Ruth Feil & Lotte Weilgaard Christensen
Department of Business Communication and Information Science, Kolding
University of Southern Denmark
Denmark

THE ThT SYSTEM - A MULTILINGUAL NORDIC SEARCH INTERFACE

Abstract

The aim of this paper is to present the results of a project the purpose of which is to develop and test a prototype of a multilingual Nordic search interface which may, in combination with existing search engines, give monolingual access to information in several Nordic languages at the same time. The project is carried out by a Nordic research network called NorNa (Nordic Navigator). The languages involved are Norwegian, Finnish, Swedish, Danish, and English, which allows subsequent addition of non-Nordic languages. The ThT system (th standing for thesaurus and t for term), which has been developed specifically for this project, is based on a term bank application provided with additional modules such as a module containing search concepts. The data on which the project is based has been obtained via terminological analysis of a number of parallel corpora, one of them in the domain of the Arctic environment. The overall aim of the project is making the results obtained via research directly useful to trade and industry as well as public institutions in the form of a language technology tool which may contribute to optimum utilisation of Nordic knowledge resources. The project has received financial support from the Nordic language technology research programme of NorFA, the Nordic Academy for Advanced Study of the Nordic Council of Ministers.

1 INTRODUCTION

The aim of this article is to present the approach of a Nordic research network to multilingual information retrieval in the Nordic languages. The purpose of our network project is to develop and test a prototype of a multilingual Nordic search interface which may, in combination with existing search engines, give monolingual access to information in several Nordic languages at the same time.

The background for the NorNa project is that we in the Nordic countries – or at least in the Scandinavian countries – can understand each other even though we speak different languages. To most of us it is not unusual that non-Nordic people react with surprise when overhearing Scandinavians in a lively discussion in which all participants speak their respective mother tongues. The starting point for our network has thus been that a Dane may ask a question in Danish and get the answer in Norwegian. A Nordic search engine should be able to do the same.

The network has received financial support from the Nordic language technology research programme of NorFA, the Nordic Academy for Advanced Study of the Nordic Council of Ministers.

At our first network meeting in September 2002, we decided that the name of our network should be NorNa which stands for Nordic Navigator. Later we discovered that Norna also is the name of a rare orchid growing in the northern parts of Sweden. We have therefore made this orchid our logo.

2 APPROACH TO MULTILINGUAL INFORMATION RETRIEVAL

The participants in the network all represent institutions focusing on specialist communication, i.e. the Norwegian School of Economics and Business Administration in Bergen and the University of Bergen (Norway), the University of Vaasa (Finland), the University of Southern Denmark, Kolding and the software developer TERMplus ApS in Copenhagen (Denmark), including observers from Stockholm (Sweden) and Ventspils (Latvia). The participants all adhere to Wüsterian terminological principles in one form or another.

This means that we have a very strong focus on LSP texts and that our multilingual approach is embedded in an onomasiological perspective. The fact that we use a concept-based method is an obvious consequence of our need to find texts which describe the same concept in different languages – rather than the different senses of a given lemma.

Our overall approach has been very pragmatic. We have chosen to define a somewhat restricted task for our project work, as reflected in the fact that our documentary basis is quite limited and at the same time clearly defined. On the one hand we want to make sure that the texts we analyse are exemplary, and on the other hand we do not want to end up with a too comprehensive material given the time and the resources available for the project.

The title of this paper begins with the phrase 'ThT system' where 'th' stands for thesaurus whereas 't' stands for term. We have chosen this name to indicate that together with terminology, I&D (Information and Documentation) plays a crucial rôle in our project. Thus the letters ThT express the basic idea behind our approach: not only is the termbase we are constructing based on documents; it is also intended to function as a multilingual retrieval system in which the terms serve as search terms.

When it comes to software the aim has been to develop an exemplary prototype with a view to future product maturing. For this purpose we have decided not to develop new software from scratch and instead chosen to:

- either start from existing software and develop it further
- or exploit facilities of existing software to meet our needs

The central piece of software is a further development of TERMplus, a terminology management system originally named DANTERM for Windows. The use of existing software in our project includes in particular modules in MSWORD and MSACCESS.

The Nordic languages that have been analysed are Swedish, Norwegian Bokmål (one of the two official Norwegian languages), Danish, i.e. the Scandinavian languages, as well as Finnish, since the Finnish network group in Vaasa has native speaker competence in both Swedish and Finnish; English is also included, initially thought of as 'pivot language'. However, in our onomasiological perspective it is more correct to designate the concept as the pivot, which implies that all languages are equal. English has rather come to be viewed as a "bridge language" in the project, since English is used in order to facilitate communication with people outside the Nordic region.

3 WORKING PROCEDURE

The overall working procedure for developing a fully implementable search interface consists of the following stages:

- the establishment of parallel specialist corpora / full-text databases
- the establishment of a classification
- the establishment of an index with search concepts in one language followed by the inclusion of equivalents in other languages
- the development of a search interface for the selected domain

3.1 Establishing corpora

Three corpora have been included in our project. In the initial project phase the network group worked with a small test corpus consisting of a TV manual, whereas our two main corpora are:

- a document on the Arctic environment in the Nordic countries. The text was selected 1) because the documentation was available in all project languages AND 2) at the same time the Arctic region is of central concern and a very topical area in the Nordic countries. The Danish network participants are "primary providers", i.e. have taken on primary responsibility for this subcorpus.

- Nordea's Year-end report. This domain has been selected since several of the project participants have specialised in economic-administrative terminology. The Norwegian School of Economics and Business Administration has for instance initiated the establishment of a knowledgebase within these domains. Here the Norwegian network participants are primary providers.

These two subcorpora have been provided with header information and basic XML mark-up.

The size of the two main subcorpora can be tabulated as follows:

The Arctic regions

Danish	65,000 tokens
English	72,000 tokens
Swedish	60,000 tokens
Norwegian	63,000 tokens
Finnish	45,000 tokens

Nordea's year-end report

approx. 9,000 – 10,000 tokens in each language

In the text on the Arctic regions, the word count is roughly equal for Danish, Norwegian and Swedish, while the number of tokens in Finnish is markedly lower, with a difference of 20,000 tokens between the Danish and the Finnish version.

Nordea's Year-end report consists of texts of approximately 10,000 tokens. This is not a very large number of tokens. Still, it is near the magic figure of 10,000 described e.g. by Khurshid Ahmad and Margaret Rogers (1994) at the University of Surrey as being sufficient to conduct some forms of LSP corpus research, such as term extraction.

3.2 Establishing classifications

When the subcorpora have been established, it is necessary to analyse and determine the content structure in the texts to develop a classification. Only then can the actual work to establish an index of search concepts for the various subcorpora begin.

The NorNa classification is based on the specific content structure of the texts. This content structure is revealed by the table of contents or, in the case of the Arctic subcorpus, by the preface, which proved to be a valuable source of inspiration for working out a classification. The latter is a strong indication of the document-based approach we have applied.

The classification of the text on the Arctic regions has resulted in the structure shown in fig. 1.

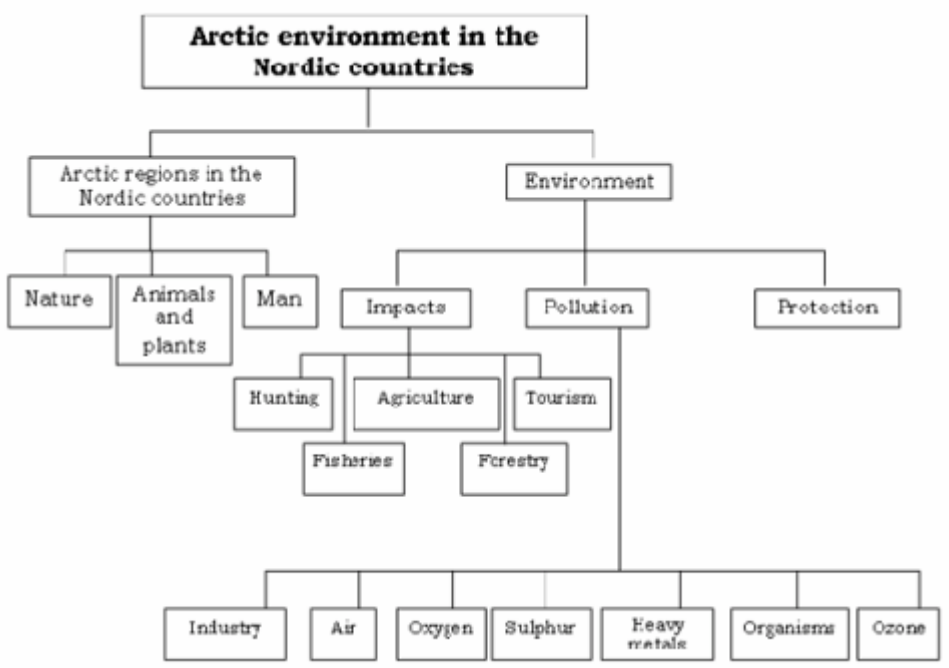

Fig. 1

3.3 Establishing indexes with search concepts

The indexes consist of search concepts represented by the terms in the five languages. The process of establishing indexes includes the following stages:

- An interactive extraction of possible search concepts (using WORD index as a tool)
- An interactive systematisation of these concepts (using a module called TREE VIEW as a tool)
- The search for equivalents (using five-language alignment as a tool)

Our extraction of search concepts is based on an interactive extraction of potential search concepts, for which we have used WORD's index facility. Possible search concepts are marked manually, and then WORD automatically marks all matches of the given string of words in the respective corpus text. The search concepts will then be grouped automatically in an alphabetical list at the end of the process. The extraction stage thus results in an alphabetical index.

The task of the primary providers of a subcorpus has been to extract search concepts in their own language(s) and distribute the list of these search concepts to the other groups, i.e. the above mentioned alphabetical index. Thus one language will be used as the point of departure when the index is established.

When an alphabetical index has been prepared for one of the languages, the extraction of equivalents in the other languages may begin. This is done by aligning the five languages. To aid the extraction of equivalents, an alignment of the texts from which the search concepts have been selected is used, that is, either the Arctic texts or the Nordea texts in all five languages.

In our terminology a search concept is represented by a term, which means that when search concepts are extracted from texts, this selection will be in the form of terms.

The five-language alignment is illustrated in fig. 2.

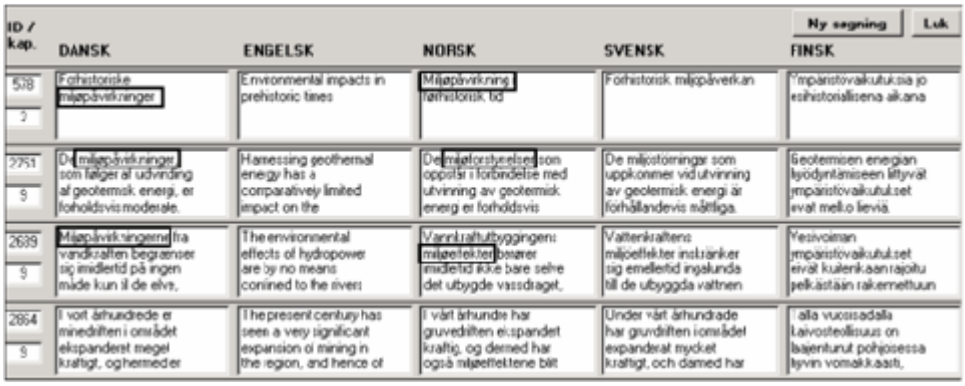

Fig. 2
The figure shows a search into the five-language alignment, which has been prepared in an ACCESS database. The search is made for the search concept environmental impact in Danish. All sentences which include this search concept are displayed together with aligned sentences in the other languages. In this way it is possible to find the equivalents in the other languages. Moreover, the method in question has proved extremely useful when it comes to recording synonyms. The example illustrates how synonyms may be recovered by means of the aligned texts. As shown in the aligned Norwegian sentences, there is a terminological inconsistency in Norwegian – in this example alone the concept environmental impact is translated in three different ways in Norwegian. It should be mentioned, though, that the translation miljøeffekt is strongly influenced by English and is considered untypical of Norwegian.

The list with the terms in all five languages is then imported into the terminology management system TERMplus, which has been enhanced with a module called TREE VIEW. In TREE VIEW it is possible to 'drag and drop' the search concepts manually to establish a hierarchical structure.

In fig. 3 an index of the conceptual structure of the text on Arctic regions is shown. This structure corresponds to what was shown in fig. 1.

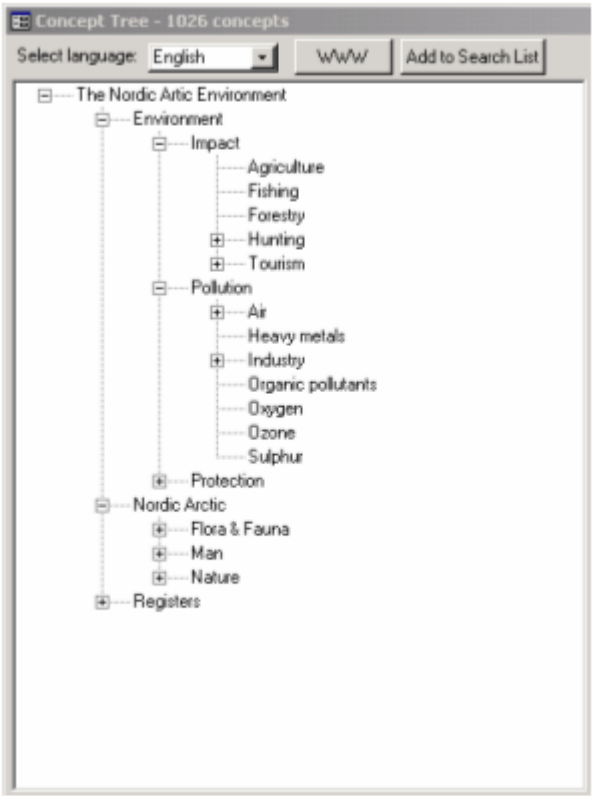

Fig. 3

3.4 Development of search interfaces for the selected domains

Based on the alphabetical index a simple index structure is built in the system, more specifically in the module named TREE VIEW in which the conceptual relations will simply appear as hierarchical, i.e. without any specification of the nature (generic or partitive) of the relation, which means that neither causal nor any other type of relations will be revealed in the structure. The relation type will, however, be specified in the termbase, to the extent that definitions or other content-based information on the search concepts have been recorded in the termbase.

In TREE VIEW it is possible to select the other four languages as well so that the above index may be shown in for instance Finnish.

4 SEARCH INTERFACE

In our opinion, it is an absolute requirement for the search interface to allow the user to identify the search concepts by widening or narrowing searches horizontally and/or vertically.

The **horizontal** dimension involves a single concept, to which one or more terms may be associated in each language, in addition to orthographic variants, i.e. a synset. The example given in fig. 4 shows the search concept waste management in English together with its equivalents in the Nordic languages.

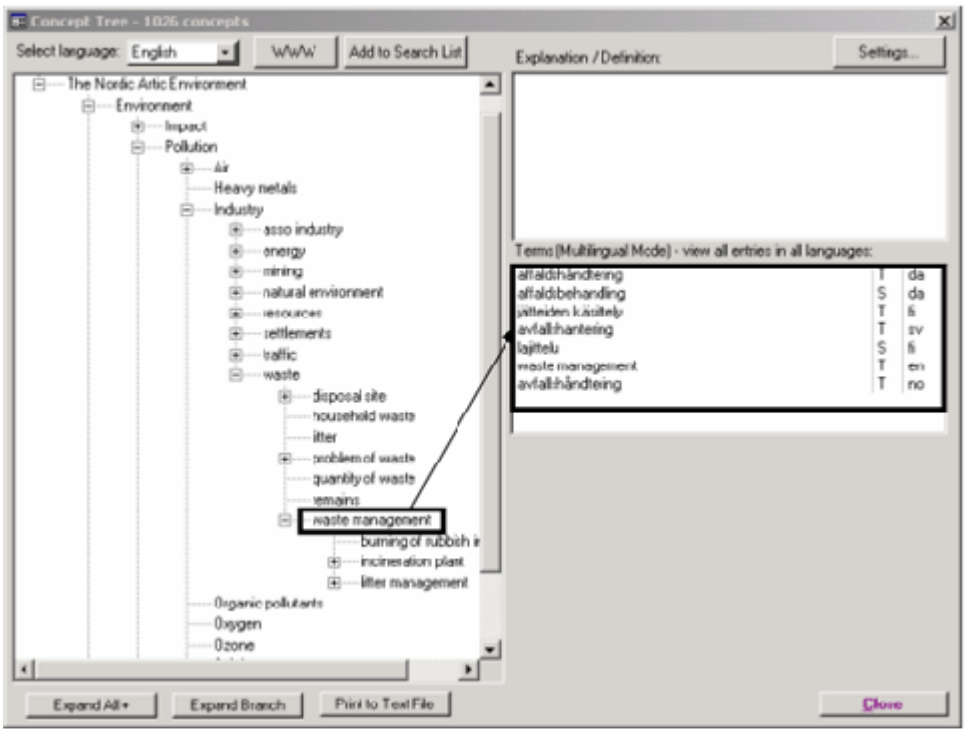

Fig. 4

Vertical identification implies that the searches may be enlarged by including superordinate concepts, coordinate concepts and subordinate concepts as shown in fig. 5.

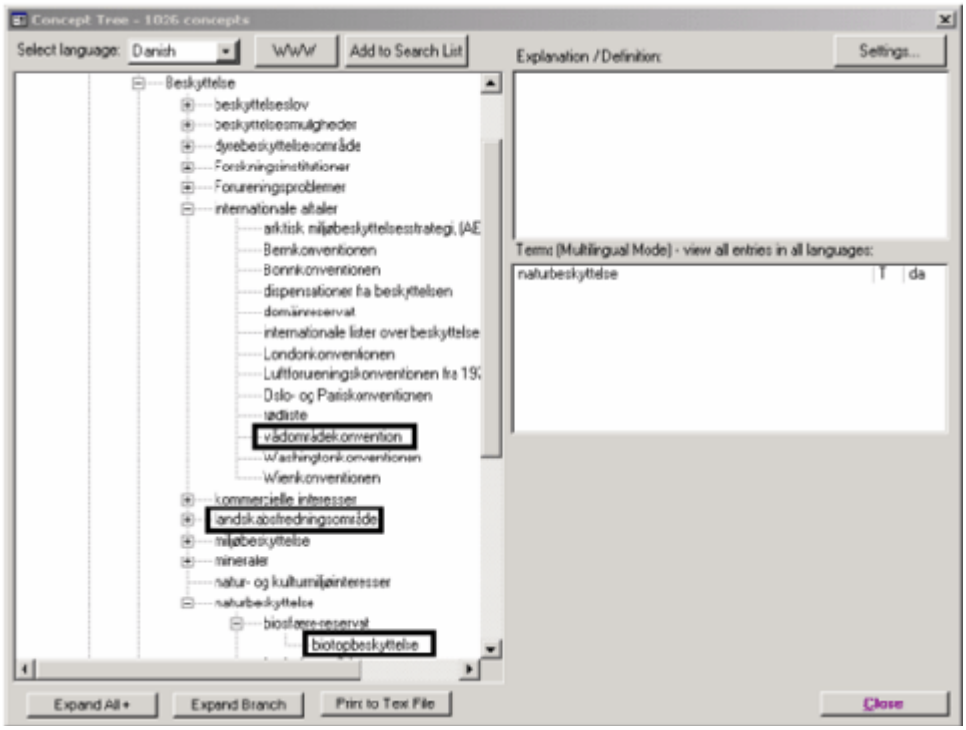

Fig. 5

Fig. 6 illustrates the search interface that contains the search concepts in several languages. First the search concepts to be searched for are selected, and then they are added to a search list, which is submitted to a search engine e.g., Google.

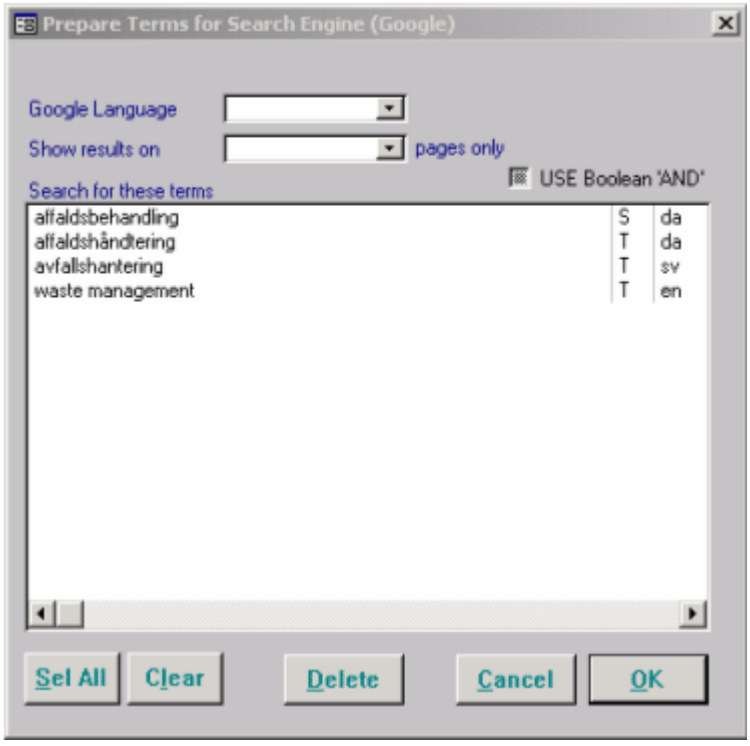

Fig. 6

At the same time it is also important to be able to validate the search concepts. This is done by consulting content related information in the terminology management system, primarily definitions and explanations, to ensure that the selected search concept holds the sense one is interested in and is searching for. Recorded definitions and explanations of the search concept are shown in the right part of

the window, as illustrated in fig. 7. After validation search concepts or selected parts are sent to a search engine.

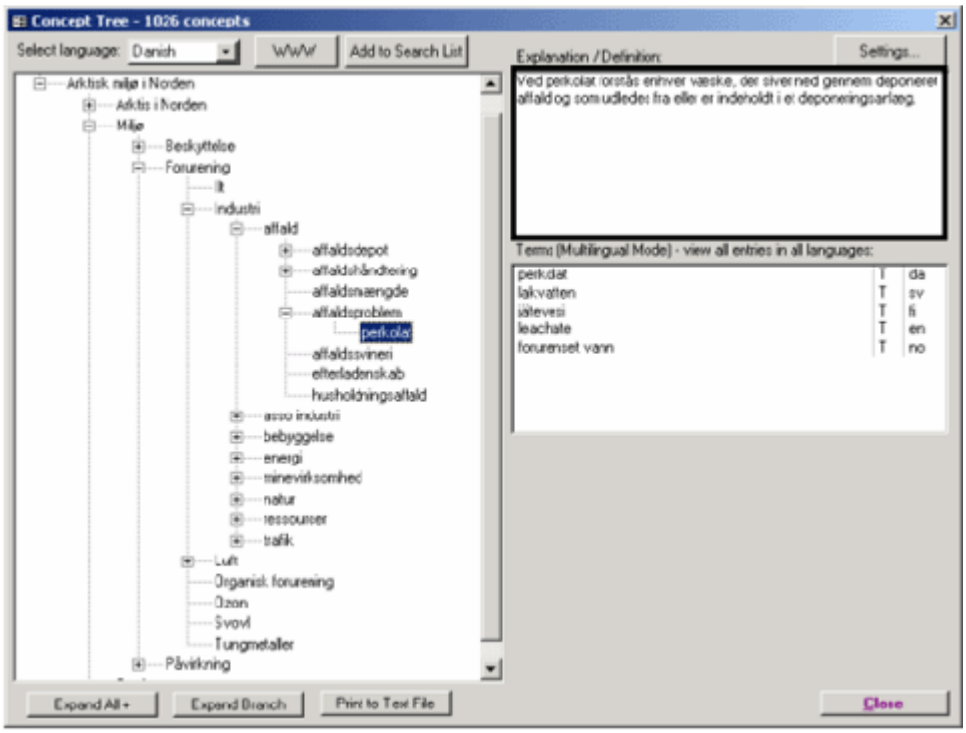

Fig. 7

5 THE THT SYSTEM IN RELATION TO I&D AND TERMINOLOGY

To conclude we should like to compare documentation thesauri, which comprise I&D search terms, with termbases. The comparison takes as its point of departure an article by Øivin Andersen (2005), our colleague at the University of Bergen, Norway. We have supplemented his comparison with the ThT system. In this article we have selected only a few of the aspects presented by Øivin Andersen.

Fig. 8 shows the similarities and differences among documentation thesauri, termbases, and the ThT system.

Documentation Thesauri	Termbases	The ThT System
Information	Communication	Primarily like I&D, but also T
Finite and relevant documents	LSP corpora	Like I&D
Incomplete hierarchies	Ideally complete hierarchies	Like I&D
No distinction as to type of relation	Distinction as to type of relation	Like I&D

Fig. 8

1. According to Øivin Andersen's article, in I&D information is regarded as the crucial aspect whereas terminology focuses on communication. In the ThT system, both are crucial since we use a term management system as a platform for cross language information retrieval.

2. One of the main differences between the two systems is that documentation thesauri are based on a finite and relevant set of documents, whereas LSP corpora for terminological purposes will comprise several central texts of a domain.

In the NorNa network we have so far only been able to work with a single document in each of our corpora, which means that for the time being, our approach is more similar to that of I&D.

3. One of the main differences is that the hierarchies of documentation thesauri are incomplete compared to those of terminological conceptual systems. In the ThT system, the hierarchies will also be incomplete, mainly as a result of our document based approach, which implies that our work is based on a single document not representative of the entire subject field.

4. Unlike terminological conceptual systems, documentation thesauri usually provide no distinction between generic and partitive hierarchies, nor between any other types of relations. As mentioned above, the user of the ThT system may also verify the exact type of relation by consulting the termbase.

What we have just said may suggest that the I&D aspect is predominant in our approach, but it is important to keep in mind that though differences exist among hierarchies, both fields nevertheless use hierarchies as their point of departure. Besides, our approach is more similar to that of the terminologist than that of the documentalist.

In our project I&D and terminology supplement each other. The point of departure of the ThT system was information retrieval, but with a terminological approach.

6 FINAL REMARKS

The project network has used the term management system TERMplus as a platform and developed it further into a cross language information retrieval (CLIR) system for the Nordic languages.

In the spring of 2005, we plan an evaluation phase in which the coverage of the ThT prototype will be tested via a Google search. In addition, we shall increase the amount of data in the termbase. On the one hand, we will first of all add more definitions and explanations in order to be able to validate the search

concepts, thus strengthening the terminological side of our approach. On the other hand we will add more descriptors, i.e. add key words the way it is done in I&D to the terms we have recorded in our document based terminology approach, i.e. the terms functioning as search concepts in the prototype. This means that in this respect, too, we plan to further develop the system both on the I&D and the terminology side.

6 REFERENCES

AHMAD, K. & M. ROGERS (1994). The analysis of text corpora for the creation of advanced terminology databases, Kolding.

ANDERSEN, ØIVIN (2005). NorNa-tool som et hybridsystem. In: Holmboe, Henrik (ed.): Nordisk Sprogteknologi, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2002-2004, Årbog 2004, Museum Tusculanums Forlag, Copenhagen, 81 - 95.

BERNES, CLAES (1996). Arktisk miljö i Norden: orörd, exploaterad, förorenad, Nordiska ministerrådet, Köpenhamn, Naturvårdsverket, Stockholm, ISBN: 92-9120-897-3 (Nordiska ministerrådet), Nord 1996:21, ISBN: 91-620-1168-5, Monitor 15

WEILGAARD CHRISTENSEN, LOTTE & GERT ENGEL (2004). Datafangst ved hjælp af en tværsproglig, nordisk søgemaskine - NorNa. In: Holmboe, Henrik (ed.): Nordisk Sprogteknologi, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2002-2004, Årbog 2003, Museum Tusculanums Forlag, Copenhagen, 243 – 247