IMPLICATIONS FOR INDEXER THESAURUS CONSTRUCTION AND DESIGN: EXAMPLE OF LIS SUBJECT FIELD

(PhD forum) for LIDA 2012

Kristina Feldvari
Filozofski fakultet u Osijeku, L. Jagera 9, 31000 Osijek. Email: kfeldvari@ffos.hr.

Abstract
Before thesaurus construction commence, some preliminary considerations concerning its purpose and scope must be made. The thesaurus compiler should have thorough knowledge of the information environment and the discourses and languages in play. Next to the characteristics of the information environment, thesaurus should also reflect the mental models and search behavior of the individuals acting in the “discourse community”. Likewise, it is important for thesaurus compiler to examine indexing policy and practice. The analysis of the existing subject indexing practice and policy will be explored by measuring inter-indexer consistency in subject library catalogues and by qualitative comparison of indexing practice with IFLA Principles Underlying Subject Heading Languages (1999). Thesaurus users or “discourse community” examined in this dissertation are indexers in Croatian libraries, and the example of subject field is LIS. This means that there is one central “discourse” that we will examine and that is “indexing discourse”. When analyzing different discourses it is essential to be aware of the specific contexts and institutions in which the discourses take place. Therefore, the emphasis is put on institutional differences (different subdiscourses). The “indexing discourse” will be explored by interviewing indexers from five libraries in Croatia, by comparison of existing internal thesauri and by conducting a simulated indexing task. Mental models of indexers will be explored by word association method. Results of the analysis of present situation of subject indexing in Croatia and “domain analysis” should provide the knowledge needed for defining implications for future indexer thesauri construction and design.

Keywords: thesaurus construction and design, indexing discourse, subject indexing

Content of the contribution

Introduction
A thesaurus is an example of a controlled vocabulary and an important aid in subject analysis. It also plays an important role in the organization of information, controls the vocabulary and is formed in a way that facilitates seeking and indexing within a specific subject area. As an aid in subject analysis, thesaurus is a set of items, phrases or words, with a set of relations between them (semantic and syntactic). In today’s world of technology, especially in e – librarianship, thesaurus is of tremendous importance and its primary role is in information retrieval (IR) (Aitchison, Gilchrist & Bawden, 2004; Soergel, 1974). The major purpose of thesaurus is to provide a map of a given field of knowledge, indicating how concepts and ideas about concepts are related to one another, to help indexers and searches to understand the structure of the field and to guide the users so that they choose appropriate index or search terms (Lykke Nielsen, 2000).

There are two types of thesauri. The first are general – purpose thesauri which are word – based and rarely used in IR systems. The second are IR – oriented or structured thesauri which can be divided into two types: end – user thesaurus and indexer thesaurus. Most thesauri in use today are indexer thesauri, which means that they are designed primarily for indexer use and only secondarily, intended for the use by the end – user (Bates, 1986). According to Aitchison et al. (2004) the primary purpose of a thesaurus is for information retrieval, which may be achieved in various ways. Furthermore, the primary use may be achieved by using the thesaurus in the indexing of a database, and/or in its searching, by various ways: 1) thesaurus used both in indexing and in searching; 2) thesaurus used in indexing, but not in searching; 3) thesaurus used in searching, but not in indexing.

When discussing IR thesauri, there are three basic issues that we have to keep in mind: thesaurus construction (manual or automatic), access (thesaurus must be accessed and used in some way to improve or expand users’ query) and thesaurus evaluation (quality and efficiency) (Jing & Bruce Croft, 1994).
Before thesaurus construction commence, some preliminary considerations concerning its purpose and scope must be made. Although thesaurus construction follows the general principles and recommendations outlined in numerous standards and guidelines established for this purpose (e.g. ANSI/NISO Z39.19-2005, 2005; Aitchison et al., 2004; ISO 2788, 1986), result of thesaurus construction process depends on several factors. Construction of a subject specific thesaurus must take into consideration the needs of the information system and the users in a domain (i.e. knowledge and understanding of the information environment) (e.g. Hjørland, 1997; Jacob & Shaw, 1998; Lykke Nielsen, 2000, 2001; Pejtersen, 1980; Pjetersen & Fiedel, 1998; Soergel, 1974, 1985). Likewise, most guidelines (e.g. Aitchison et al., 2004) also recommend that thesaurus construction should be based on next considerations: user group, subject field, quantity and type of literature along with language considerations, resources available etc. The thesaurus construction also depends on indexing policy and practice and decision concerning the two cannot be studied separately (Lykke Nielsen, 2000).

**Background of the study**

With the appearance of World Wide Web it has become apparent that the more digital libraries, data bases, portals etc. are developed on the net, the more important subject analysis and thesauri become in order to aid these searches (Hert, Jacob & Dawson, 2000; Shiri & Revie, 2000). Despite the obvious fact that thesauri are needed to aid indexing and searching, methodology of its construction is also important. According to Miller (2003b) only a methodologically accurately constructed thesaurus essentially provides certain significant advantages to the whole retrieval system, which it supports.

The process of thesaurus construction traditionally consists of the collection of concepts and terms within a given subject area, formation and definition of concepts and terms (terminological and semantic control) and finally organization of concepts and terms (distinguishing and displaying relationships between them) (Aitchison et al., 2004; Lykke Nielsen, 2004; Soergel, 1974). While conducting these technical processes thesaurus constructor must apply principles and guidelines for thesaurus construction. On the other hand, the existence of variations in nature and quality of different thesauri brings impact of the guidelines on actual thesaurus construction into question (Gilchrist, 1991). This is one of the reasons why thesaurus constructor primarily has to study and examine information environment and apply the guidelines according to “knowledge domain” or “discourse community” and its characteristics. The thesaurus should be developed according to the nature and needs of the information environment in which it is going to be used. This implies a thorough knowledge of the information environment and the discourses and languages in play. This knowledge may be gathered by different methodologies. However, it is generally accepted to take a “holistic approach” and study the information environment and its users as an interacting whole (Lykke Nielsen, 2004). When analyzing different discourses it is essential to be aware of the specific contexts and institutions in which the discourses take place (Talja, 1999).

Two types of knowledge are needed for thesaurus construction – “domain knowledge” and “conceptual knowledge”.

“Domain knowledge” covers knowledge about the environment and the situation in which thesaurus is going to be used, knowledge about work tasks, knowledge about information use and information needs, and knowledge about different approaches to the subject field. “Conceptual knowledge” is linguistic, semantic knowledge about naming, form, meaning and relations between concepts and terms and it allows choosing, controlling, and structuring the vocabulary of thesaurus (Lykke Nielsen, 2000, 2001). Next to the characteristics of the information environment, thesaurus should also reflect the mental models and search behavior of the individuals acting in the “discourse community”. Even though individuals have been formed by the domain, they have also been shaped socially and will approach the knowledge domain from a perspective which depends both on their actual situation and on their individual characteristics (Allen, 1997; Bates, 1998; Furnas, Landauer, Gomez & Dumais, 1987; Ingwersen, 1996; Lykke Nielsen, 2000, 2001). This dissertation will be designed using that approach, trying to capture individuals as well as collective information behavior. As early mentioned, Lykke Nielsen (2000) and Jacob & Shawn (1998) state that thesaurus construction also depends on indexing policy and practice that also have to be examined carefully.

Motivated by these last statements and considering situation of the indexing policy and practice in Croatia, the intention of this study is to detect problem areas in Croatian subject indexing and emphasize the need for creation of national indexing rules and common indexing aids such as thesaurus. The overall aim of the research presented in this dissertation is to give implications and to establish firm basis for indexer thesaurus construction and design.

According to Williamson (1996), and in accordance with Lykke Nielsen (2000), standards, guidelines, and rules are needed to ensure consistency and quality in the design and construction of indexing language. Specifically, in

---

1In this study knowledge domain or discourse community is understood according to Hjørland (1997) as a scientific, scholarly or professional domain with unique structures of communication and publication, unique types of documents, specific terminology, information structures, information systems, literature and relevant criteria.

2Different domains can also be seen as different subcultures (e.g. indexers) having their own discourse (Nykyri, 2010)
Croatia there are no national rules for subject indexing and every library uses its own internal rules and its own internal thesaurus. Usage of different rules and indexing aids in every library causes problems like inconsistency with international indexing principles and guidelines and low inter – indexer consistency (Doležal, 2006; Dominković, 2008). Few papers explained the need for re – examining the existing subject indexing practice and policy in Croatia by taking a clear stand regarding the principles, guidelines, standards etc. and their applicability in Croatian subject indexing practice (Doležal, 2006; Dominković, 2008; Mikacic, 1991, 1996; Strbac & Vujić, 2004). The work on the national subject indexing policy should be intensified providing detailed instructions and guidelines for subject indexers, so it is primary essential to conduct a comprehensive analysis of the present situation and future developments of subject indexing in the libraries in Croatia. The analysis of the existing subject indexing practice and policy will be explored by interviewing representatives of subject indexing departments, by measuring inter – indexer consistency in subject library catalogues and by qualitative comparison of indexing practice with IFLA Principles Underlying Subject Heading Languages (1999).

Next to the analysis of present situation of subject indexing and prior the thesaurus construction it is essential to examine “discourse community”. Thesaurus users or “discourse community” examined in this dissertation are indexers in Croatian libraries, and the example of subject field is LIS. This means that there is one central “discourse” that we will examine in this dissertation and that is “indexing discourse”3. As previously mentioned, when analyzing different discourses it is essential to be aware of the specific contexts and institutions in which the discourses take place. For instance, guidelines for indexing used in one library presumably differ from guidelines in other library due to their function, environment and conditions of use. Therefore, the emphasis is put on institutional differences (different Croatian libraries). The “indexing discourse” will be explored by interviewing indexers from five libraries in Croatia, by analysis of existing internal thesaurus and by conducting a simulated indexing task. Mental models of indexers will be explored by word association method.4

Results of the analysis of present situation of subject indexing in Croatia and “domain analysis” should provide the knowledge needed for defining implications for indexer LIS thesaurus construction and design, and consequently provide guidelines for future indexer (domain specific) thesauri construction and design.

**Previous research, knowledge and literature review**

Since the topic of this dissertation is the thesaurus construction, special emphasis is put on current literature and knowledge of process of thesaurus construction. The process of thesaurus construction is traditionally divided into three technical stages: collection of concepts and terms, formation and definition of concepts and terms and organization of concept and terms (Aitchison et al., 2004; Lykke Nielsen, 2004, Soergel, 1974). These stages are well described in various guidelines and standards for the construction of monolingual and multilingual thesauri (e.g. ANSI/NISO Z39.19-2005, 2005; Aitchison et al., 2004; ISO 2788, 1986).

Guidelines and standards may suggest different instructions for thesaurus construction, e.g. alternative types of relationship, but it is domain – related factors that make it possible to choose among the alternatives. Therefore, the thesaurus compiler should have thorough knowledge of the information environment so he could be able to determine purpose, scope and implications for thesaurus construction. To gain knowledge about information environment it is essentially to conduct “domain analysis”.

During the eighties a user-centered approach to thesaurus construction emerged, which introduced strategies that based the development of indexing systems, including thesauri, on field studies of the users’ information use and information needs. The thesaurus construction process was considered as a system design process, and focus was on information systems and thesauri supporting people in their daily activities. Pejtersen (1980) was one of the first to base the development of an indexing system on empirical field studies of users' request and together with Fidel he developed a framework for work – centered evaluation and design of information systems (Pejtersen & Fidek, 1998). Soergel (1985) also recommends the study of information needs as a basis for information system design. Later, the user-centered approach turned into a holistic view of all of the interactive communication processes, stressing the importance of the situational context that surrounds the user. This development of the user-centered approach is called the “person-in-situation” approach or the holistic cognitive approach (Allen, 1997; Ingwersen, 1996, 2001; Lykke Nielsen, 2000; Kuhlthau & Vakkary, 1999; Saracevic, Kantor, Chamis & Trivison, 1988.; Wilson, 1999). The user-centered and the cognitive approaches have been criticized for being defective in being based solely on studies of individual users detached from their context. It is argued that the individual user’s knowledge structures and behavior are shaped through participation in socially grounded domains. The critics stress the importance of seeing the individual user as member of a particular knowledge domain. Individuals do not operate in isolation but work,

---

3 At general level discourse can be defined as referring to language use in social context but it also means doing something and it is therefore an active process (Potter& Wetherell, 1990).

4 About methodes see chapter Methodology and research design
instead, within the dynamical contexts provided by complex socio-cultural and historical systems (Hjørland, 1997, 2002; Hjørland & Albrechtsen 1995; Jacob & Shaw, 1998; Talja, Heinisuo, Luukkainen & Järvelin, 1997). Ingwersen (1996) and Wilson (1999) illustrate how searches are bound by different kinds of barriers. Some barriers are personal, others are role related and others are related to the interpretation of specific environment in which the searcher is acting.

It seems to be generally accepted that the study of information behavior must take into account the users, their particular role or function and, especially, the information environment that influences information seeking and use. In this study, users of thesaurus are indexers in Croatia, so we examine “indexing environment” (indexing work tasks, indexing behavior, indexing needs, resources, aids, problems etc.) (Bates 1998; Bertrand, Cellier & Giroux, 1996; David & Giroux, 1995; Farrow, 1991; Lancaster, 2003; Mai, 2005; Shauperl, 2004 etc.). To gain semantic knowledge methodology of corpus analysis (lexico-semantic), word association test and discourse analysis are going to be used (López – Huertas, 1997; Lykke Nielsen, 2000; Lykke Nielsen & Ingwersen, 1999; Nykyri 2010; Talja, 1999).

**Objective and hypotheses**

The overall aim of the research presented in this dissertation is to give implications and to establish firm basis for indexer thesaurus construction and design.

Specific objectives of this study are:

- to give ‘the state of the art’ situation of subject indexing in Croatian libraries
- to conduct “domain analysis”
  - by analyzing Croatian indexing environment (indexing work tasks, indexing behavior, indexing needs, indexing aids, indexing resources, indexing problems etc.) with the emphasis on institutional differences
  - by analyzing internal thesauri (focus on LIS field) in Croatian libraries

Implications for thesaurus construction are examined through attempts to find the answer for the following research questions:

**Indexing policy and practice analysis**

1. Are there any differences in indexing practice in Croatian libraries? If there are, what are they?
2. What affects the differences in indexing practice in Croatian libraries?

**Domain analysis (LIS subject field)**

3. In what way can the analysis of Croatian indexing environment (indexing work tasks, indexing behavior, indexing needs, indexing aids, indexing resources, indexing problems etc.) contribute to defining implications for indexer LIS thesaurus construction?
   - What are the differences and similarities on institutional level?
4. In what way can the indexer thesaurus be based on the indexing discourse?
   - Are there any conceptual and terminological differences (considering equivalent, hierarchical and associative semantic relations) in LIS subject field in internal thesauri in Croatian libraries?
5. Will the methodology used for obtaining implications for LIS indexer thesaurus provide the necessary knowledge for a future construction and design of an indexer thesaurus in general?
**Hypotheses**

1. Differences in indexing practice in Croatian libraries are present and visible in subject library catalogues. Inter – indexer consistency between libraries in Croatia is low.

2. Usage of different internal indexing rules and different internal indexing aids as well as noncompliance to international indexing rules, affects the differences in indexing practice in Croatian libraries.

3. Analysis of Croatian indexing environment (indexing work tasks, indexing behavior, indexing needs, indexing aids, indexing resources, indexing problems etc.) can contribute to defining implications for structure and design of indexer LIS thesaurus

4. Discourse analysis method, word association method and corpus analysis (lexico-semantic) help to build up the semantic knowledge and to define implications for structure and terminology of indexer LIS thesaurus.

5. The methodology used for obtaining implications for LIS indexer thesaurus will provide necessary knowledge (domain and conceptual) for a future construction and design of an indexer thesaurus in general

**Methodology and research design**

The research questions, the methods used and the material are summarized in the table below:

**Table 1**

Summary of the research questions, methods and material

<table>
<thead>
<tr>
<th>Research question</th>
<th>Data Collection and Analysis Methods</th>
<th>Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are there any differences in indexing practice in Croatian libraries? If there</td>
<td>Experimental quantitative methodology to investigate inter – indexing consistency (relationship between</td>
<td>50 documents from in 5 Croatian libraries</td>
</tr>
<tr>
<td>are, what are they?</td>
<td>indexing consistency/exhaustivity)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Qualitative comparison with IFLA guidelines</td>
<td></td>
</tr>
<tr>
<td>What affects the differences in indexing practice in Croatian libraries?</td>
<td>Semi-structured interview</td>
<td>Representatives of subject indexing departments of 5 Croatian libraries</td>
</tr>
<tr>
<td>In what way can the analysis of indexing environment (indexing work tasks,</td>
<td>Focused interview</td>
<td>Indexers from 5 Croatian libraries</td>
</tr>
<tr>
<td>indexing needs, indexing aids, indexing resources, indexing problems etc.)</td>
<td>Simulated indexing task method (three complementary methodological approaches : analysis of concurrent</td>
<td>5 documents (LIS field) indexed by indexers from 5 Croatian libraries</td>
</tr>
<tr>
<td>contribute to defining implications for indexer LIS thesaurus?</td>
<td>verbal report (think aloud method), retrospective verbalization, and an interview involving peer evaluation</td>
<td></td>
</tr>
<tr>
<td>§What are the differences and similarities on institutional level?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In what way can the vocabulary of the indexer thesaurus be based on the indexing</td>
<td>Word association test</td>
<td>Indexers from 5 Croatian libraries</td>
</tr>
<tr>
<td>discourse?</td>
<td></td>
<td>Results of focus interviews,</td>
</tr>
</tbody>
</table>
Are there any lexical and semantic differences (considering equivalent, hierarchical and associative semantic relations) in internal thesauri in Croatian libraries in LIS subject field? | Discourse analysis | simulated indexing task and co-word analysis |
---|---|---|
Corpus analysis (lexico-semantic) | Internal thesauri (focus on LIS field) in 5 Croatian libraries |
Co-word analysis | |

Will the methodology used for obtaining implications for LIS indexer thesaurus provide the necessary knowledge for a future construction and design of an indexer thesaurus in general? | Conclusion | Theoretical discussion + empirical studies |
---|---|---|

This study uses quantitative and qualitative methods. Two different methodological approaches are needed in order to answer the research questions of this study. Quantitative and qualitative methods are thus used to supplement each other and to give answers partly to the same and partly to different research questions. In its methodological nature the whole study is a qualitative study. Qualitative methods are used to examine the discourses and the nature of the problems occurring in the research material. Since the problems are partly overlapping and complex ones, several methods are also used simultaneously. A combination of multiple search methods will reveal a more varied and valid picture of the research object (Fidel, 1993; Vakkari, 1997). The material should allow multiple strategy approaches and represent different types of subdiscourses, which should be taken into consideration when constructing an indexer thesaurus.

The first and the second phase of research will refer to investigation of indexing practice in Croatian libraries. The analysis of the existing subject indexing practice and policy will be explored by measuring inter-indexer consistency in subject library catalogues and by qualitative comparison of indexing practice with IFLA Principles Underlying Subject Heading Languages (1999).

Second stage will include interviewing representatives of subject indexing departments of five Croatian libraries to investigate what affects the differences in indexing practice in Croatian libraries. Studied libraries are: National and University library in Zagreb, The Zagreb City Libraries, Public and University library in Osijek, Public Library Metel Ožegović in Varaždin and Library of the Faculty of Philosophy in Zagreb. These libraries were selected because of four different library information systems used there: Aleph, Crolist, Zaki, Metel and Koha. Furthermore, different primary user – groups and their information needs may affect the indexing policy and practice and our sample contains libraries of various types. There are two public libraries (The Zagreb City Libraries, Public Library Metel Ožegović), one academic library (Faculty of Philosophy in Zagreb) and two libraries with dual function (National and University Library in Zagreb and City and University Library in Osijek).

The third and the fourth phase of research will refer to “domain analysis”. Focused interviews with Croatian indexers will be used to gain information about environment in which thesaurus is going to be used (indexing work tasks, indexing behavior, indexing needs, indexing aids, indexing resources, indexing problems etc.). As a part of interview a simulated indexing task method and word association test will be used. Interviewing and simulated indexing task mostly provide insight into the context of indexers' information needs and behavior. The quantitative word association method is also used as a part of the interview method in order to obtain information about the studied terms semantic network (semantic knowledge) and as a starting point for further discussion about the terms meaning and usage. It reveals respondents' mental models and provides knowledge about respondents' way of relating terms and their use (Lykke Nielsen & Ingwersen, 1999). To study five different internal thesauri and to gain semantic knowledge (from LIS subject field) corpus analysis (lexico-semantic) and co-word analysis will be used. In this study, co-word analysis is used to study indexing frequency and the lexical network of the indexing terms. The final aim is to see how these practices diverge and how they can be used for thesaurus construction and design.

One of the methods used is discourse analysis (Hjørland, 2002; López – Huertas, 1997; Talja, 1999). In this study discourse analysis is understood not only as a method, but as a theoretical framework, that allows different emphasis
in the focus and methodological applications. The discourse analytic approach offers a possibility to study differences and similarities both on the individualistic and community level. Discourse analysis studies practices of producing knowledge and meanings in concrete contexts and institutions, it pays attention to the way in which discourses produce and transform social reality, and makes it possible to evaluate the practical consequences of different ways of approaching a particular phenomenon (Talja, 1999).

Expected contribution

The aim of this study is to give thorough ‘state of the art’ analysis of subject indexing in Croatian libraries. At present, Croatian libraries have neither national indexing rules nor common indexing aids. Therefore, the intention of this study is to detect problem areas in Croatian subject indexing and emphasize the need for creation of national indexing rules and common indexing aids such as thesaurus. This study offers implications for construction and design of indexer LIS thesaurus as an example for future thesaurus construction and design.

References and Citations


Curriculum Vitae

Education

PhD student in information sciences "Društvo znanja i prijenos informacija (Faculty of Philosophy, University of Zadar, 2009)

Graduated in history and librarianship (Faculty of Philosophy, University of Osijek, 2004-2008.)

Publications

2010.
