# EMPTY CATEGORIES IN DOCUMENTARY BASED THESAURUS CONSTRUCTIONS

ØIVIN ANDERSEN

UNIVERSITY OF BERGEN,

DEPARTMENT OF SCANDINAVIAN LANGUAGES AND LITERATURE,

SECTION OF NORWEGIAN LEXICOLOGY,

STRØMGATEN 53,

N-5000 BERGEN,

NORWAY.

| Surname | First name |
|---|---|
| Title | University |
| Mailing address | |
| Postal code, City State, Zip | Country Tel no |
| Co-author | |
| Research interest | |


Abstract

Empty categories in documentary based Thesaurus constructions

A descriptor with no bibliographical reference in a documentary based thesaurus is here called an empty category (EC). ECs are usually inserted in a thesaurus in order to establish useful links between descriptors in the thesaurus hierarchy, primarily for ease of retrieval. This paper discusses some aspects of the use of ECs in this field. Comparisons are drawn to ECs in syntax. The empirical justification of these categories constitutes a problem. How can we give principles to restrict their use?

A candidate for empty thesaurus categories may be found in E. Rosch's basic level categories. A proposal as to how basic categories might be justified as ECs will be introduced as well as suggestions as to how they can be tested as regards efficiency in retrieval.


PTO

Prof. Dr. Christer Laurén,
LSP Symposium 1987,
School of Modern Languages,
University of Vaasa,
Raastuvankatu 31,
SF-65100 Vaasa,
Finland

Øivin Andersen,
University of Bergen,
Departement of Scandinavian
Languages and Literature,
Section of Norwegian Lexi-
cology,
Strømgaten 53,
N-5000 Bergen,
Norway.

Telephone number: 212950.

General:

1. Language : This paper will be read in English.
2. Topic of the paper : Empty categories in documentary based Thesaurus constructions.
3. Subsection : 4. Terminology and Lexicography

A descriptor with no bibliographical reference in a documentary based thesaurus is here called an empty category (EC). ECs are usually inserted in a thesaurus as auxiliary concepts in order to establish useful thematic links between different vertical levels of descriptors in the thesaurus hierachy, primarily for ease of retrieval, but also to increase the completeness of the hierarchies. An EC is characterized by lack of thesaurus reference, i.e. it cannot be traced back to any of the indexed documents of the thesaurus. They are not indexing words, like the descriptors, Its thesaurus bibliographical reference is thus empty, and their function in a thesaurus hierarchy is of a mediating character between descriptors.

The use of such categories represents some problems:
Firstly, the efficiency of search is may be reduced because a direct bibliographical match is impossible.

Secondly, we face the problem of restricting the use of such categories: Where should they be inserted and where should they be avoided?

The first problem is related to the difference between conceptual classification (e.g. taxonomies) and thematic classification (e.g. thesaurus classification). The difference between these two types of classification will be discussed in order to clarify the context of this problem.

As to the second problem, the problem of demarcation and justification of empty categories is well known in recent transformational syntax. These syntactic problems will be related to the ones mentioned above, which are linked to thesaurus constructions for comparison.

In a thesaurus context the question is essentially the same as

in syntax: Can we imagine cases which could justify the use of empty categories like ECs? In which way can they be said to be useful?

A candidate for ECs may be found in E. Rosch's basic level categories. These categories have been studied intensively in modern cognitive psychology in connection with taxonomies. Rosch claims that the perceived world posess a high correlational structure and that certain attributes (semantic features) are strongly assosiated, i.e. chairs are assosiated with sit-on-able-ness and birds are assosiated with wings etc. Not all vertical levels in a taxonomy are equally useful in categorization. The most basic level of categorization is the level at which categories best mirror the structure of the attributes perceived in the world. At the horizontal level in the taxonomy the distinctiveness (i.e. disjoint character) of the categories are maximized at this basic vertical level.

If this theory is valid: is it possible that this level should be represented in a thesaurus, in case the thesaurus contains hierarchies with this level of inclusiveness, even if they are ECs? Is there a tendency among the on-line useres of a thesaurus to use basic level categories, like chair, bird etc. as points of departure, or entry words, in a document retrieval search in a thesaurus?

In this article I will describe the use of so called empty categories in documentary based thesauri seen in the light of the crucial distinction between thematic and conceptual classification systems as described among others in Wüster (1985). Examples will be drawn from the monolingual Norwegian Petroleum Thesaurus Petrus, which was revised and restructured by the Norwegian Term Bank at the University of Bergen in 1986. The total number of terms in Petrus is now 6567, of which 5787 have the status of descriptors. The thesaurus consists of ten large hierarchies based (with some modifications) on the American Exploration and Production Thesaurus, University of Tulsa, Oklahoma.

The term empty category (EC) is borrowed from transformational syntax. ECs in syntax are nominal phrases without lexical material. They are postulated in the surface sentence structures in order to establish a transformational path from underlying to surface structure. The underlying structure is thus postulated in order to account for certain syntactic properties that sentences and elements are claimed to have (such as subcategorisation, morphological case marking, concord, selection restrictions, phonological contraction and the syntactic behaviour of reflexive pronouns. A comprehensive description of this is given in Radford (1981) ).

The basic point here is that ECs cannot be postulated ad hoc, but must meet certain correlated empirical principles.

The nature and function of ECs in thesaurus constructions may be seen in the light of the distinction between thematic and conceptual classification systems mentioned. Conceptual and thematic classifications may look similar at the surface, but the basis of classification is quite different. A basic distinction can be made between direct and indirect reference. The concepts of a conceptual classification refer directly to extensional classes in the non-linguistic world (i.e. real world objects, states and events), whereas concepts of a thematic classification (including descriptors of documentary based thesauri) refer to classes of thematic units of a set of documents, i.e. to linguistic entities, which I propose to call thesaurus reference. The thematic units, in turn, refer to real world objects, states and events.

Thus, conceptual hierarchies are structured according to man's view of the external world directly, whereas thematic hierarchies represent an analysis of a set of authors' view of the world. Aspects such as thematic weighting and degree of thematic centrality of concepts in documents become crucial in thematic classifications, and irrelevant in conceptual classifications. The principle aspect in thematic classifications is not semantic relatedness, but thematic connections between concepts. One of the consequences of this difference is that conceptual classifications tend to be more comprehensive and complete than thematic ones because some concepts are not treated explicitly (or only mentioned shortly) in documents.

ECs have no thesaurus reference, i.e. they are non-indexed descriptors. This means that they cannot be traced back to any of the indexed documents of the thesaurus. They can only be motivated on the basis of a specified underlying conceptual system, like ECs in syntax are based on a specified underlying sentence structure.

Petrus contains 56 ECs (also called auxiliary concepts). They were introduced by the librarians as mediators between various levels in the hierarchies in places where they were considered desireable in order to obtain more complete and logically structured hierarchies.

Figure 1 (in the appendix) gives an example of a conceptual classification dealing with geological time and the corresponding thematic concepts of Petrus. The illustration represents a unification of the two systems where categories occuring in both classifications occur in common types, categories occuring exclusively in the conceptual classification are given in round brackets, categories occuring exclusively in Petrus are given in bold types, and ECs are underlined. The same system is represented as partially overlapping classes in figure 1a.

At a closer look we see that sen kritt (Upper Cretaceous) is an empty category mediating between maastricht (Maestrichian) and cenoman (Cenomanian) on the one hand and kritt (Cretaceous) on the other hand. But we can also read from figure 1 that ECs are inserted as terminal categories (i.e. categories belonging to the lowest level of thematic subcategorisation, as exemplified by sen paleocen) (Upper Paleocene).

One of the reasons for looking into ECs in Petrus was to see if it was possible to derive some principles of their distribution and their retrieval function in the hierarchies. Obviously, the basic idea seems to be that the user groups of a thesaurus posess an underlying mental conceptual system to which he relates the thesaurus hierarchies in the search process, i.e. a kind of unification or mapping the two systems. But this unification cannot be ad hoc: it must be subject to some kind of systematic logic. This means that ECs cannot be postulated ad hoc, or everywhere. If we do, the consequences will be like in figure 2. Figure 2 illustrates a classification of various pumps subcategorized according to liquid (væske), form and function (funksjon). Again, the common type categories represent the set of concepts appearing in both systems, whereas the round bracket categories are absent in Petrus. If we incorporate all the round bracket categories in Petrus as ECs in order to obtain a more complete system, we certainly complicate and impede document retrieval. Obviously, this is not what we want.

Figures 3 and 4 give a set theoretical survey of the relationship between the class of descriptors and non-descriptors, and, in relation to this, the presence of LSP and LGP categories in Petrus. The variables x (LSP - categories) and y (LGP - categories) of propositions 5 and 6 of figure 4represent the set of the potential candidates for ECs in Petrus. This framework enables us to ask the question: Which of the x's and y's (if any) may qualify as ECs in Petrus?

There are three possible solutions to this question:

1. There is no distinction between thematic and conceptual classification. ECs are allowed (but marked as such) in all x and y positions in 5 and 6 of figure 4.
2. Pure thematic classification. No ECs are allowed.
3. ECs are allowed in some x and y positions, but not in others.

Solution 1 gives a good survey of the underlying conceptual system, and may be both interesting and useful to some users. But it complicates retrieval to a very large extent and increases the possibility of mismatches in documentary search. Moreover, the number of concepts of the lexicon and the number of concepts of technolects of any language are, if not infinite, indeterminably large. In solution 1 there is thus no way to restrict the use of ECs, as the pump examples in figure 2 illustrated.

Solution 2 is the common solution in many documentary based thesauri. It is the best solution in cases where the discrepancy between the conceptual and thematic structures is not too large.

Solution 3 may be a good solution if the conceptual hierarchies are very incomplete (e.g. if several levels in the vertical ladder is missing, and if these levels are important in retrieval.

Let us examine the consequences of solution 3. If we choose this solution, we have to answer the question above (Which potential EC positions may be permitted in propositions 5 and 6 of figure 4?). A possible answer to this demarcation problem may be found by asking which terms the enquirer is liable to use when beginning his search in an information retrieval system, i. e. his approach terms (Buchanan 1976). At this point research and insight from modern cognitive psychology may help us to find possible answers.

There are at least two theories dominating modern decompositional conceptual classification: the classical definitorial view and the modern prototype or cluster concept view. The definitorial view is the more prevalent of the two in concept theory of terminology. It is decompositional in the sense that the intension of a concept is seen as the aggregate of all the characteristics which constitute it (cf. DS/ISO R/1087).

The finite set of characteristics are both necessary and sufficient to pick out all, and only, the referents which are to be subsumed under the concept (i. e. its extension) and no referents which are not to be subsumed under the concept. The characteristics are supposed to be able to pick out for instance the set of dogs from everything else in the world. In other words the set of characteristics determines category membership, and all members (or referents) of a concept have equal status as to membership.

The prototype view dates back to Wittgenstein's Philosophical-Investigations (1953) and his notion of family resemblance. Rather than postulating a well defined set of category members (the extension of the concept) with equal status to their set of characteristics (the intension of the concept), conceptual membership in prototype theory becomes a question of degree.

A family resemblance relationship can be visualized as in figure 5. It consists of a set of items of the form AB, BC and CD. Each referent has at least one (very often several) characteristics in common with one or more other referents of a concept. But very few characteristics common to all referents of the same concept. Thus, a chair for instance, is said to be a better or more prototypical referent of the concept furniture than vase, and vase is more prototypical than ashtray. Generally speaking, the more prototypical a referent is of a concept, the more characteristics it has in common with other members of that category, and the less it has characteristics in common with contrasting categories.

The concepts of family resemblance and prototypicality have been studied rather intensively by E. Rosch and her colleagues (cf. the reference list). Logic or generic hierarchies (taxonomies) are based on the principle of extensional inclusion in the sense that the extension of broader concepts include the extension of narrower concepts which they dominate.

One of the basic ideas of the prototype theory is that there exists a consistent basic hierarchic level of inclusion where referents of a concept have an optimal degree of family resemblance to one another and a minimal degree of family resemblance to referents of other concepts. In figure 6 this basic level is shown for 4 of 9 taxonomies of concrete concepts. In a series of experiments (cf. Rosch et. al 1976) it is shown that basic categories are more imagelike than other categories (i.e. we can form a mental picture of a prototypical chair, a basic level category, but not so easily a picture of a prototypical piece of furniture. In identifying objects at a level subordinate to (or more specific than) basic level concepts, like a Chippendale chair, they are first and immediatly recognized as basic level categories (i.e. a chair).

The important aspect here is that the basic level seems to be the most useful level of reference in identification of more specific or subordinate level objects.

Moreover, basic level concepts seem to be the first ones learned by the child (e.g. guitar and drum in figure 6 are learned long before musical instrument and folk guitar and kettle drum). (For further details, methods applied and experimental results cf. Rosch et. al 1976).

All this suggests that concepts subordinate to the basic level of classification are too specific and requires more expert knowledge on the part of the user of a thesaurus. When for instance trying to identify a mechanical product with several specific characteristics, we may ask ourselves: What is the most useful and common level of reference of this mechanical product? Perhaps we do not understand the details, but we know that it is most probably a pump. In a retrieval context, then, pump should perhaps be represented as an EC if it is not indexed from the documents. Or, if a non-expert, average user of a thesaurus were given the task of finding documentary information on Chippendale chairs: Would his approach term be more likely to be chair than furniture?

To take a final example: In figure 7a we have a pure thematic classification of vehicles, whereas 7b illustrates the same classification with the empty category car. Most people know that Volvos and Nissans are cars. So, car would be a very likely candidate for the approach term here. In 7a we cannot use car as an approach term (unless it appears in the alphabetical part as a non-preferred term). By giving it status as a non-indexed descriptor we are lead directly to the relevant hierarchy.

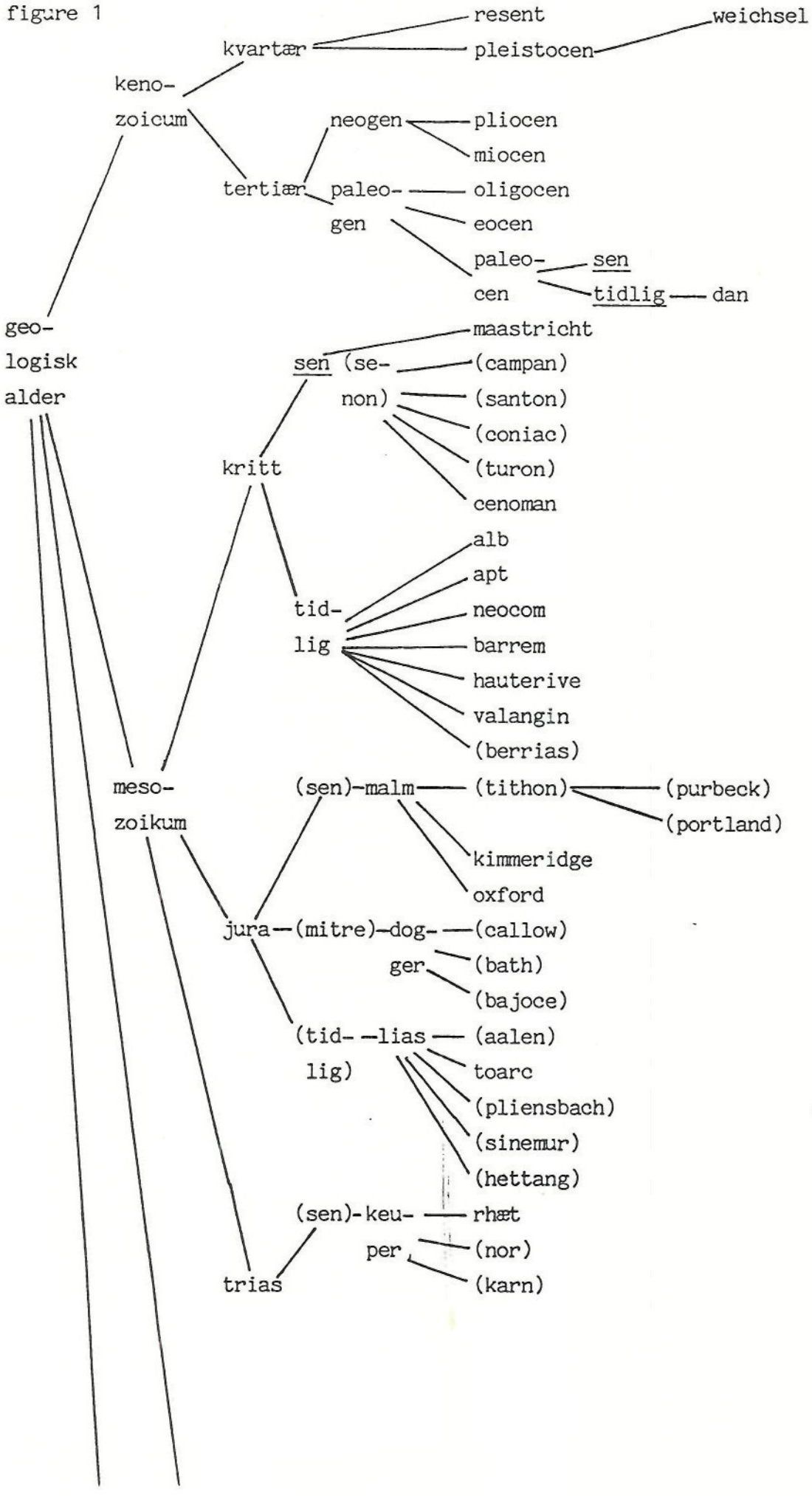
From this it is, of course, not possible to conclude that basic level categories are useful as ECs in a thesaurus. However, numerous experiments in cognitive psychology suggest that there is a consistent correlation between basic level concepts on the one hand and ease of identification, frequency of reference and salience on the other hand.

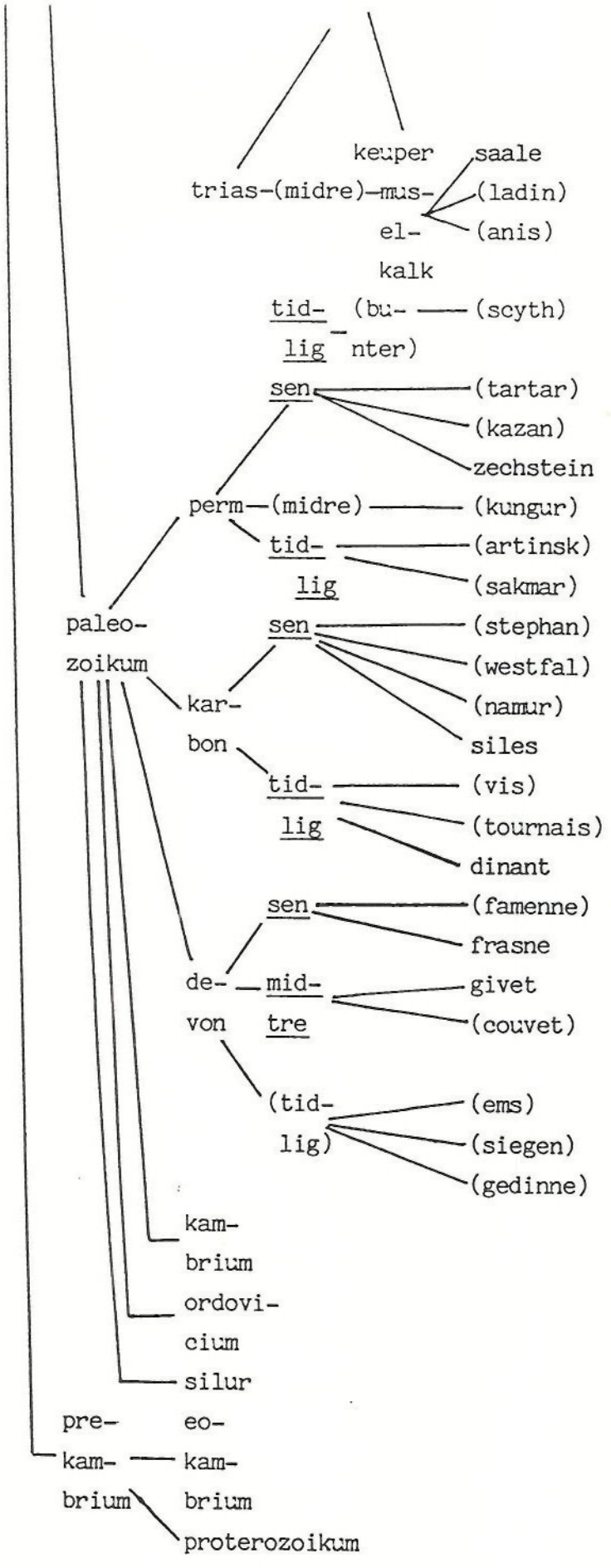
The question in our context is: Is there also a consistent correlation between basic level concepts on the one hand and storing, retrieval, category searching and the use of approach terms in thesauri on the other hand?

It is very reasonable to assume that man, as a classifying being, actively assigns characteristics to concepts in order to achieve the maximum degree of contrast between them in order to be able control his environment. The basic level categories seem to mirror this contrast more clearly than other levels of inclusion.

# APPENDIX

figure 1


geo-
logisk
alder
keno-
zoicum
kvartær
resent
pleistocen
weichsel
tertiær
neogen
pliocen
miocen
paleo-
gen
oligocen
eocen
paleo-
cen
sen
tidlig
dan
meso-
zoikum
sen (se-
non)
maastricht
(campan)
(santon)
(coniac)
(turon)
cenoman
kritt
tid-
lig
alb
apt
neocom
barrem
hauterive
valangin
(berrias)
jura
(sen)-malm
(tithon)
(purbeck)
(portland)
kimmeridge
oxford
(mitre)-dog-
ger
(callow)
(bath)
(bajoce)
(tid-
lig)
-lias
(aalen)
toarc
(pliensbach)
(sinemur)
(hettang)
trias
(sen)-keu-
per
rhæt
(nor)
(karn)


keuper
saale
trias-(midre)-mus-
(ladin)
el-
(anis)
kalk
tid-
(bu-
(scyth)
lig
nter)
sen
(tartar)
(kazan)
zechstein
perm-(midre)
(kungur)
tid-
(artinsk)
lig
(sakmar)
paleo-
sen
(stephan)
zoikum
(westfal)
kar-
(namur)
bon
siles
tid-
(vis)
lig
(tournais)
dinant
sen
(famenne)
frasne
de-
mid-
givet
von
tre
(couvet)
(tid-
(ems)
lig)
(siegen)
(gedinne)
kam-
brium
ordovi-
cium
silur
pre-
eo-
kam-
kam-
brium
brium
proterozoikum

common types : + P + RTT
round brackets : - P + RTT
bold types : + P - RTT
underlined types: EC

P = Petrus - The Norwegian Petroleum Thesaurus
RTT = Rådet for teknisk terminologi (Norwegian Council for Technical Terminology)
EC = empty category
+ P = The category is included as a descriptor in Petrus
+ RTT = The category is included in the conceptual classification of the Norwegian Council for Technical terminology
- P = The category is not included as a descriptor in Petrus
- RTT = The category is not included in the conceptual classification of the Norwegian Council for Technical Terminology

figure 1a SET THEORETICAL RELATIONS BETWEEN PETRUS CATEGORIES AND RTT - CLASSIFICATION SYSTEM

| B | | | A - B | A |
|---|---|---|---|---|
| weichsel | kenozoikum | oxford | senon | westfal |
| neocom | kvartær | toarc | campan | hamur |
| saale | resent | rhæt | coniac | vis |
| zechstein | pleistocen | trias | turon | tournais |
| siles | tertiær | tidlig- | berrias | tidlig devon |
| dinant | neogen | trias | sen jura | tamenne |
| | paleogen | perm | midtre jura | couvet |
| | pliocen | sen- | tidlig jura | ems |
| B - A | miocen | perm | tithon | siegen |
| | oligocen | tidlig- | purbeck | gedinne |
| | eocen | perm | portland | |
| | paleocen | devon | callow | |
| | sen- | sen- | bath | |
| | paleocen | devon | bajoce | |
| | tidlig- | midtre- | aalen | |
| | paleocen | devon | pliensbach | |
| | kritt | frasne | sinemur | |
| | sen- | givet | hettang | |
| | kritt | pre- | nor | |
| | maastricht | kambrium | karn | |
| | cenoman | kambrium | sen trias | |
| | tidlig- | ordovi- | midtre trias | |
| | kritt | cium | bunter | |
| | alb | silur | ladin | A = RTT conceptual classification |
| | apt | proter- | anis | |
| | barrem | ozoikum | scyth | B = Petrus thematic classification |
| | hauterive | | tartar | |
| | valangin | | kazan | A - B = Categories not included in Petrus |
| | mesozoikum | | midtre perm | |
| | jura | | kungur | |
| | malm | | artinsk | B - A = Categories not included in RTT |
| | dogger | | sakmar | |

figure 2

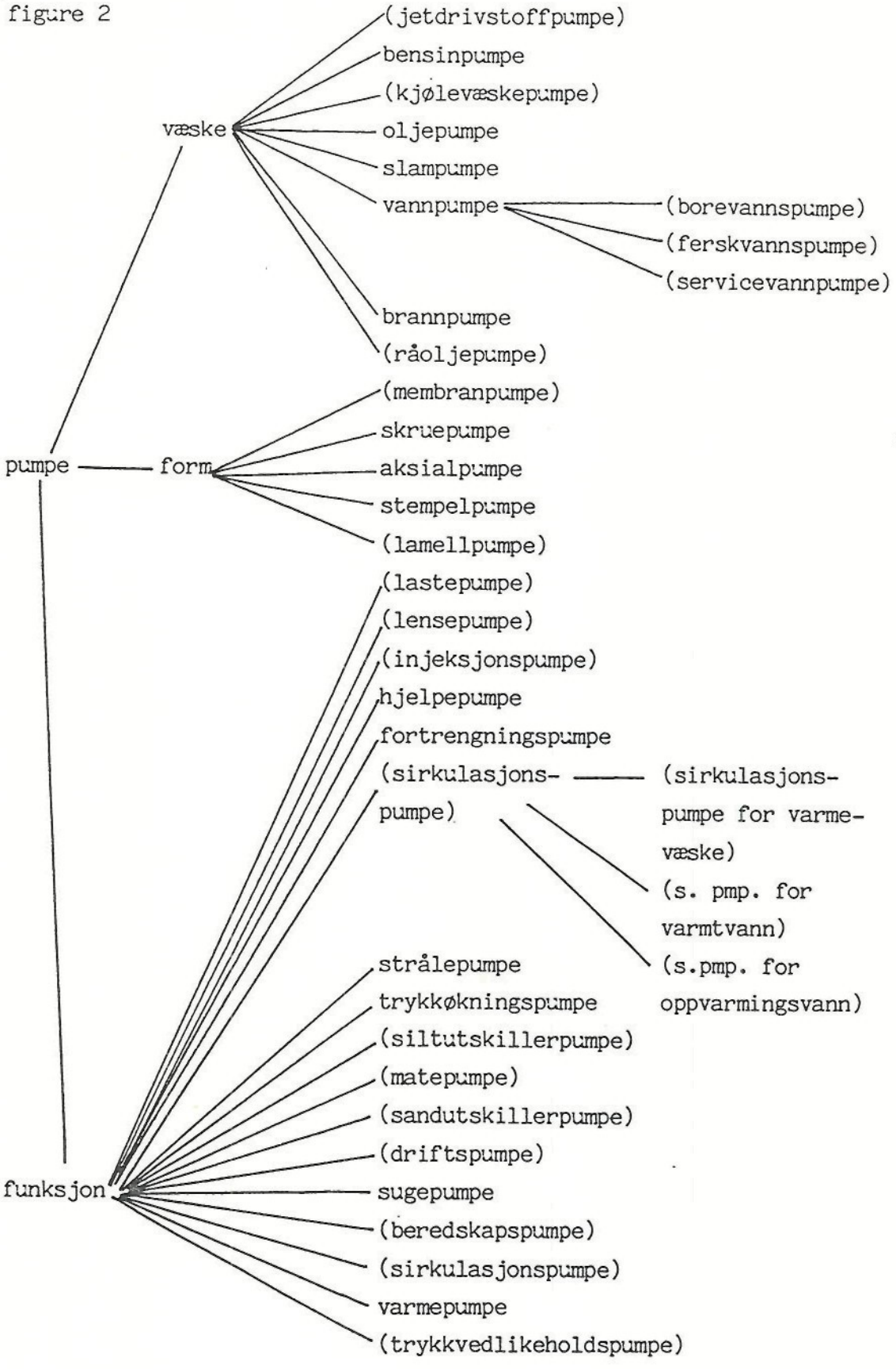
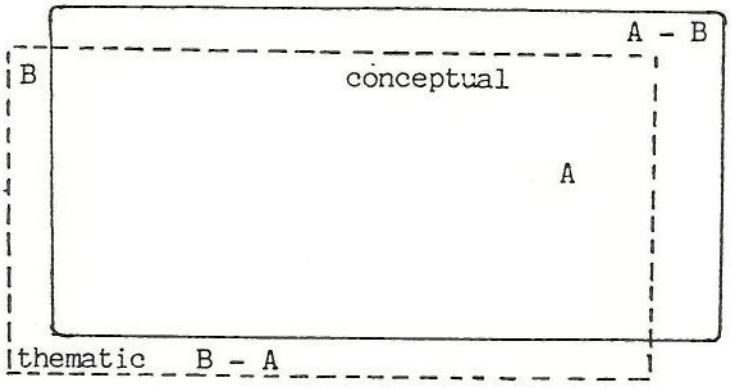

pumpe
væske
(jetdrivstoffpumpe)
bensinpumpe
(kjølevæskepumpe)
oljepumpe
slampumpe
vannpumpe
(borevannspumpe)
(ferskvannspumpe)
(servicevannpumpe)
brannpumpe
(råoljepumpe)
form
(membranpumpe)
skruepumpe
aksialpumpe
stempelpumpe
(lamellpumpe)
funksjon
(lastepumpe)
(lensepumpe)
(injeksjonspumpe)
hjelpepumpe
fortrengningspumpe
(sirkulasjons-pumpe)
(sirkulasjons-pumpe for varme-væske)
(s. pmp. for varmtvann)
(s.pmp. for oppvarmingsvann)
strålepumpe
trykkøkningspumpe
(siltutskillerpumpe)
(matepumpe)
(sandutskillerpumpe)
(driftspumpe)
sugepumpe
(beredskapspumpe)
(sirkulasjonspumpe)
varmepumpe
(trykkvedlikeholdspumpe)

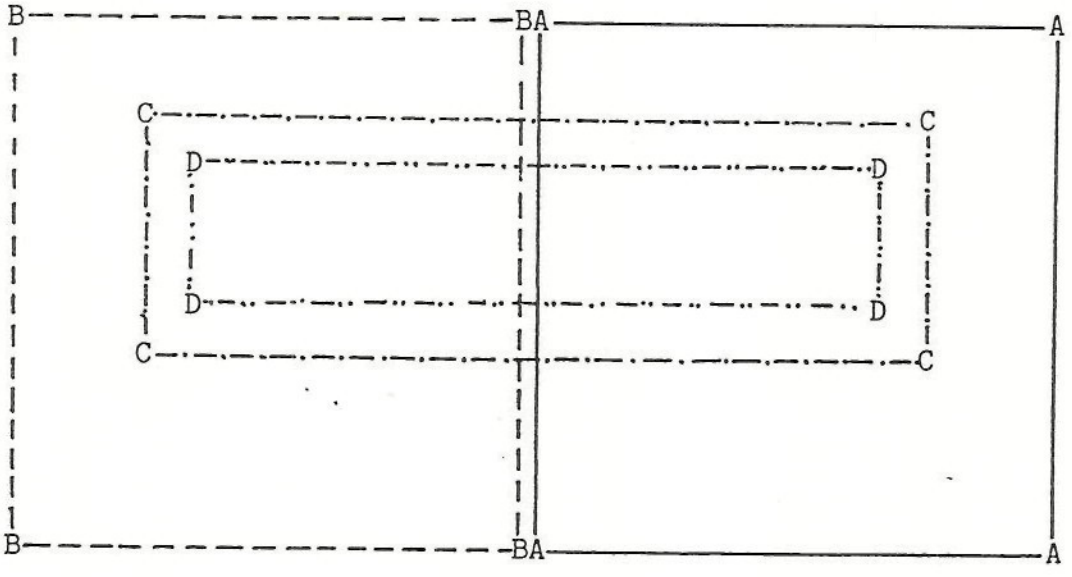
figure 3




A = Conceptual classification

B = Thematic classification

figure 4




A = LGP - categories (the total set of the Norwegian lexicon)

B = LSP - categories (the total set of the Norwegian conceptual systems in all technolects)

C = The total set of entry terms in Petrus

D = The total set of descriptors in Petrus

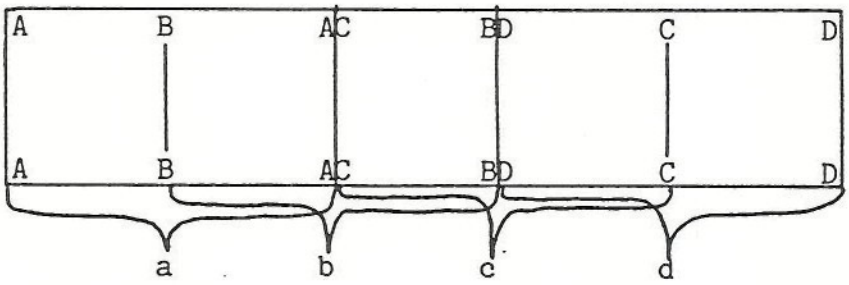
Propositions

| | | |
|---|---|---|
| 1. | $A \cap B = \emptyset$ | The intersection between the set A and the set B is empty (disjunct sets) |
| 2. | $D \subset C$ | The set D is properly included in the set C |
| 3. | $A \cap C \neq \emptyset$ | The intersection of the set A and the set C is not empty |
| 4. | $B \cap C \neq \emptyset$ | The intersection of the set B and the set C is not empty |
| 5. | $\exists x \in B \ \& \notin D$ | There exists a category x which is a member of the set B and not a member of the set D |
| 6. | $\exists y \in A \ \& \notin D$ | There exists a category y which is a member of the set A and not a member of the set D |

Three possible solutions:

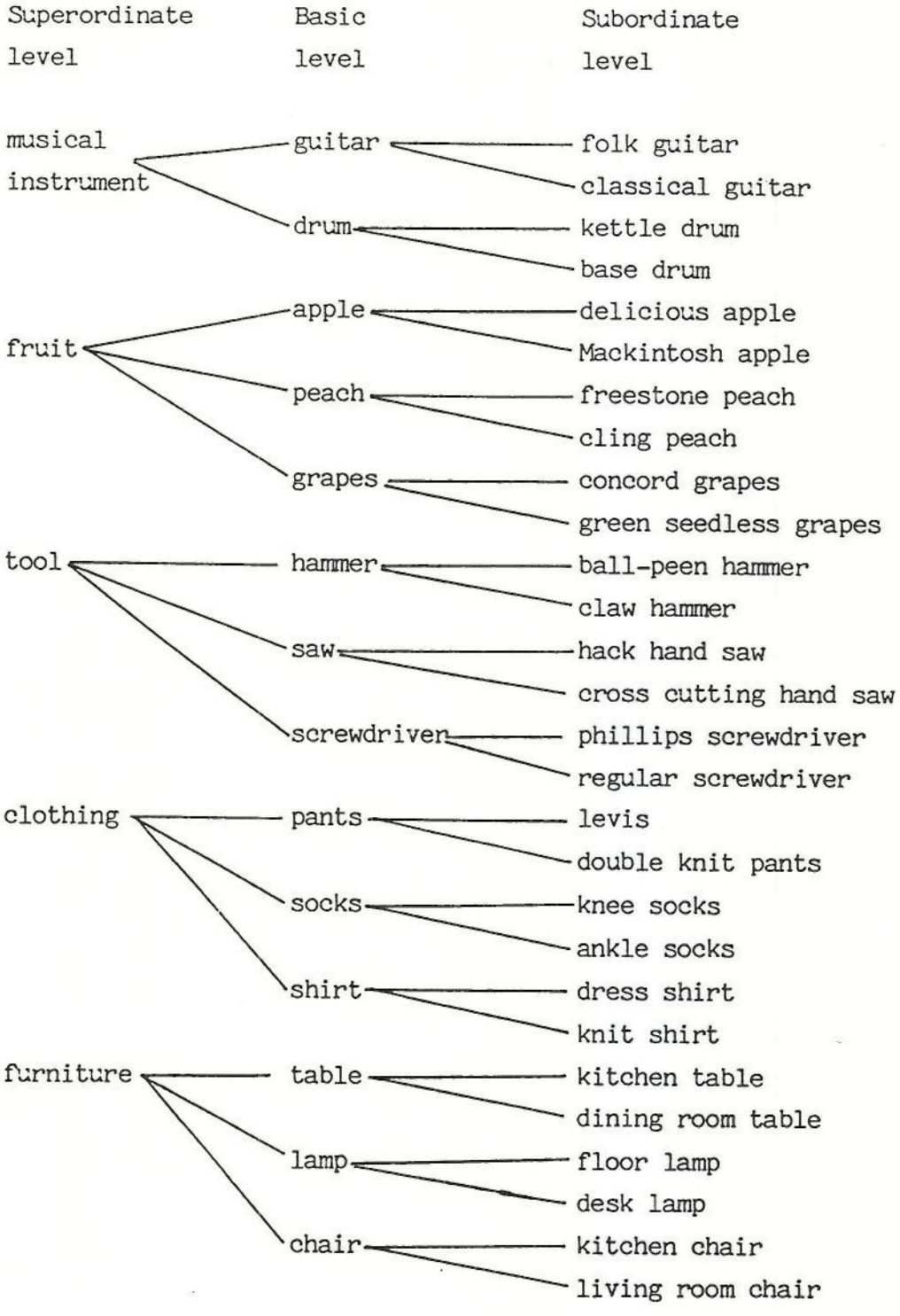
1. There is no distinction between thematic and conceptual classification. ECs are allowed (but marked as such) in all x and y positions in 5 and 6.
2. Pure thematic classification. No ECs are allowed.
3. ECs are allowed in some x and y positions, but not in others.

figure 5




A, B, C and D represent the sets of characteristic features of the concepts a, b, c and d.

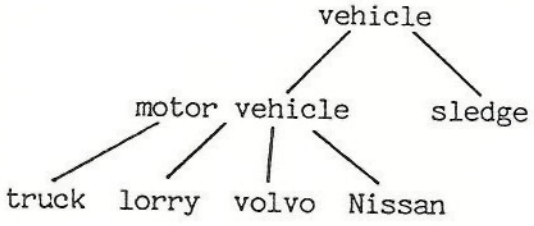
figure 6



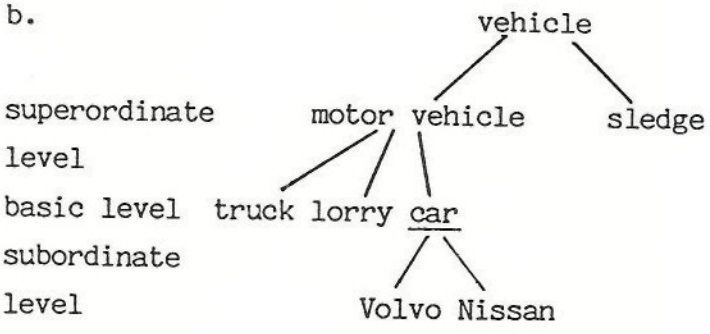

From Rosch et al (1976)

figure 7

a.


vehicle
motor vehicle
sledge
truck
lorry
volvo
Nissan

b.


vehicle
superordinate level
motor vehicle
sledge
basic level
truck
lorry
car
subordinate level
Volvo
Nissan

REFERENCE LIST


Ashcraft, H. (1978): "Property Norms for Typical and Atypical Items from 17 Categories: A Description and Discussion". In: Memory and Cognition 6, 227 - 232.

Armstrong, S.L., L.R. Gleitman, H. Gleitman (1983): "What Concepts might not be". In: Cognition 13, 263 - 308.

Battig, W.F., W.E. Montague (1969): "Category Norms for Verbal Items in 56 Categories". In: Journal of Experimental Psychology (monograph). 1 - 46.

Buchanan, B. (1976): A Glossary of Indexing Terms. London.

Dahlgren, K. (1985): "The Cognitive Structure of Social Categories." In: Cognitive science 9. 377 - 398.

DS/ISO/R 1087 (1976): "Terminologiens Terminologi". Dansk Standardiseringsråd. København.

Exploration and Production Thesaurus (E & P Thesaurus).(1986): University of Tulsa. Oklahoma.

Ordliste Gullfaks (1985). Preliminary version. Statoil. Bergen.

Radford, A. (1981): Transformational Syntax. Cambridge.

Reinton, J.E. (1987): "Petrus - Prosjektet 1985-86. Sluttrapport fra revisjonen av en petroleumstesaurus". Norske Språkdata 14.

Rosch, E. (1978): "Principles of Categorization". In: Rosch, E., B.B. Lloyd (1978), 27 - 48.

Rosch, E., B.B. Lloyd (eds) (1978): Cognition and Categorization. Hillsdale, New Jersey.

Rosch, E., C.B. Mervis (1975): "Family Resemblances: Studies in the Internal Structure of Categories." In: Cognitive Psychology 7. 573 - 605.

Rosch, E., C.B. Mervis, W.D. Gray, D.M. Johnson, P. Boyes-Braem (1976): "Basic Objects in Natural Categories". In: Cognitive Psychology 8. 382 - 439.

Rådet for teknisk terminologi (1976): Ordbok for Petroleumsvirksomhet. Universitetsforlaget. Oslo.

Tversky, A., I. Gati (1978): "Studies of Similarity" In: Rosch, E., B.B. Lloyd (1978). 79 - 95.

Wittgenstein, L. (1953): Philosophical Investigations. New York.

Wüster, E. (1985): Einführung in die allgemeine Terminologielehre und Terminologische Lexikographie. Copenhagen.