[From: *Human translation, machine translation*. Papers from 10th annual conference on computational Linguistics, Odense, Denmark, 22-23 November 1979, ed. Suzanne Hanon and Viggo Hjørnager Pedersen.]

121.
THE SETTING UP AND OPERATION OF A DANISH TERMINOLOGICAL DATA BANK (THE DANTERM PROJECT)

by Lene Frandsen and Bodil Nistrup Madsen

I (Lene Frandsen) and my colleague, Bodil Nistrup, who are members of the DANTERM project group at the Copenhagen School of Economics and Business Administration, Faculty of Modern Languages, intend to give you a short report on the DANTERM project, i.e. The Danish Terminological Data Bank.

First I shall state the objectives of the term bank and then give you a short description of the initial phases of work as well as a description of some of the information categories of the DANTERM record. Thereafter Bodil Nistrup will focus on the ongoing system development.

OBJECTIVES

As terminology makes out an essential part of the LSP*) taught at the Copenhagen School of Economics it was natural to let work on a terminological data bank be done there. The Danish "Group of Terminology" - which is a nationwide co-operation in theoretical and practical terminology work - supported the setting up of a project group of research fellows at the Copenhagen School of Economics. Work was started in 1975.

One of our own motives at the Copenhagen School of Economics was a desire to record the results of the terminological research which is carried out by students and teachers at the four language departments in such a form that it could be made accessible to a large public. Besides discussions in the "Group of terminology" made it clear that trade and industry as well as the public sector would welcome such a tool for overcoming the growing language barriers.
As a result of the discussions the project group formulated the following objectives of DANTERM:

1) to set up a data bank comprising Danish terms and their equivalents in one or more foreign languages. If possible, these entries shall at least be accompanied by definitions of concepts and examples of usage of the terms.
2) to keep up and extend the data bank.
3) to make information in the data bank available to the following users:
   - translators and lexicographers
   - trade and industry
   - the public sector
   - international organisations including other data banks
   - members of the Danish Group of Terminology
   - teachers and students at the Language Departments of the Danish Schools of Economics and other higher institutes of learning.

The following services shall be available to the users:

- on-line retrieval from the user's own terminal
- off-line retrieval e.g. print-outs as a response to requests for batch retrieval
- production of glossaries and dictionaries covering specific subject fields
- answering of written and telephonic inquiries about the meaning and translation of individual terms.

---

*) Language for Special Purposes

The bank may also provide terminology for "common language" dictionaries containing LSP entries. Apart from that it will probably be possible to use the data bank for various linguistic studies.

We hope that the term bank can begin operation in 1980. To be efficient from the very beginning we shall concentrate on covering a few selected subject fields.

DATA BANK SURVEY

In the first phase the project group made a comparative study of existing terminological data banks, how they were set up and operated.

The study comprised:

The TERMDOK SYSTEM, Tekniska Nomenklaturcentralen, Stockholm,
EURODICAUTOM, Luxembourg,
TEAM, Siemens Sprachendienst, Munich,
The term bank of the Bundessprachenamt, Hürth.

The conclusion of the study was that it would not be possible to take over any of the systems studied, mainly because their objectives and user groups were different from those envisaged in DANTERM.

However, it was clear that the information categories of the DANTERM record should as far as possible be compatible with existing data bank records in order to facilitate exchange of terminological information.

DEVELOPMENT OF THE DANTERM RECORD

In the second phase – the development of the record of information categories - we therefore endeavoured to adopt all relevant information contained in the term sheets we had studied as well as the information necessary to cover our own special needs. With a view to future exchange of data we also studied MATER (i.e. ISO Draft Proposal 6156 "Magnetic Tape Exchange Format for Terminological/Lexicographical Records"), but found that MATER was only of very moderate use to us, as it foresees lots of information on morphology and syntax which we do not want to supply. On the whole we think MATER is confusing because some categories are not transparent and the multilingual aspect is not clear.

For test purposes we have transferred a number of terminological theses to our record format which in the course of time has undergone several changes. By the end of 1978 we had prepared what we consider the final version of the DANTERM record - although small corrections naturally may be necessary in the course of the system development. It will be possible to transfer all information from other term bank records, of which we know, to our record, but there is not full compatibility.

Survey of the information categories of the DANTERM record (figure 1)

Term with additional information

Naturally the term itself is the main entry, which gives rise to all the other entries most of which determine the term or the concept behind the term.

First of all the term is specified as to language. As regards the number of languages in the bank our plans so far comprise the main languages taught at the School of Economics, i.e. English, French, German, Spanish, and Danish. Later on we may include Italian and Russian.

Our starting point was that we would establish a bilingual bank because that is what our resources until now enable us to do. For instance a research fellow's preparation of a special subject in normally two languages, or a student's thesis, normally also only in two languages, results in bilingual material. In other words the outcoming terminological product is built up of language pairs and one of the languages of any language pair will normally be Danish.

Material prepared according to the bilingual method will no doubt be utilizable for multilingual retrieval, i.e. retrieval between two or more foreign languages, but only with some limitations. As in several subject fields, e.g. law, education, social sciences there will very often not be the same degree of equivalence between for instance a French and a Danish term on the one hand and the same Danish term and its English equivalent on the other hand, free and unlimited combination of all languages in the bank would pose many problems and often lead to unreliable output information. After much discussion of the problems we have decided that we will not prevent anybody from getting any language combinations he may want, only we store information making the computer automatically tell the user who asks from one foreign language to another whether there is full equivalence or only partial equivalence between the two foreign terms, or the computer says "degree of equivalence not examined". At multilingual retrieval "degree of equivalence not examined" will most often be the case and then the user must find out himself, by reading the definitions supplied, whether the foreign terms in question are sufficiently equivalent for his translation purpose.

Naturally the ideal thing would be to prepare multilingual material from the very beginning. We do not know whether at some time we shall have resources for that ourselves. There is no hindrance, however, to the bank's taking over and utilizing material which was prepared multilingually, e.g. from dictionaries or other term banks.

Classification

As all of you will know some kind of classification is necessary to distinguish homonyms. But, to facilitate the exchange of data, existing and future term banks should agree on a common classification system. Great efforts are going on in a special European classification working group to create a common classification system specially fit for terminology. The group, however, endeavours to make only a basic system which can then be further developed by the various term banks in accordance with their different needs. DANTERM is represented in that working group.

Grammatical information on the main entry

We intend to give only little grammatical information, i.e. part of speech, gender and inflexion of the term in as far as the information is useful or interesting from a terminological point of view.

We have discussed the possibility of including far more detailed linguistic information, inspired by MATER for instance a detailed description of the complexity of single and multiword terms. Such information might be used for teaching and research purposes, but from our tests of the record in practice we learned that it is far too time-consuming to make the terminologist supply such detailed information, so the idea was abandoned. But it will always be possible to add such information at a later stage.

Special field of application

A freely chosen verbal remark, meant as a supplement to the classification of the term. The terminologist may refine the classification by stating that the term is used for instance in a special type of machine.

Sources

As we do not intend to standardize terminology, but naturally know that a term bank will always have some normative effect, it is very important that all terms, definitions, and other texts are provided with sources so that the user - if he wants to - can judge the reliability of a given term for himself.

Texts

The purpose of definitions or explanations is clear. "Other texts" are for instance collocations. One subdivision is the socalled "parallel texts". Texts in different languages are socalled parallel texts if the texts as a whole or expressions in the texts resemble each other so much that they can be used as translations of each other although they are not recorded as terms.

Conceptual relations

Most of the terminological theses include systems of concepts which might be a useful guide to the user in his search for the right term for translation. For some time these systems will be kept in a manual file, but in the record a code giving the number or name of the system is supplied. We hope that some day it will be possible to include in the bank a file containing e.g. graphical representations of systems of concepts, tables, and illustrations. Part of the systems of concepts is reflected in the record by means of the categories broader term, narrower term etc. These can normally be found as main entry in their own full record.

Comments

It is possible to give comments on texts, terms, sources or whatever necessary.

After this short survey of our information categories you may think that we have too high ideals. Time will show, but actually we do not think that this record is unhandy and too comprehensive. We do not intend to demand all information for each main entry. The bank will no doubt contain lots of records which besides the term, classification and language only supply a definition or an explanation or maybe some text examples. By this comprehensive record we have, however, established the possibility of supplying much more information so that the bank - at any rate in some subject fields - will be able to serve all potential needs which our diversified user categories may have.

SYSTEM DEVELOPMENT

Prerequisites

During the development of the DANTERM record the project group continued collaboration with the term banks included in the above-mentioned comparative study. In that way we have gained valuable experience of term bank systems.

For instance we have studied the TEAM system in detail and then made an adaptation of the DANTERM categories to the format of TEAM. Finally we tested a series of TEAM programmes on a number of DANTERM test records.

Furthermore the installation of a terminal connected to the EURODICAUTOM system in Luxembourg has given us useful practical experience in the use of a term bank.

In order to facilitate communication between existing and future Scandinavian term banks a special working group has been set up by NORDTERM. NORDTERM is a cooperation between Scandinavian terminology organisations with the purpose of encouraging and coordinating terminology work in Scandinavia.

At the beginning of 1979 the Copenhagen School of Economics got a new computer system, consisting of two computers, namely PRIME 550 and 450, connected in parallel. In connection with the grant for this computer system the Danish committee on a.d.p. capacity made a special grant for development of DANTERM software.

System requirements

The first step in the system development was to work out a specification of the requirements, as to search procedure and output-forms, which are to be met by the DANTERM system.

This specification of system requirements has formed the basis of intensive discussions with various a.d.p. consultants concerning realization either in the traditional way, i.e. by means of 'tailored' programmes, or by means of a data base management system, as e.g. the DBMS of PRIME. We are still studying advantages and disadvantages of these two different methods.

One result of the cooperation with the a.d.p. consultants was a logical structuring of the DANTERM information categories.

This data structure can to some degree illustrate the philosophy and the system requirements of the DANTERM project group.

Data structure

The data structure (cf. figure 2) is represented by means of a Bachman diagram. Each box represents a record class and each arrow represents a set class. The sets (chains) connect owner records and member records.

In each box the name of the record class is written in capital letters and the information actually contained in each record class is written in small letters.

As will appear from the diagram three different types of relationship between owners and members are used:

1) hierarchical relationship between two or more superordinated record classes (owners) and one subordinated record class (member).

The boxes LANGUAGE, DANTERM CLASSIFICATION, OTHER CLASSIFICATION and PROJECT are all owners of the box TERM.

2) hierarchical relationship between records of the same record class.

The link HIERARCHICAL CONCEPTUAL RELATIONSHIP connects two terms of the same language, which are super-/subordinated.

The box TYPE OF HIERARCHICAL RELATION qualifies the conceptual relation between the two terms in question, as being either a generic, a part-whole, or a not identified or other relationship.

4) relationship between pairs of records of the same record class.

The link EQUIVALENCE connects terms of different languages.

Obviously this link is of great importance. We have discussed very much the problems in connection with equivalence between the languages. Here I can neither describe the pairing of terms nor how the system is to interpret the information on equivalence supplied by the terminologist.

The record class TYPE OF EQUIVALENCE contains records stating either that there is full or partial equivalence, that the degree of equivalence has not been examined, or that one of the terms in a language pair is a translation (which may be proposed if the concept in question does not exist in one of the languages) . The records of the class EQUIVALENCE contain any further comments on the degree of equivalence and information on reversibility.

The link EQUIVALENCE also connects synonyms of the same language and antonyms.

The link CONCEPTUAL RELATIONSHIP BETWEEN PAIRS connects terms in a successive relationship.

In the link PARALLEL TEXTS the above-mentioned special parallel texts in two languages are linked together.

Search fields

Search fields are the fields in which the user can search the source language term of which he wants the equivalent in a given target language.

We intend to offer the user some pre-defined sets of search fields. But we also want the possibility of composing individual sets of search fields.

Of course a pre-defined set of search fields will first of all include the term. Since the queried term might as well be recorded as a synonym, an abbreviation or a full form of an abbreviated term, or as an orthographic alternative to a term, the search must normally comprise those categories too. Also the so-called parallel texts might belong to a basic set of search fields.

The basic set of search fields can be extended by adding two subsets of information, 1) terms representing super- or subordinated concepts and 2) all texts in extenso.

Search terms

If the search term is a single-word term that cannot be retrieved in the data base, the system shall render any multi-word terms which include the search term.

If the search term is a multi-word term that cannot be retrieved in the data base, the system shall render any parts of the multi-word term irrespective of the order of the individual words. The user can specify the relative positions of the individual words of the search term.

Furthermore it must be possible to search components of compounds and derivations.

It shall be possible to use 'positive' and 'negative' truncation at the search. This means that the system shall render an answer, even when the retrieved term is longer or shorter than the search term.

Examples:

'positive' truncation

question: car
answer: cars

'negative' truncation

question: haulage
answer: haul

both 'negative' and 'positive' truncation

question: magnetic storage
answer: magnetic store

The user specifies himself how much the answer must deviate from the question. At a later stage we might define an automatic function between the length of the question and the answer. The EURODICAUTOM system operates with such predefined functions.

Selection criteria

In order to limit the scope of search some types of information can be used as selection criteria both when the user queries a specific term and in the opposite case, when one does not query a specific term, but wants to select some subsets of the database (i.e. for production of glossaries and dictionaries).

When a user queries a term he must be able to limit the number of answers by specifying his question by means of one or more selection criteria, first of all of course language. Another very important criterion is classification. If the user specifies his question by the correct classification, homographs belonging to other subject fields will be omitted.

For the selection of subsets of the data base the most important criteria will probably be language, classification and name of project.

That is the reason why exactly these information categories have been made owners of the term in the data structure.

Many other information categories are potential selection criteria. Selections can be based on a specific character string of a field (i.e. the wording) or on the occurrence of a field (i.e. whether the field has been filled in or not).

If for instance one wants all terms from a certain source, the selection is based on the character string of the field. If, on the contrary, one wants all terms which are supplied with a definition, the selection is based on the occurrence of the field.

Information on occurrences of the queried term

If the queried term occurs several times in the data base, the system has to inform of number of occurrences and, if desired, list the subject fields. The system shall also be able to list the search fields in which the queried term was actually retrieved. By means of a selection criterion the user can then limit the number of answers, if he wants to. As a matter of fact we learned from our use of the EURODICAUTOM system that it is terribly irritating
to get endless lists of answers to a question.

Information profiles

For many purposes it is unreasonable to let output comprise all information recorded for one term.

Therefore we have tried to define some standard sets of information, which the user can choose, the so-called information profiles. We have defined a basic information profile that supplies the minimum information which, in our opinion, any user ought to receive.

By means of certain standard commands this basic information profile can be extended by pre-defined subsets of information.

However, it must be possible to compose individual sets of information, i.e. the users (preferably terminologists) shall be allowed to define their own output by combining categories according to their own immediate needs.

Figure 1:

Survey of the Information Categories of the DANTERM Record

1. Classification

DANTERM classification
Other classification
Name of project

2. TERM with additional information

Language
Term, main entry
Pronunciation
Full form
Abbreviated form
Orthographic alternative
Grammatical information on the main entry
Language region
Special field of application
Stylistic information
Sources
Comments on degree of equivalence
Information on reversibility

5. Texts

Definition
Explanation
Other texts

4. Conceptual relations

Number or name of the system of concepts to which the information in this category refers
Position of the term in the system of concepts
Broader term (generic relationship)
Broader term (part-whole relationship)
Broader term (no information on type of relationship available)

Narrower term (generic term)
Narrower term (part-whole relationship)
Narrower term (no information on type of relationship available)
Preceding term (in a successive relationship)
Succeeding term (in a successive relationship)
Other relations
Antonyms

5. Synonyms

Synonyms
Quasi-synonyms

6. Comments

7. General information

Date of first input
Bureau of origin
Copyright
Name of terminologist
Date of terminological preparation
Working method
Information on up-dating

figure 2: THE DANTERM DATA STRUCTURE