Content Management - Concept and Indexing Term Equivalence in a Multilingual Thesaurus

Susanna Keränen
Åbo Akademi University, Turku, Finland

skeranen@abo.fi

Abstract

Languages and the thinking they reflect stem mainly from cultural needs for expression. A controlled vocabulary, thesaurus, can be seen as a cultural product. The focus of this study is **the translatability of British-English social science indexing terms into Finnish language and culture on a conceptual, term and indexing term level.** The emphasis is on Finnish language and human factors. The study is quantitative-qualitative and the perspectives are both linguistic and sociological – a combination through which a broader understanding of the phenomena is being aimed at in the general frame of information science. The study uses multiple cases aiming at theoretical replication. It is thus an empirical case study and the goal is to illustrate a new theory of “pragmatic indexing (term) equivalence”. Several data collection and analysis methods will be used in order to construct a theory by triangulation of evidence. The aim of this research is a doctoral thesis in information studies.

Keywords: multilingual thesauri, translation problems, equivalence, controlled vocabularies, social science terminology

Introduction

This study concentrates on cultural and linguistic problems in multilingual thesauri. The paper begins with a background and the need for this kind of study, next is reported the planned study and last are discussed the results of a case study made.

The need for this kind of research has grown rapidly in the multicultural environment of the Internet. As Milstead (1998) states,

> “The information retrieval world has changed dramatically in recent years, with the immense increase in availability of searchable full text -- and the increasing availability of powerful engines for searching the text. It is reasonable to ask whether there is any place left for thesauri in this new information retrieval environment. I believe there is a place for thesauri -- or something like them - but they must change in order to continue to be of value, and it is hard to predict just what the changes will be. --- Today it is beginning to seem as if all information is available in full text. However, this is not true, nor will it be true in the immediate future. Vast numbers of legacy documents remain, and converting these to searchable text is an expensive, long-term proposition. Furthermore, many documents are still being produced only in printed form. Therefore, thesauri and indexing will continue to have a place - at least for awhile - in facilitating access to documents for which electronic text is not available. Their long-

> run value, however, depends on integration with full-text search."

Although searching only full texts, metadata still has its place also today (see also Miller, 2000; Mulvany, 1997):

> "While metadata has become a buzzword in the information business, the concept is important for both authors and seekers of electronic information. Used effectively, it makes information accessible by labelling its contents consistently. Metadata leaves a pathway for users to follow to find the information they need—all in one place. In invisible cyberspace, this is even more important than in a library where desperate users at least have shelves to browse." (Milstead & Feldman, 1999)

The rapid development of modern societies changes language and terminologies in the social sciences. Together with a widening globalisation, this increases the need for high quality documentation tools for locating up to date information from multilingual resources. Jorna and Davies (2001) state that most current forms of multilingual information access are inadequate to answer the needs of increasingly diverse user groups from different cultural and linguistic backgrounds, and that a new form of multilingual thesaurus is therefore required.

Information retrieval and seeking over national borders are constantly on the rise. The success of creating and using international information resources still depend on common tools and an understanding of the concepts used. Along with expanding use of international databases, cross-cultural indexing is becoming more common and tools like multilingual thesauri are needed.

Traditionally, a thesaurus can be defined as a documentary language and it is developed to be a tool in information retrieval and documentation. Terms in a thesaurus have many equivalents in natural language – especially in the spoken language. The manner of representation of concepts in a thesaurus is artificial – hierarchical and associative relationships being the context of the thesaurus terms. This kind of representation aims at helping an information seeker to broaden or specify the search.

Guidebooks and standards for thesaurus construction are very strongly focused on linguistic matters when defining a good indexing term. – A good term should be transparent, consistent, practical, brief, productive, easy to pronounce and spell, and unambiguous. It should also otherwise be linguistically suitable and preferably based on one's own language. (Aithchison & Gilchrist, 1987, p.12; Haarala, 1981, p.37; ISO, 1987, p.12; SFS, 1988, p.172; TSK 1989, p.74-9.)

The aspects related to multilingualism in Library and Information Science (LIS) are today studied more from the perspective of cross-language information retrieval, CLIR, and machine translation (see e.g. Gollins & Sanderson, 2001; Pirkola, 2001; Sperer & Oard, 2000). Thesaurus matters are studied from the perspective of constructing methods (see e.g. Nielsen, 2001, Schneider, 2001), information retrieval (see e.g. Greenberg, 2001), automatic semantic indexing (see e.g. Hofmann, 1999) and, perhaps closest to this research, from the viewpoint of semantic problems (see e.g. Doerr, 2001; Jorna & Davies, 2001).

Although these other aspects have a great significance and further research is needed, there is still need and room for other, overlooked, aspects, too. - Problems related to both the multilingual aspects (human communication, communicating cross cultures) and the thesaurus construction (equivalence problems, human effort) in our field have stayed in shadows, even though controlled vocabularies and thesauri are nowadays commonly developed also for use on the WWW. For example, the EU has developed a multilingual Eurovoc thesaurus (see EU, 2002) and European social science data archives have constructed a multilingual ELSST thesaurus (Miller & Matthews, 2001). Also monolingual thesauri have been developed recently and are in use. The Social Science Information Gateway, for example, uses a number of monolingual subject specific thesauri (SOSIG, 2000) and the Inter-university Consortium for Political and Social Research (ICPSR) has developed a monolingual controlled vocabulary system (includes a Subject Thesaurus) to assist users searching its social science data archives (ICPSR, 2002). Thesauri need to be treasured - many studies support the usefulness of thesauri also in the age of WWW (see e.g. Doerr, 2001; Greenberg, 2001; Tudhope, Alani, & Jones, 2001).

Background

Very little attention has been paid to cultural aspects and the problems of minor languages in the digital environment. Nevertheless, constructing a multilingual thesaurus demands operating distinctly in a multicultural environment. When studying thesaurus construction we are operating with old problems, but ever so current.

Equivalence

In multilingual and multicultural terminology work one of the key problems is **equivalence**. In defining the equivalence level a thesaurus constructor tries to anticipate what kind of term will most likely be used in information seeking and retrieval.

Usually, equivalence means similarity, (perfect) reciprocity (see e.g. Itkonen, 1990, p.83). Terminology and thesaurus construction standards and guidebooks provide very little details and consideration for equivalence. In International Organization for Standardization’s standards *Documentation - Guidelines for the establishment and development of multilingual thesauri* (ISO, 1985, p.8) equivalence is divided into exact equivalence, inexact equivalence, partial equivalence, single-to-multiple equivalence and non-equivalence.

In translation studies equivalence is understood in many different ways, but a general consensus today seems to be that there is no sense in demanding ”mirror-translation”. Still, even today, translating is not seen as an easy act. A translator is commonly seen as a ”prisoner” of his/her own culture. Eugene Nida and William Reyburn (1981) have found, that a translator usually understands the message in accordance with his/her own cultural-linguistic context. Usually a translator is aware of that and tries to solve the problem by using foreign terms. Regardless of that, a concept in the source language is not always semantically equivalent with the same concept in the target language. Generally, descriptive phrases are therefore better (more equivalent) than foreign terms in translating a message into another culture. (Nida & Reyburn, 1981, p.21-25) This can also be seen in the thesaurus construction standards and guidelines where loan words are not recommended (ISO, 1985; ISO, 1986; SFS, 1988).

Language and Culture

The language pair of this study is Finnish-English and the subject fields studied are the social sciences. The fact that English is the most common language on the Internet (see e.g. Grefenstette & Nioche, 2000; Helsingin Sanomat 2001) and also the common working language between international social science documentation providers like the members of the Council of European Social Science Data Archives (CESSDA) adds to its importance also in the international environment of social sciences and information seeking. The social sciences are connected not only to the development of science but also to the development of their surrounding culture and society. In a social sciences thesaurus this phenomenon is seen more clearly than in, for example, thesauri of technology or medicine. Language is not static and therefore the language and documentation of social sciences is tied with culture and time (see e.g. Aitchison, 1984; TSK, 1989; Varantola, 1990; Wierzbicka, 1997).

In this research **culture** refers to a **conceptual** level and can thus refer to:

1. a (geographical) culture

Culture can then be - on a very general level - defined as a framework for our lives, something that affects our values, attitudes and behaviour. Attention is paid to language and communication styles as a dimension of cultural differences between Finnish and British culture.

2. a subculture

Different groups, for instance, Finnish and British indexers versus Finnish and British specialists. Attention is paid to institutional differences.

Language refers to an **expressional** (verbal) level. **Linguistic** differences can thus occur between different:

1. languages

Finnish versus British-English. Their characteristic problems differ from each other. On the WWW English is a dominant language and Finnish very minor (Grefenstette & Nioche, 2000). A common and growing trend in Finland's academic life is to write and publish in English and there has been a wide and lively debate on whether to use Finnish or English. On one hand it is important to have a bigger audience, on the other, it is essential to publish research results in our native language and thus also ensure that the special terminologies will be up-to-date also in Finnish (see e.g. Leiwo, 2000; Oittinen & Väyrynen, 2001).

In free-text searching the richness of noun case inflections in Finnish is often considered a problem: Finnish has 15 cases. In Finnish the inflection of words is done by adding grammatical affixes instead of using prepositions like in English and other Germanic languages. An illustrative (and typical) example within the field of information studies is English *search* versus Finnish *haku*:

- *in searches – hauissa* (means also *in pikes*, whose basic, nominative form is *hauki/pike*)

- *searches – haut* (nominative) / *hakuja* (partitive)

In controlled vocabularies the noun inflections are not a problem, but another problem (influencing also free-text search) are compounds. In Finnish, it is often not clear whether to write compound words together or not.

2. discourses

Each information search in a database covers at least five different languages: the authors, the indexers, the synthetic structure, the users and the search strategy (Buckland, 1999), which all represent a type of discourse. An indexer's or a specialist's ways to express their ideas and thoughts on a certain social environment differ from each other and in indexing this can cause problems.

According to Suojanen (1993) the world's information culture has commonly approved of ”a neo-language” or ”an euphemistic language” in public and official discourse – technology, politics, economics, religion – especially when the context of a phenomenon feels heavy, scary, threatening. With this euphemistic language we try to guide the thinking of familiar things from a new perspective by the choice of words or a new compound term. Especially the media talks e.g. about ”the poor” with words like ”low-paid”. (Suojanen, 1993, p.23) A specialist may thus in some situations use euphemisms and in some cases more unambiguous terms, whereas indexers may aim to conform with guidelines and controlled vocabularies in some institutions more tightly than in others.

ISO and SFS standards (ISO, 1987, p.12; SFS, 1988, p.172) define what kind of term is a good indexing term. They may sometimes greatly differ from other guidelines and common language usage and are not always easy to follow. The time dimension is also considered to be one of the major factors in communication styles.

Multi**cultural** thesauri refer to the using environment of the thesauri studied in this research. The thesauri are meant to be used extensively in European social science resources.

Multi**lingual** thesauri are tools for information retrieval and documentation, where indexing terms have equivalents in one or more foreign languages. They are not necessarily multicultural in a sense that they may be bound to the surrounding culture, e.g. the British.

Thesauri in a Multilingual Environment

ISO standard 5964 - *Guidelines for the Establishment and Development of Multilingual Thesauri* recognize three approaches to the construction of multilingual thesauri:

1. Ab initio construction, i.e. the establishment of a new multilingual vocabulary without direct reference to the terms or structure of an existing thesaurus;
2. Translation of an existing monolingual thesaurus;
3. Reconciliation and merging of existing thesauri in two or more working languages. (ISO, 1985)

According to Michele Hudon (1997) the problems traditionally associated with multilingual thesauri are:

1) that of stretching a language to make it fit a foreign conceptual structure to the point where it becomes barely recognizable to its own speakers;
2) that of transferring a whole conceptual structure from one culture to another whether it is appropriate or not;
3) that of literally translating terms from the source language into meaningless expressions in the target language, etc.

One way to create a multilingual thesaurus is translating an existing monolingual thesaurus. Hudon points out that a monolingual thesaurus is always culturally biased, and a straight translation might lead to a form of “cultural imperialism”. (Hudon, 1997)

Also Doerr (2001) states that even though semantic heterogeneity of terminological resources has frequently been referred to as a problem, a systematic analysis of its intellectual basis and structure has not been carried out. According to him, translated thesauri are thesauri, “--- where each concept is optimally interpreted in words of another or multiple languages, to allow speakers of those languages to understand better and use concepts of this thesaurus more effectively”. He also stresses that such translations are in general not established indexing terms of the target language. (Doerr, 2001)

On the other hand, controlled vocabularies reduce linguistic problems even in a monolingual searching environment and the benefits are the same – or even greater - in a multilingual environment. According to Milstead and Feldman (1999) metadata attacks three language problems that cause poor precision: polysemy, synonymy and ambiguity. When operating with a foreign language, these problems are even more difficult to solve without any vocabulary or terminological help.

> “The larger the information domain, the more important is to find an effective and efficient way to define narrower domain for searching. One of the major causes of false hits in retrieval is homographs, that is, words that look the same but have different meanings. The advantage of searching within a specific domain is that terms are often ambiguous across several disciplines, but seldom have multiple meanings within a particular discipline or subject domain.”(Chan, Lin, & Zeng 2000, p.188)

Thesauri usually include scope notes to define or clarify the meaning and use of ambiguous terms in that certain context and thus help the indexer and seeker to select a proper equivalent for their purposes. Also different virtual libraries (like the Finnish Virtual Library, *Virtuaalikirjasto,* and the British SOSIG) aim to restrict information retrieval to narrower and more relevant domains.

Research in Finland

The theme of this research is very topical, because today many multilingual terminologies and thesauri are being developed (Forsman & Keränen, 2000). The problems in their construction have not been studied much. Translation studies as a science is also very young and multicultural and cross-cultural communication as a research topic has become more common in Finland only since the 1980s. In Finland information science concentrates today more on problems in information seeking rather than in storage and documentation. Yet, **an effective content analysis and documentation of information is a condition for effective information retrieval.**

In the Department of Information Studies at the University of Tampere there are three on-going research projects on multilingualism: Dictionary-based cross-language information retrieval (Ari Pirkola), Swedish language information retrieval (Turid Hedlund), TEQIR - translating and expanding queries for information retrieval (Kalervo Järvelin, Ari Pirkola, Turid Hedlund, Jaana Kekäläinen). These projects, in contrast to this study, focus on information seeking and machine translation and thus do not overlap with this work. Mirja Iivonen (1989, 1995) has earlier touched the subject field from the perspective of indexing and information seeking and Jarmo Saarti (1999) from the perspective of content description of fictional literature. The indexing of Finnish articles in two databases has been compared by Mirja Iivonen and Katja Kivimäki (1998). The documentary languages in four international databases in the subject field of biotechnology have been compared by Sara von Ungern-Sternberg (1994, 1998).

The Study

Aim

The focus of this study is the translatability of English social science indexing terms into Finnish language and culture on a conceptual, term, and indexing term level.

The perspectives are both linguistic and sociological – a combination through which a broader understanding of the phenomena is being aimed at within the general frame of information science.

Research Questions and Hypotheses

Translatability is examined by trying to answer the following research questions:

Equivalence:

- What kind of equivalence type is aimed at and is that kind of equivalence possible to achieve in a documentary language translation?
- How do indexing term equivalence and concept equivalence differ from each other?
- What kinds of terms are considered problematic?
- Do equivalence problems mainly stem from the nature of language and/or culture or from the nature of the documentary language or domain?
- How are terms understood and defined by people representing different cultures and subcultures?
- In what way do equivalents given by specialists and indexers differ from each other?

Indexing practices:

- How do indexing practices differ from each other in Finnish, British and international databases?
- How do Finnish and British indexers index the same document?

- How are the terms studied represented in well-known and widely used thesauri?

The central **hypotheses** (propositions) of this research are:

1 Intercultural misunderstandings are often caused by

   - Subconscious blinders. People are not aware of their own assumptions and their cultural basis.
   - The lack of cultural self-awareness. It is a common belief that the main challenge in cross-cultural communication is to know the foreign culture, when in fact it is to know our own culture and how it affects our behaviour and thinking.
   - Projected similarity. In real life, other people differ more from oneself than is assumed. Another person's situation differs also more than is usually assumed. Differences can then be expected, imagined and discovered as similarities.
   - Cross-cultural misevaluation. We use our own culture as a standard of measurement. (Adler, 1997, p.78-87)

2 Languages and ideas expressed by languages are created mainly in accordance with expression needs of the surrounding culture (Suojanen, 1993, p.16) and thus, a thesaurus can be seen as "a cultural product". The content of a thesaurus depends on the documentary needs of the data sources in the surrounding culture. Finnish indexers are guided by the Finnish culture and its traditions. In Finland, the most important and widely used indexing term source is the Finnish general thesaurus (Yleinen suomalainen asiasanasto, YSA). Indexers usually find the Finnish equivalents for English terms in YSA and thus conform to the Finnish culture and the conventions in information retrieval and storage.

3 Translation problems of Finnish and non-Finnish concepts occur for instance on the basis of the thesaurus' structure, the differences between languages and cultures, and the time factors which influence the content of concepts (see Suojanen, 1993, p.19).

4 Indexing terms are part of a documentary language and because of this they can sometimes greatly differ from natural language and common usage. Specialists' and indexers' equivalents differ from each other. When selecting an indexing term indexers are more likely to take into consideration the standards' and guidebooks' requirements for a good indexing term.

5 Indexing term equivalence and concept equivalence are not necessarily the same. We can have concept equivalence between a Finnish and a British concept, but on the indexing term level equivalence levels differ from each other and vice versa.

Methods

This study is **quantitative-qualitative**. Qualitativeness is emphasised in developing the analysis method for this specific research material. Traditional terminology sources will be supplemented by the theories and concepts of translation studies, linguistics and communication theories in order to achieve a broader and more pragmatic perspective. Thus, a new research method can be developed. Linguistic and cultural matters cannot be clearly separated, because they go hand in hand. Therefore, in this study, concentrating on cultural or linguistic matters is more like an emphasis or a perspective.

The study uses multiple cases aiming at theoretical replication. It is thus an empirical case study and in its nature descriptive, generating hypotheses and illustrating a new theory of a "pragmatic indexing (term) equivalence". The samples studied are theoretical (not random) and thus cases assumed to give answers to the questions of this study.

Several data collection and analysis methods will be used in order to construct a theory by triangulation of evidence. The research process is thus linear, starting from hypotheses (propositions) and ending through observations and generalisations at new hypotheses and a construction of a theory.

The study examines Finnish equivalents for British-English indexing terms. The key terms of the research corpus will be studied:

- by **interviewing** people representing different cultures (Finnish versus British) and subcultures (indexers, specialists, multilingual thesaurus constructors);
- by a **modification of co-word analysis** (the use and **indexing practice** of the key terms studied in Finnish, British and international databases);
- by comparing several **thesauri** and;
- by **component analysis** (the recognition of the equivalence type).

Main tools for analysis on the indexing term level

Reasons and motives for the chosen equivalents will be studied by **interviewing** specialists, thesaurus constructors and indexers.

Explanations and datasets are compared by **classification** of the terms into different groups. Comparisons inside and between these groups are made on the basis of e.g. their type of equivalence, nativeness in the Finnish language and culture-boundness.

Problems can be caused by, for example

1 Society
The same kind of phenomenon or a concept as in the British society does not exist in the Finnish society. (*’homemaker’*)

2 Language

The concept exists in both cultures, but to find a term to describe it conforming to common Finnish usage is problematic. (*’illegitimate births’*)

3 Euphemisms

Socially difficult concepts are described with euphemisms. (*’family disorganization’*)

4 Institutional differences

To express the concept conforming to Finnish indexing language practices a thesaurus constructor has to use factoring. (‘*married women workers’*)

5 Citation loans, foreign words

There is not a Finnish native word to describe the concept and/or on the term level a citation loan or foreign word is commonly established. (‘*nationalism*’)

Finnish, British and multicultural indexing practices will be studied and compared with the key terms of the corpus. Indexing frequency of the key terms of this study and their equivalents will be studied by examining the use of terms and their synonyms in several databases using a **modification of a co-word analysis**.

Main tools for analysis on the conceptual level

To achieve a broader perspective and a better understanding of the content (connotation and denotation) of the key terms, people representing different cultures and subcultures will be **interviewed**. The aim is to find out the tacit assumptions and practices, which are linked with the key terms. Terminologies and thesauri usually operate on the denotation level of the words, but human communication – both formal and informal – uses also connotative meanings.

> “Connotation refers to what is associatively suggested by the word; denotation, on the other hand, points to the object which is meant by the word. The denotation of ”moon” is ”earth’s natural satellite”; the connotation of ”moon” would be ”cold, distant, lonely, longing…” (Hörmann, 1986, p.142)

The recognition of the equivalence type is in this research done by **component analysis**. It is commonly used by translators and known also as denotation analysis. In component analysis, the meaning of the word (denotation and connotation) is divided into smaller components, semantic characteristics. A component analysis is helpful especially in polysemy cases and in co-ordinated and related cases.

In component analysis one can use, for instance, a semantic characteristics matrix, where words are defined by semantic characteristics. With a matrix we can recognize the common characteristics of different words and the distinctive components:

	woman	work at home	mother	married	unmarried
housewife	+	+	0	+	-
kotiäiti	+	+	+	0	0
kotirouva	+	+	0	+	-

Table 1: Semantic characteristics matrix sample

Also the connotative meanings of Finnish and British-English terms can be studied with similar methods and tools.

Material

The terms for this study are selected according to three criteria:

1. For one English term there are several (quasi-) synonyms in Finnish:

1.1 *nationalism* versus *kansallisuusaate, kansallisuuskiihko, kansallismielisyys, nationalismi*

1.2 *homelessness* versus *asunnottomuus, kodittomuus*

1.3 *abuse of the elderly* versus *vanhusten hyväksikäyttö, vanhusten pahoinpitely* etc.

2. There is no exact Finnish equivalent for an English term

2.1 *white collar workers*

2.2 *family disorganization*

2.3 *illegitimate births*

3. A theme - terms related to gender (~parenthood and gender in educational and working life)

(15 preferred-terms, 16 non-preferred terms. The focus is on terms 3.9-3.11.)

3.9 *homemakers*

3.10 *housewives*

3.11 *married women workers*

The terms above will be studied in three ways:

1) Five databases are selected in order to get pools of documents on the topics in the groups mentioned above: Reference Database for Finnish Articles ARTO and Finnish Virtual Library and; English Social SciSearch®, Sociological Abstracts and The Social Science Information Gateway (SOSIG).
2) Nine thesauri are used to examine and compare the indexing terms used in the received documents in the database searches: The UNESCO Thesaurus, Eurovoc, The OECD Macrothesaurus, ELSST, ERIC, Thesaurus of Sociological Indexing Terms, HASSET, ICPSR Subject Thesaurus and YSA.
3) Finnish and British indexers will be asked to index the same documents, that is, articles retrieved from Finnish and British databases having the studied terms among the central ones (e.g. in title and/or as a descriptor and thus likely to be used in indexing).

A Case Study

A small-scale case study was done to illustrate what kind of linguistic and cultural differences and similarities are found in Finnish and British indexing on a terminological level from the perspective of a cultural (monolingual) and a multicultural (multilingual) thesaurus constructor and user.

The first step was to use one key term *nationalism*, and its Finnish equivalents (*nationalismi*, *kansallismielisyys, kansalliskiihko*, *kansallisuusaate*) to collect a sample useful in studying the differences in indexing practices at the national level, that is Finnish versus British emphasis on the Finnish language and practices. Datasets were obtained from three databases, Finnish ARTO and English Social SciSearch® and Sociological Abstracts using *nationalism* and the Finnish equivalents as search strategies. The Finnish datasets were limited to a maximum of 200 documents published after 1994. The English datasets were made to be commensurate with the retrieved equivalent Finnish ones.

The second step was to take for further examination the datasets retrieved with keywords *nationalism*, *nationalismi* and *kansallisuusaate* in order to find out what terms were used in indexing with the indexing term "nationalism" in the Finnish and English databases and to see if they were found in a certain multilingual thesaurus (ELSST). The final aim was to find out national characteristics in indexing and in a multilingual thesaurus (ELSST). In this case study the method used was a modification of co-word analysis. The software used was Bibexcel (Bibexcel 2001).

According to the case study we can state that the **Finnish indexing is highly coherent and conforms to the indexing guidelines and the common vocabulary tool (Finnish general thesaurus, YSA) used.** Documents having no indexing terms or terms not conforming to common practice were distinctly exceptional. On average, the retrieved documents had eight different indexing terms. The use of linked indexing terms (precoordinate indexing) occurred frequently. The use of indexing terms referring to place-names and time was quite common in the retrieved datasets. Also the top-ten lists of the most used indexing terms were rather similar – although not identical.

The terms used in Finnish and British retrieved datasets were very heterogeneous in their nature. The Finnish ones were more general and related to the surrounding society and its history, while the British ones were more complex and sociological. The cultural differences were not so common and clear as supposed.

For example the 166 documents indexed with *nationalismi* had 443 unique indexing terms in all, used altogether 1228 times. 343 of the used indexing terms (793 occurrences in all) did not have an equivalent in ELSST and a 100 indexing terms used (435 occurrences) had an equivalent in ELSST, which is almost one fourth of the terms used.

To a great extent, 75 terms with 234 occurrences in all, the terms with no equivalent in the ELSST thesaurus represented a place-name or were numeric expressions of time and thus represented rather *transliteration* than *translation* problems.

The retrieved sample included nineteen terms used 56 times in all that were related to Finnish society and/or history from the perspective of a multicultural thesaurus user/constructor. They were:

1. names of nationalities [*Finns, Russians, Estonians*];
2. names of ideologies or isms or features [*Russianism, Lapua movement* (an extreme right-wing organization in Finland, 1929-1932)*, Fennomania, Panslavism, Finnish nationalist movement, Finnishness* (i.e. the sense of Finnish identity)];
3. verbal expressions of time periods related to Finnish history [*Age of Autonomy* (from 1809 to 1917)*, Period of Oppression* (first from 1899-1905, second 1908-1914)];
4. other proper names, nouns [*Civil War* (In YSA the term Civil War – *kansalaissota* - refers only to the Finnish civil war, 1918, and it is defined in the term's scope note. The broader term and the plural form civil wars, *sisällissodat,* can be used to refer to any civil war. In Finnish we thus have different terms for the English plural and singular forms), *Continuation War* (*Jatkosota,* 1939-1945), (the) *Porvoo diet* (was held in 1809), *Winter War* (*Talvisota,* 1939-1940*), image of Finland*, *Havis Amanda* (name of a statue), *Turun Lilja* (name of a statue)], *Finnish*.

The multilingual thesauri studied were more general than the monolingual ones. The most culture-bound thesaurus was the Finnish general thesaurus, YSA.

In YSA *nationalismi* has a broader term *ideologies* (*ideologiat*) and it belongs to the thematic group [65]: Political science, Politics, International politics. Related terms are *Ethnocentrism (etnosentrismi), Fennomania (fennomania), Patriotism (isänmaallisuus), ~idea of nationality (kansallisuusaate), Nationality questions (kansallisuuskysymykset), Panslavism (panslavismi), Svekomania (ruotsinmielisyys), Slavophiles (slavofiilit)* and *Finnishness (suomalaisuus).*

The other Finnish equivalent and preferred term *kansallisuusaate* belongs also to the same thematic group [65] and additionally also to [52]: History. The related terms were the same as *nationalismi* has with certain exceptions: instead of *kansallisuusaate* is *nationalismi*, and in addition *kansallisuusaate* has also as related terms *Tribeism* (*heimoaate), National minorities (kansalliset vähemmistöt)* and *Separatism (separatismi).*

The Unesco Thesaurus, Eurovoc, ELSST and OECD Macrothesaurus represent the term *nationalism* in a neutral, universal way and thus they are not so bound with the surrounding culture, whereas ERIC, Thesaurus of Sociological Indexing Terms and HASSET can be seen as culturally biased having as related terms *Black Power, Colonialism, Scottish Nationalism, Zionism* and *Imperialism*, which all had occurrences in the retrieved datasets. Nonetheless the relationships in English thesauri were not so much related to the history of e.g. United Kingdom or United States as the Finnish ones were related to its own.

In all the loss of indexing terms used in ELSST was about 80 % in the datasets retrieved. The retrieved Finnish datasets missed more general terms while the retrieved British datasets more complex terms.

The English sample retrieved from the Sociological Abstracts database and indexed by indexers (seven in all) with a vocabulary tool was more homogeneous when comparing the number of indexing terms used. It was also more homogeneous when comparing the amount of concepts with more matches to the other indexing tool studied (ELSST) than the sample retrieved from the Social Scisearch® database and indexed by the authors.

It has also been claimed that a monolingual thesaurus is always culturally biased, and a straight translation might lead to a form of "cultural imperialism" (see Hudon, 1997). However, ELSST has been constructed in a close co-operation with European social science data-archives. According to the results of this case study ELSST seemed to be so general and universal that information losses occurred also in the British datasets. The other multi*lingual* thesauri studied were also multi*cultural* in the respect that they were so general that they did not feature any national characteristics. This leads us to a further question of principle and practice: how the national/cultural needs and characteristics should and could be treated when constructing multicultural thesauri?

When comparing only the amount of indexing terms used in Finnish and British indexing we can say, that the depth and specificity of indexing is about the same. However, there were **clear differences in the nature of indexing**. The Finnish and British samples differed mostly in the nature of the selected indexing terms in the sense that the British keywords retrieved often represented more complex and abstract concepts than the Finnish ones. The complex concepts in Finnish datasets were indexed using precoordinated indexing, where several more general terms were linked together. Although *kansallisuusaate* and *nationalismi* were commonly used in Finnish indexing they were not common in titles.

The main result of this case study was that Finnish indexers conform to the Finnish indexing practices and guidelines when using the Finnish general thesaurus YSA. It is therefore essential to aim at constructing multilingual thesauri including Finnish in as total harmony with YSA as possible. There were also cultural differences in the content of the concepts. Mostly differences stemmed from different indexing practices. When constructing a multicultural indexing vocabulary it is good to be aware of nationalistic characteristics in existing indexing practices.

Significance of the Results

Research on multilingual thesauri is useful for all the aspects of production, management, and use of digital information resources. This research aims at defining the nature of multicultural and multilingual terminology work and clarifying traditional standards and guidebooks for multilingual thesaurus construction and multicultural content management. A cross-domain approach will bring new perspectives to the research field and helps to develop the research analysis method by borrowing expressive and useful concepts and tools especially from translation and communication studies.

Research results will be useful in a large variety of different fields – from private and public sectors to the developers and actors in global Internet communication. The project has also novelty value describing a small language area’s culture and language bound problems in development and studies of multilingual indexing tools and vocabularies.

Acknowledgements

The author would like to acknowledge Sara von Ungern-Sternberg, Maria Forsman, Markku Mustaranta, Marianne Wikgren, Eeva-Liisa Eskola, Olle Persson and the blind reviewers for valuable comments and suggestions, Sami Borg and FSD, Ken Miller, Pam Miller, Alex Skinner, Lorna Balkan and UKDA for the generous access to their ELSST material and valuable comments.

References

Adler, N. (1997). *International Dimensions of Organizational Behaviour* (3rd ed.). Canada: South-Western College Publishing.

Aitchison, J. (1984). *Language Change: Progress or Decay?* (2nd ed.). Suffolk: Fontana Paperbacks.

Aitchison, J. & Gilchrist, A. (1987). *Thesaurus Construction: A Practical Manual* (2nd ed.). London: Aslib.

Bibexcel – a toolbox for bibliometricians. Version 2001-05-21. Developed by Olle Persson, Inforsk, Umeå University, Sweden. World Wide Web http://www.umu.se/inforsk/Bibexcel/index.html

Buckland, M. (1999). Vocabulary As A Central Concept In Library And Information Science. Preprint of paper published as "Vocabulary as a Central Concept in Library and Information Science" in: *Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science* (CoLIS3, Dubrovnik, Croatia, 23-26 May 1999). Ed. by T. Arpanac et al. Zagreb: Lokve, pp. 3-12.

Chan, L., Lin, X., & Zeng, M. (2000). Structural and Multilingual Approaches to Subject Access on the Web. *IFLA Journal*, Vol. 26(2000)3, pp.187-197.

Doerr, M. (2001). Semantic Problems of Thesaurus Mapping. *Journal of Digital information*, vol. 1 issue 8. Retrieved April 23, 2001 from the World Wide Web http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/

The European Union (EU) (2002). *Presentation of the thesaurus*. Retrieved February 6, 2002 from the World Wide Web http://europa.eu.int/celex/eurovoc/cgi/sga_doc?eurovoc_dif!SERVEUR/menu!prod!MENU&langue=EN

Forsman, M., & Keränen, S. (2000). Monikieliset tesaurukset ovat nyt ajankohtaisia. *Tietopalvelu* 15 (2000):5, pp.20-22. Helsinki: Tietopalveluseura.

Gollins, T., & Sanderson, M. (2001). Improving Cross-Language Information Retrieval with Triangulated Translation. In W. Bruce Croft, David J. Harper, Donald H. Kraft, Justin Zobel (Eds.), *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 90-95, September 9-13, 2001, New Orleans, Louisiana, USA: ACM.

Grefenstette, G., & Nioche, J. (2000). Estimation of English and non-English Language Use on the WWW. *RIAO'2000* (Recherche d'Informations Assistee par Ordinateur), Paris, April 12-14, 2000. Retrieved February 1, 2002 from the World Wide Web http://www.xrce.xerox.com/research/mltt/publications/Documents/P19137/content/RIAO2000gref.pdf

Greenberg, J. (2001). Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science*; 52 (6) Apr 2001, pp.487-98.

Haarala, R. (1981). *Sanastotyön opas*. Helsinki: The Research Institute for the Languages of Finland, Publication Series No 16.

Helsingin Sanomat (1.4.2001). Kieleni rajat ovat internetin rajat. Monikielisen internetin toteutuminen vaatii vielä vuosia. Retrieved February 5, 2002 from the World Wide Web http://www.helsinginsanomat.fi/uutiset/juttu_t.asp?id=20010401KU1

Hudon, M. (1997). Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge and concepts. *Information Services & Use*, Vol. 17 Issue 2/3, 1997, p. 111, 13 pp.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50-57), August 15-19, 1999, Berkeley, CA, USA: ACM.

Hörmann, H. (1986). *Meaning And Context. An Introduction to the Psychology of Language*. (Ed. and with an Introduction by Robert E.Innis.) New York: Plenum.

The Inter-university Consortium for Political and Social Research (ICPSR) (20.1.2002). *Announcements*. Retrieved February 5, 2002 from the World Wide Web http://www.icpsr.umich.edu/ORG/announce.html#thesaurus

International Standards Organization, ISO (1985). *Guidelines for the Establishment and Development of Multilingual Thesauri* (1st ed.). Geneva: ISO.

International Organization for Standardization, ISO (1986). *ISO International Standard 2788. Documentation - Guidelines for the establishment and development of monolingual thesauri*. Geneva: ISO.

International Organization for Standardization, ISO (1987). *Principles and methods of terminology*. Geneva: ISO.

Iivonen, M. (1989). *Indeksointituloksen riippuvuus indeksointiympäristöstä*. Tampere: University of Tampere, Department of Library and Information Science, Studies No. 26/1989.

Iivonen, M. (1995). *Hakulausekkeiden muotoilun yhdenmukaisuus onlineviitehaussa*. Tampere: University of Tampere, Acta Universitatis Tamperensis. Ser. A, vol. 443. (Includes summary in English)

Iivonen, M., & Kivimäki, K. (1998). Common Entities and Missing Properties: Similarities and Differences in the Indexing of Concepts. *Knowledge Organization* 25(3), pp. 90-102.

Itkonen, T. (1990). *Vierassanat. Kielenkäyttäjän opas*. Vaasa: Kirjayhtymä.

Jorna, K., & Davies, S. (2001). Multilingual Thesauri For The Modern World - No Ideal Solution? *Journal of documentation* 2001 57 (2) pp.284-295.

Leiwo, M. (2000). Suomen kieli 2000-luvulla: Voiko kielen kehitystä ennustaa. In Kalaja, P, & Nieminen, L. (eds.) (2000). *Kielikoulussa – kieli koulussa. AfinLAn vuosikirja 2000*. Jyväskylä: Suomen soveltavan kielitieteen yhdistyksen julkaisuja no. 58, pp. 387-404.

Miller, K., & Matthews, B. (2001). Having the Right Connections: the LIMBER Project. *Journal of Digital information*, vol. 1 issue 8. Retrieved August 27, 2001 from the World Wide Web http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Miller/

Miller, P. (2000). I say what I mean, but do I mean what I say? *Ariadne* Issue 23/March 2000. Retrieved January 23, 2002 from the World Wide Web http://www.ariadne.ac.uk/issue23/metadata/

Milstead, J. (1998). Use of Thesauri in the Full-Text Environment. Based on a paper presented at the 34th Clinic on Library Applications of Data Processing in Cochrane, P., & Johnson, E. (eds.) *Visualizing Subject Access for 21st Century Information Resources; Proceedings of the 34th Annual Clinic on Library Applications of Data Processing*, March 2-4,1997. Champaign, IL: Graduate School of Library and Information Science, University of Illinois, 1998. pp. 28-38. Retrieved March 23, 2001 from the World Wide Web http://www.jelem.com/useof.htm

Milstead, J., & Feldman, S. (1999). Metadata: Cataloging by Any Other Name … *ONLINE*, January 1999. Retrieved May 7, 2001 from the World Wide Web http://www.onlineinc.com/onlinemag/metadata/

Mulvany, N. (1997). What's going On in Indexing? Retrieved October 20, 2000 from the World Wide Web http://www.bayside-indexing.com/jcd.htm. Originally published in the ACM *Journal of Computer Documentation*, May 1997.

Nida, E., & Reyburn, W. (1981). *Meaning Across Cultures*. American Society of Missiology Series, No.4, New York: Orbis Books.

Nielsen, M. L. (2001). A framework for work task based thesaurus design. *Journal of Documentation*; 57 (6) Nov. 2001, pp.774-97

Oittinen, V., & Väyrynen, K. (2001). Englannin hegemonia ja humanistinen tutkimus. *Tieteessä tapahtuu* 8/2001 [*What happens in science*]. The Federation of Finnish Learned Societies. Retrieved December 19, 2001 from the World Wide Web http://www.tsv.fi/ttapaht/018/kesk.htm#kes

Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation*; 57 (3) May 2001, pp.330-48.

Saarti, J. (1999). *Kaunokirjallisuuden sisällönkuvailun aspektit. Kirjastoammattilaisten ja kirjastonkäyttäjien tekemien romaanien tiivistelmien ja asiasanoitusten yhdenmukaisuus.* [Aspects of Fictional Literature Content Description : Consistency of the Abstracts and Subject Indexing of Novels by Public Library Professionals and Clients.] Acta Universitatis Ouluensis. B, Humaniora, 33. Oulu: Oulu University. Accessible also via World Wide Web http://herkules.oulu.fi/isbn9514254767/ (Includes summary in English)

Schneider, J. W. (2001). *Thesaurus Construction by use of Bibliometric Methods: Theoretical Framework and Preliminary Methodological Steps*. A paper represented In Nordis – Net workshop: Selecting Theoretical Frameworks for Doctoral Research Projects. November 22 – 25 2001, Lithuania, Vilnius University, Faculty of Communication. Accessible also via World Wide Web http://www.kf.vu.lt/nordis/J_Schneider.doc

The Finnish Standards Association, SFS (1988). *Suomenkielisen tesauruksen laatimis- ja ylläpito-ohjeet.* SFS 5471. Helsinki : Suomen standardisoimisliitto.

The Social Science Information Gateway, SOSIG (2000). *Searching by Thesauri*. Retrieved February 6, 2002 from the World Wide Web http://sosig.ac.uk/help/thesaurus.html

Sperer, R., & Oard, D. W. (2000). Structured Translation for Cross-Language Information Retrieval. In Nicholas J. Belkin, Peter Ingwersen, Mun-Kew Leong (Eds.): *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 24-28, 2000, Athens, Greece. ACM.

Suojanen, P. (1993). Identiteetti, perinne, vallan viestit. In Suojanen, P. & Suojanen, M. K. 1993: *KULTTUURIN KALEIDOSKOOPISTA. Kirjoituksia kielestä ja kulttuurista*. Kangasala: Antrokirjat.

The Finnish Centre for Technical Terminology, TSK (ed.) (1989). *Sanastotyön käsikirja. Soveltavan terminologian periaatteet ja työmenetelmät*. (TSK 14, SFS-handbook 50.) Jyväskylä: The Finnish Centre for Technical Terminology (TSK).

Tudhope, D., Alani, H., & Jones, C. (2001). Augmenting Thesaurus Relationships: Possibilities for Retrieval. *Journal of Digital information*, vol. 1 issue 8. Retrieved April 5, 2001 from the World Wide Web http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Tudhope/

Varantola, K. (1990). *Tekniikan suomi yhdentyvässä Euroopassa. Sanastotyön merkitystä koskeva selvitys.* Helsinki: The Finnish Centre for Technical Terminology (TSK).

Wierzbicka, A. (1997). *Understanding Cultures Through Their Key Words. English, Russian, Polish, German, and Japanese*. New York: Oxford University Press.

von Ungern-Sternberg, S. (1994). *Verktyg för planering av tvärvetenskaplig informationsförsörjning. En tillämpning på ämnesområdet bioteknik i Finland*. Åbo: Åbo Akademis tryckeri. (Includes summary in English.)

von Ungern-Sternberg, S. (1998). Knowledge organisation and a macro language for indexing in biotechnology. *Proceedings of the 6th ASIS SIG/CR Classification Research Workshop* held at the 58th ASIS Annual meeting Chicago, Illinois, October 8, 1995. Ed. by Schwarrz R. P. ASIS Monograph Series Information Today, Medford, 1998. Advances in Classification Research 6, pp.181-194, 1998.

Biography

Susanna Keränen, M.Soc.Sc. 2000 Tampere University, since 2001 a doctoral student in the Åbo Akademi University, Department of Social Sciences / Information Studies. She currently works as a researcher in the project *Cultural and linguistic differences in digital storage and retrieval of information* (see World Wide Web http://www.abo.fi/instut/diginfo/index.html). Before that she worked as a planning officer in the Finnish Social Science Data Archive (FSD) in a multilingual thesaurus project (8/2000-5/2001). This research aims at doctoral thesis in information studies at the Åbo Akademi University, Department of Social Sciences / Information Studies. It is supervised by Sara von Ungern-Sternberg and financed by the Academy of Finland (targeted programme on the Production, Management and Use of Digital Information Resources, 2001-2004) and by the Finnish Cultural Foundation.