC and commissions to process new external material.

Cost per stored record, search facilities and the prospect of networking with other term banks will in a near future decide the direction of this expansion. The issue is whether the term bank should continue to reside in a dedicated system within the precincts of TNC, or rather be transferred to a larger, general system.

Irrespective of which choice this issue will lead to, it will most probably be closely associated with a decision to input some 30000 additional term records. These will to a large extent be derived from national standard publications. Then, at the level of some 75 000 term records (non-authoritative external collections not included), no large-scale retroactive input will be required. The future growth of the term bank from that stage on will roughly correspond to the annual increase of new, authoritative technical terminology in Sweden.

2.5 Direct access and interactive processing

A set of programs, unique to the TERMDOK system, is the instrument for direct access in a term bank. Written in assembler language, the program system implements an inverted-file technique with hierarchical tree search and Boolean logic. It performs with short response times even at large term banks and its operation requires no specialized knowledge.

An updating program extracts index words from a number of pre-determined fields in the input term records. Some 100 grammatical words are exempt from index generation. Both the selection of these stop-words, of fields yielding index words and the conventions governing the limits between index words can easily be changed in the program. The choice of stop-words and fields to be indexed are made with a view to balance the need for search flexibility against obvious limitations in storage capacity. A cross-reference facility is used mainly to take care of orthographic variants (search on "colour" will yield all references to both "colour" and "color" simultaneously). It is the intention, however, to make this facility the carrier of an embedded thesaurus structure in the term bank, providing references to broader, narrower and related terms.

Term records can be deleted by an operator message during the updating or by means of delete messages in

the input text file. Naturally, deletion implies all corresponding changes in the index file. Unless provided by the operator, the system will assign a record number in increasing order from a chosen initial value to each new input item belonging to the same term collection. Internally, all records are accessed by a combination of project designation and record number.

When running the updating program, the operator will be asked to state the name of the term bank (several banks can be accommodated simultaneously), the name of the input text file and the name of the index file. Also, messages to the system are accepted at this stage. During the updating run, the system will display the actual number of term records, the number of deleted records and information on total space occupied by the bank.

When no messages to the system are offered, updating will in practice be performed by a chain file which also invokes sorting and index updating programs. This means that the operator can be completely ignorant of the mechanism, all that is necessary is a call for the chain file. A complete updating run will provide additional statistics on the number of index records and transaction records as well as the number of deletions and cross-references in the index file.

Search in the term bank is performed in dialogue between operator and system. Upon naming of a word, the primary response is a number, representing the count of term records containing the given string in any of the indexed fields. Expansion or reduction of this number is accomplished with Boolean logic.

The search format can be expressed in the following way.

(AND/OR) (xnnn)aaa

Everything except the search word, aaa, is optional. AND signals a reduction of the search space, OR an expansion. The string xnnn represents a field code in the term record. The presence of this string restricts search to the corresponding field.

When a desired number of term records is scored, output can be ordered over CRT display or printer or both. Also, output of an appropriate part of the index file can be invoked for inspection purposes.

Commands are available for copying of the accessed term record set onto a working file. Modifications in the records can be made with the use of a standard program, followed by updating as described above.

The primary use of direct access on the existing hardware is as a support to the editing of terminological data. Further integration with terminological work in a broader sense necessitates a more powerful storage capacity than the one now available to TNC.

Term bank interaction has proved to be of value in the processing of moderately sized collections of terms. The typical case is the tracing and performing of a series of secondary changes in different term records, arising from one single primary change at some place in the bank. Each such set of connected changes can be done in one work cycle, as opposed to serial processing with its need of pre-ordering in a determined sequence. The ease with which classification codes and term record subset designations are used in the search procedure is also of undisputable value to the editor of terminological data.

2.6 Bringing linguistics into the system

Methodological questions of a more or less overt linguistic nature turn up at each juncture in the development of the TERMDOK system. It is true that linguistic technique is present in terminological practice as such, and not only belongs to its data-processing tools. However, computerization tends, in a manner of speaking, to pose clear-cut questions and to demand clear-cut answers. Also, negligence to methodological problems in a computer context is known to lead to a relatively high penalty. Examples abound in actual terminological practice.

Establishment of a morpheme inventory is of great intrinsic value to an activity like TNC's. Information on semantic and compounding properties etc. can be summarized and made easily accessible once the principles of morpheme analysis are decided on with sufficient clarity. Also, such technique will alleviate search in a term bank. Permutation of morphemes will lead to the possibility to retrieve terminological information which only nearly satisfies a search query. Discrimination of levels of such near hits implies an associative search method. The vital point, evidently, is to find the optimum degree to which morpheme analysis should be carried. Extensive experimentation at TNC during the last three years has failed to disclose any serviceable, complete set of criteria on the recognition of morphemes relevant to terminological work.

Still more difficult to identify and to handle in a satisfactory manner in terminological work is the syntagma: a (set of) word(s) having some syntactic relation to some other (set of) word(s). The syntagma must be susceptible to identification and search because of its ability to be terminologized, i.e. turn into a technical term. The instruments to achieve this goal are more or less non-existing to-day, at least in TNC's practice. Again, morpheme analysis is a pre-requisite. A technique can be envisaged, involving permutation of morphemes, recognition of inflected words and long-range comparison of linguistic entities. One supposedly profitable use of this technique is to search technical texts for recurring locutions of potential terminological interest. The implications are important to TNC: candidate objects of study can be identified 'in situ' at an early stage in the terminologization process.

Great advantages can be gained with better linguistic control of definitions. A deliberate attitude to the vocabulary used in definitions will, together with formalization of references to other term records, constitute a point of departure for several time-saving techniques in the processing of terminological material. Automatic checking of contradictions and considerably augmented search mechanisms are advances within reach.

These few gleanings into questions now facing TNC may serve to underline the need for co-ordinated development work on linguistic techniques in terminology. Plans are made up within TNC to approach professional linguists with an outline of a project work in this vein. It is felt that a mutual exchange of experiences can yield many valuable results, and it is certainly true that the competence to judge linguistic questions must be cultivated at TNC. Even if solutions have been found to most, if not all, questions touched upon here, these solutions have to be adapted to terminological practice.

2.7 Interchange and dissemination

Ultimately, an integration is seen between the term bank system as an instrument in terminological work proper and as a means to disseminate terminological recommendations to all interested parties, including the interconnection of TERMDOK and other similar systems.

The coming into existence of public data networks employing packet switching technique offers new possibilities here. Fast and cheap access from geographically distant locations and without human intervention at the term bank certainly represents a major advance over the present mode of operation. But this facility alone will hardly motivate any consuming interest from an institution like TNC as long as all but a few of its regular customers are foreign to the technology involved. It is only when viewed in conjunction with the possibility of communication between a number of independent, large-scale producers of terminology that the use of a data network as a means to disseminate term records gains its prime importance.

A terminology network has been conceived on a common Nordic basis. A couple of circumstances have contributed to this thought. Terminology agencies exist in Denmark, Finland, Norway and Sweden, each with approximately the same goal and general mode of operation. A formal cooperation framework has been established in the NORDTERM scheme, see section 1.3 above. The technical prerequisite has come into existence with the realization of an experimental data network specialized on information and documentation needs, SCANNET.

A guiding principle behind the concept of a common Nordic terminology network is that joint elaboration of specialized terminologies can be freed from the present need for simultaneity in actual operation. However, some minor standardization measures are necessary in term record format and classification system. The most important benefit will probably be considerable savings in manual work each time new terminology is elaborated in co-operation between two or more institutions. Also, the possibility of direct interrogation in several national term banks implies savings both at the central institutions themselves and to a number of other interested parties.