TKE '99

TERMINOLOGY
and
KNOWLEDGE ENGINEERING

edited by

Peter Sandrini

Proceedings

Fifth International Congress on

Terminology and Knowledge Engineering

23 - 27 August 1999
Innsbruck, Austria

organized by

Association for Terminology and Knowledge Transfer (GTW)
International Information Centre for Terminology (Infoterm)
International Network for Terminology (TermNet)
University of Innsbruck

TermNet

Vienna 1999

Øivin Andersen
University of Bergen, Norway

From Conception to Concept in a Terminological Data Base

1. Introduction

The Norwegian terminological data base NOT is the largest collection of terms in Norway. Terminology from various professional fields has been accumulated over the last 15 years. Today the base contains about 90 000 terms distributed in 30 000 term posts. Recently the base has been developed into a relational data base and the terminology structure has been enriched both with morphological information and argument structure information associated with deverbal nominal terms (cf Andersen 1998). Parts of the base in its redefined form is currently being incorporated into the Norwegian national computational lexical data base NorCompLex (Myking1998).

However, so far little has been done to develop the conceptual side of the base, i.e. to develop the structuring of the terminological concepts into systems, and to develop definitions of the terms.

Ten years ago a first effort was made to develop such ideas in conncetion with the third revised edition of the International Gas Union's dictionary of the gas industry. This edition was edited in 1994 containing between about 8000 entries in several different languages including Norwegian.

The terminology included not only traditional and new gas terminology, but also terminology related to administration and economy. Fields related to offshore drilling, gas exploitation and pipelines were considerably extended. The dictionary contains definitions in all cases where at least one of the languages has no equivalent vocabulary item. The definitions were taken in all languages from original sources, which implied that they differed from language to language.

The Norwegian Term Bank at the University of Bergen was responsible for providing Norwegian terms and definitions.

Technical definitions and other sources of documentation have of course currently been used in the process of forming motivated Norwegian terms in oil related technolects, but the task of actually forming practical definitions of terminological concepts was a new experience to the terminologists in Bergen.

Confronted with this task, the need to work out some guidelines for definition writing was immediatly felt. A set of guiding principles for conceptual construction as a basis of terminographical definition writing was established and tested. A few small-scale pilot projects were launched in those days in order to test the usefulness of the principles. But practical and financial problems prevented the terminologists from developing these ideas further. The revisions and extensions of the NOT base have once again made these efforts important. In this article I will present my ideas as to how this part of the job may be done.

Traditional terminological concept theory has had a tendency to view the terminological concept as a discrete static unit of thought. Many terminologists find this view very difficult to accomodate in their practical terminology work, basically because it seems contraintuitive.

More recent approaches focus on the status of the concept as a knowledge unit (Dahlberg) and as a dynamic epistemological unit evolving processually (Oeser 1988).

I will suggest a model of concept representation which has some resemblance to Budinet al's model (1988). As my point of departure, though, I will rather draw on Wierzbicka's (1985) distinction in lexicography between concept minimum and concept maximum. These terms reveal the difference between an individual's conception of a terminological concept and the traditional static concept.

From a cognitive point of view concept formation can be seen as a hypothetico-deductive device where knowledge accumulation, knowledge modification and knowledge classification are simultanious ongoing processes.

The basic question I will try to answer is: How can these processes be represented in a terminological data base in such a manner that the representations will meet the varied and changing needs of terminological practice?

2. Some basic assumptions

1. Following Dahlberg (1976) concepts are viewed primarily as elements of knowledge, and not significantly as elements of thought.

2. Following Wierzbicka, (1985) concept formation is best understood by adopting a scalar model of conceptuality. This type of model is compatible with the fact that there is no sharp line of demarcation between conceptual knowledge and world knowledge (or more generally between lexicon and encyclopaedia).

Moreover, as several LSP-studies have illustrated. the border lines between LSP and LGP are fuzzy. The same is also true of adjacent technolects of LSP. Even domains within the same technolect display much more heterogeneity than one would initially expect.

Wierzbicka distinguishes between concept minimum and concept maximum (1985: 214ff). Her notion of concept maximum is compatible with other well known concepts such as Rommetveit's meaning potential (Rommetveit 1972: 1977) and Fillmore's interpretive frame (Fillmore 1985). Wierzbicka's two meta-concepts are related to Putnam's notion of shared (collective) stereotypes (Putnam 1975) applied to LGP.

This model may also be adopted in an LSP context. Specialized communication will also include sets of shared stereotypes specific to a group of speakers, but the sets of stereotypes may vary even within one single technolect. As figure 1 in the appendix illustrates, the concept maximum is viewed as the total collection of knowledge which can be found on a specific concept. This collection may be seen as a set of possible features for the concept in question.

The concept minimum. on the other hand, constitutes the minimum of knowledge a person must have in order to be able to apply it in a specialist context. If a person's knowledge is lower than this lower limit, that person cannot be said to "know" the concept. Knowledge of a concept, including its world correlates, is then a gradable phenomenon: It is something which one possesses to a larger or a smaller degree.

As a consequence, it may vary from person to person. Of course, a special concept may be said to be a discrete unit for a special group of people based on salience. which at rock bottom is controlled by special interests and special needs. But the terminologist as a "concept discoverer" who is a non-expert in the field, will not have a map of the knowledge fields with their borders. His/her task is to discover the fields by the help of initiated authoritative experts in the field. So the model in figure 1 is a practical working tool for the terminologist rather than a metaphysical statement about the status of concepts.

3. Concept formation is a cumulative process. As Dahlberg (1976: 105) points out, all kinds of practical definition work have to be based on what has been done before in the field of investigation. A tentative division of knowledge areas may be established as the concept formation work is in progress, since both these processes are mutually dependent.

As stated above, concepts are regarded as elements of knowledge. Since these knowledge elements must in some way be glued to the world of experience, they should be formulated in hypothetical form as synthetic propositions representing some possible world, i.e. a world of possible experience (Popper 1980: 39). Popper's criteria of intersubjective testability and refutation are regarded as crucial.

Following Dahlberg, then, conceptual formation is seen as collecting and synthesizing possible propositions which (following Popper) must be intersubjectively testable and refutable

4. The process of concept construction is not regarded as decomposition but as synthesis. Decomposition presupposes the general existence and availiability of a discrete concept or system of concepts. To insist on a conceptual view based on discreteness would in many cases result in a hopeless atomization: It would yield an enormous number of possible worlds (or domains) with the same number of conceptual systems. These systems would be likely to change rapidly as new insight was gained and concepts would be discarded as fruitless and substituted by new ones.

5. Knowledge elements of concepts are types of semantic features. Dahlberg (1976: 90) makes a distinction between essential and accidental features. The essential features are interpreted as necessary, whereas accidental features are said to be "additional" features.

One of the problems with such a distinction is that the terminologists do not know in advance which features are which. Moreover, the set of essential features will vary from context to context (or from project to project).

This also illustrates the fact that the typical unambiguous character of terminological concepts is achieved by imposing restrictions on the context: Heavy restrictions result in a high degree of unambiguousness and vice versa: The burden of information is simply shifted from context to concept. The features constituting such concepts will be relevant in some contexts and irrelevant in other contexts.

Hence we draw a distinction between possible and relevant features; a distinction which is inspired by Wittgenstein's concept of family resemblance (Wittgenstein 1953). The physiognomies of a family have a distinct family resemblance, but if you take each feature individually (as the design of the nose, cheek, chin, eye colour etc.), you will be likely to find that the nose feature, for instance, is common to person A and B, the characteristic design of the cheek is common to B and C, the design of the chin is common to C and D, and eye colour common to D and E. and so on.

There is no single feature, seen in isolation, recurring in every family member. So it is the combinations of these features which are characteristic of the family. These combinations are necessary to identify the family, not the individual features themselves. This represents a very important aspect of modern terminology structure (cf Weissenhofer 1995).

3. The proposed principles

The set of context dependent conceptual features for different purposes may be thought of (in our project) as similar types of Wittgensteinian combinations. and are referred to as relevant features.

These combinations will form the basis for definition writing of the various oil related Norwegian terminological concepts in the NOT base.

Figure 2 (in the appendix) illustrates the proposed set-up for feature recording in the definition field of the Norwegian Term Bank record. The list of possible features is open-ended, and new features may be added. The list will contain all the features which the terminologists are able to collect from various sources, and which may possibly contribute to the concept.

The relevant features for different domains will be different subsets of the possible feature list. In figure 2 the difference between the subsets is exaggerated for the purpose of illustration: Whatever is relevant to project 1 may be relevant or irrelevant to project 2.

An advantage with this set-up is that, as the possible feature list accumulates. it will steadily grow more project independent.The main objective is to approach the concept maximum on the knowledge scale in figure 1 as much as possible.

The point of departure is very often a term x. Our first question will be: What do we know about x already? In other words: Which position do we occupy on the knowledge scale of figure 1?

The most common starting point is either at or somewhere below the threshold of concept minimum. The small amount of information we do have will occupy the fl slot in figure 2 (if the source of information is acceptable). Then we consult all our available soures and persons (specialists, all sorts of encyclopedias, dictionaries, technical handbooks and other kinds of specialized documentation in the field) in order to establish a list of possible features.

Two principles are regarded as crucial in this process: The principle of accumulation and the principle of disjunctivity. The principle of disjunctivity is crucial to the elicitation of unambiguous information, and is often regarded as typical of conceptual taxonomies, as Miller & Johnson-Laird (1976) and later Jackendoff (1985) point out.

The principle of accumulation means that source 2 will be registered only if it gives additional information beyond source 1. The principle of disjunctivity means, in the ideal cases, that no features must overlap, i.e. contain information listed in other possible features on the list. This is illustrated in figure 3b, where the information for source 2 is B. but where f2 is the set- theoretical difference between the set B and the set A. If the information in source 2 is included in the information in source 1, as figure 4 illustrates, source 2 will not be registered. Registration of f3 will then follow from the principles of accumulation and disjunctivity, as shown in figure 5.

In this manner we hope to be able to establish a list of possible features, feature by feature. The possible feature list will constitute the basis for the further

conceptual construction to be performed in committies consisting of terminologists and initiated experts. The committe will have three main tasks to perform:

a. Evaluation of sources.

At this stage the terminologists will have to identify the various sources of information and evaluate them in cooperation with experts in the field. Is the source an authoritative one in the field to be investigated? Is the source of information frequently consulted by the experts? Is it easily available in their daily work? Is it easily available on the market? Is the source copied or literally translated from another source? What is the authority of the author or the originator of the source? Does he/she apply a deviant or a commonly accepted terminology? (cf Lading 1979). These questions are often difficult and time consuming. but of fundamental importance.

b. Completion of the possible feature list

The list must be as complete as possible in order to reduce the possibility that features which may be relevant to a current project are absent. This task is primarily the responsibility of the experts in the current project.

c. Identification of relevant features.

In the possible feature list there will most probably be features which are relevant to the concept maximum, but which are irrelevant in a specific project related to for instance petroleum production, medicine or law, and others which are relevant. Definitions will then be formed for the different projects on the basis of the sets of relevant features (cf figure 6).

4. The Pilot Projects

The two pilot projects were performed by two terminologists at the Norwegian Term Bank. They investigated two off-shore terms relating to technical equipment which is considered to be crucial in both oil and gas production: kelly and flange. Moreover a central administrative term operating company was treated. The list of applied sources ranges from general dictionaries with some technical terms, like Duden 10.Bedeutungswörterbuch and The Heritage Illustrated Dictionary of the English Language to technical dictionaries like McGraw-Hill Concise Encyclopedia of Science and Technology, and highly specialized dictionaries like Composite Catalogue of Oil Field Equipment & Services.

5. Conclusions of the pilot projects

The two pilot projects made it quite clear that further small-scale test projects were needed before a large-scale project could be embarked upon. But the preliminary conclusions to be drawn from the projects could be summarized in 6 points:

1. The sources are in some cases contradictory. The principles of accumulation and disjunctivity presupposes that all collected information is logically and factually consistent. This would function in an idealized situation like the Kuhnian "normal science" (Kuhn 1979). Of course, we would like our sciences to be like that, but science almost never is to 100%. Instead we often get a source situation bristling with inconsistencies and contradicions: A kelly, for instance cannot significantly have a hexagonal shape and be 3, 4, 6 or 8 sided at the same time. It would seem that the principle of strict disjunctivity is exceedingly hard to adhere to, partly because the projects are not always unambiguously defined. Consequently, our principles must be regarded as idealizations; a goal which should be reached for, although we realize that in many cases it is impossible to achieve.

2. The types of information conveyed by the sources are so heterogeneous that a crude semantic classification of feature types are required in order to control the information.

3. A (set of) standardized way(s) of formulating the features in the possible feature list should be established. They ought preferably to be formulated as empirical hypotheses about the concept to be evaluated by initiated experts as to both truth value and relevance.

4. Based on their previous knowledge of various fields of professional terminology, the terminologists should, to some extent at least, be able to decide (or at least be able to have a justified opinion about) whether two (or more) formulations are equivalent or not. In the event of non equivalence they must try to establish whether the formulations are logically compatible or logically contradictory.

5. The applied method of concept formation turned out to be far more time consuming than expected, although the pilot projects only covered the first stage of the method: the establishment of a possible feature list.

6. Although the second stage of the method has not been tested yet, it seems obvious that further problems will have to be faced in connection with the elicitation of relevant features. Some of these problems may eventually have to be solved by the initiated experts themselves. But it will require training and skill in conceptual analysis. As Rommetveit puts it (cited by Linell 1989) word meanings are conceived of as potentials and as "drafts of contracts concerning

categorization and attribution", i.e. as some kind of procedural knowledge abstracted from across instances and with an open and indeterminate, multiple determinable character (op . cit. 1989).

A scalar concept model would be required to depict such a view.

After all, our methods do assume some kind of epistemologically transparent technical oil world "out there". but what we very often face is a multifaceted, only partially shared and only fragmentarily known world.

Bibliography

ANDERSEN, Ø. (1998): Argumentstruktur og nominalisering. In: LexicoNordica 5.

BUDIN, G./ GALINSKI, CH./ NEDOBITY, W./ TIIALLER, R. (1988): Terminology and Knowledge Data Processing. In CZAP, H./GALINSKI, C. (eds): Terminology and Knowledge Engineering. Supplement. Frankfurt a. M.

DAHLBERG, I. (1976): Über Gegenstände, Begriffe. Definitionen und Benennungen. In: Muttersprache 2.

FILLMORE C. J. (1985): Frames and the semantics of understanding. In: Quaderni di Semantica. vol. VI no. 2.

JACKENDOFF, R. (1985): Semantics and Cognition. MIT-press. Cambridge.

KUHN, T. (1970): The Structure of Scientific Revolutions. London.

LADING, B. (1979): Et terminologisk-fraseologisk materiale vedrørende pantsætning i dansk og engelsk ret. Department of English. The Copenhagen School of Economics.

LINELL, P. (1989): The embeddedness of decontextualization in the contexts of social Practices. In: Heen Wodl, A. (ed.): Festschrift für Ragnar Rommetveit.

MILLER, G.A./JOHNSON-LAIRD, P. N. (1976): Language and Perception. Cambridge.

MYKING, J. (1998): Norsk terminologisk database - NOT. In: LexicoNordica 5.

OESER, E. (1988): Terminologie als Voraussetzung der Wissenstechnik. In: I.H. CZAP & CH. GALINSKI (eds.): Terminology and Knowledge Engineering. Supplement. Frankfurt a. M.

PICHT, H. (1987: Fachsprachliche Phraseologie. Die terminologische Funktion von Verben. In: Terminologie et traduction, No 3. Commission des Communautés Européennes.

PILKE, N. (1997): Dynamiska begrepp. I: Terminologi som vetenskapsgren. Studentlitteratur. Lund.

POPPER, R.(1980): The Logic of Scientific Discovery. London.

PUTNAM, H. (1975): Mind, Language and Reality; Philosophical Papers. vol 2. Cambridge.

ROMMETVEIT R. (1972) : Språk, tanke og kommunikasjon. Oslo.

ROMMETVEIT, R.(1977): On Message Structure. London.

WEISSENHOFERP. (1995): Conceptology in Terminology Theory. Semantics and Word-formation. Vienna.

WIERZBICKA, A. (1985): Lexicography and Conceptual Analysis. Ann Arbor.

WITTGENSTEIN, L. Philosophical Investigations. New York. 1953.