Marianne Lykke Nielsen  
Royal School of Library and Information Science, Copenhagen, Denmark

Anna Gjerluf Eslau  
H. Lundbeck, Valby, Denmark

Corporate Thesauri – How to Ensure Integration of Knowledge and Reflection of Diversity

Abstract: This paper evaluates and compares three thesaurus construction methodologies: literary scanning, word association tests, and involvement of subject expert groups. The evaluation concentrates on exploring advantages in relation to the sub-processes: collection, formation and structuring of concepts and terms. Quantitative as well as qualitative analyses have been carried out. The analysis shows that the methods are complementary each providing distinct conceptual information from respectively a domain-oriented and a scientific viewpoint. The combination of methods provides a thesaurus, at the same time, mapping authoritative language and reflecting the diversity of language.

1. Introduction

Basically, a corporate thesaurus serves as a mechanism to facilitate communication, improve learning and support information retrieval across the organisation. Recently, in relation to knowledge management and implementation of corporate intranets, portals and document management systems, there has been an increased interest in conceptual tools like thesauri providing "... a semantic road map to individual fields and the relationships among fields, to relate concepts to terms, and to provide definitions ..." (Soergel, 1996, p.167). Business corporations consist of divisions and subsidiaries situated in various locations, nationally and internationally, with staff coming from different knowledge domains, having different cultural, social, educational and professional backgrounds. In large organisations there exists a range of discourse communities, each of which having specific tasks and objectives, generating their own jargons (Cabre, 1999; Morgan, 2000, Gilchrist and Kibby, 2000).

In corporate information environments, conceptual tools like thesauri should describe the languages in play and bridge the boundaries between the individual user and the set of jargons. The compiler of such tools faces and must consider ambiguity problems like subject crossover and overlap, interdisciplinary borrowings, disciplinary or domain-specific focus, cultural differences, linguistic differences demanding thorough knowledge about the corporate language use (Gilman, 1997).

The present paper compares thesaurus construction methods, based on literary warrant, with methods based on user warrant with the aim to develop a framework for thesaurus construction and maintenance combining the two approaches. The paper is structured as follows: Section 2 discusses the conditions...
and problems related to thesaurus construction. Section 3 describes the purpose of the research project and the evaluation method, and Section 4 discusses the empirical findings. The concluding Section 5 outlines a framework for thesaurus construction combining literary- and user-oriented construction methods.

2. Thesaurus construction

Thesaurus construction involves three primary processes: collection, formation and structuring of concepts and terms. Principles for each sub-process have been developed and embodied in a variety of standards and guidelines (see e.g.: Soergel, 1974; International organization for Standardization, 1986; Aitchison, Gilchrist & Bawden, 2000). Particular problems are related to each process, and the guidelines suggest alternative solutions. However, it is often domain-related factors, which make it possible to choose among the alternatives. This implies a thorough knowledge of the information environment and the discourses and languages in play.

Professional jargons are specialized languages, based on and derived from general language, presupposing special education and restricted to communication among specialists in the same closely related fields (Cabre, 1999). Professional jargon is characterised of being well defined with little ambiguity of meaning and use. Ideally, each term of the vocabulary refers to one concept, and concepts are named by only one term. Relationships among concepts are clear and well defined. However, within the same organisation several jargons will be at work, and it may be difficult to capture, separate and describe the domain-specific languages in play and present them in an understandable form. Also “double-coded terms”, terms adapted from everyday vocabulary to professional, specialized languages with a changed meaning, may cause problems as corporations consist of professionals approaching the subject domain with different levels of perspective and subject knowledge (Haas and Hert, 2000).

The thesaurus compiler face several problems. Concerning collection it is difficult to define the topical focus and decide what concepts and terms to include. Decisions regarding form pose other problems: grammatical forms, choice of preferred terms, and concept definitions. Concerning structuring it is difficult to capture perspectives of the organisation and relate concepts and terms so that the relationships provide conceptual understanding of use, mapping subtle, but important distinctions and relationships. Many concepts are poly-hierarchical, and deciding what viewpoints to reflect by the hierarchies is difficult avoiding both information overload and mislead. The level of specificity of hierarchies constitutes another important problem.

Sources for collection of conceptual information for thesaurus construction are normally divided in written, literary sources and unwritten sources, for instance, knowledge obtained from experts and users by interviews, workshops, etc. Each methodology provides distinct information, supports the thesaurus compiler differently, and involves different competences and resources processing the information collected. Standardised sources require less effort in gathering and collecting well-defined disciplinary-oriented and authoritative information. Unwritten, non-standardized sources make possible collection of a set of terms reflecting the information environment actual usage from a more current viewpoint (Soergel, 1974).
3. Purpose and evaluation method

The present research project aims at evaluating methods respectively based on literary warrant and on user warrant. The hypothesis behind the project is that nature of conceptual information gathered by literary sources and human sources differ as regards topical perspective, specificity and resources for processing. The project focuses on investigating the nature of data collected in relation to the three primary processes in thesaurus construction: collection, formation, and structuring. The information collected is compared and strength and problems of the two (literary and human) sources to conceptual information is evaluated.

The analysis includes the following methods: a) scanning of internal or external literary sources; b) word association tests, and c) cooperation with groups of subject experts (Soergel, 1974, Aitchison, Gilchrist and Bawden, 2000, Lykke Nielsen, 1997).

The three methodologies under investigation have been used for thesaurus construction in a real-life thesaurus design project in a large research and development company within the pharmaceutical industry. The methods have been applied and evaluated according to findings of a domain study exploring purpose and role of the corporate thesaurus. According to the domain study the objective of the thesaurus is twofold. The thesaurus shall support information retrieval by referring to retrieval synonyms (strongly related terms appropriate for query expansion), and the thesaurus shall support learning and knowledge exchange providing insight into the specific jargons used within the corporation. The thesaurus should mediate the preferred corporate language and the internationally recognized scientific and authoritative language, but the thesaurus should also expose the diversity and vocabulary developments over time (Lykke Nielsen, 2001).

The literary scanning aimed at identifying and comparing vocabularies recognised within the corporation and internationally. Primarily, glossary lists were revised. Concepts were picked in order to cover the standardised vocabulary, and preferred terms and definitions were chosen to reflect the scientific, authoritative language use. The first version of the corporate thesaurus was constructed on basis of the literary scanning.

The purpose of the word association tests was to gather retrieval synonyms and capture term relationships reflecting the viewpoint of the work domain. A primed, continued test methodology was used in order to capture as many relationships as possible. The respondents, representing four information-oriented work task groups within the area of research and development were asked (primed) to make associations in relation to their work tasks and frame of interest. A total of 100 descriptors from the literary-based thesaurus were presented to the thirty-five respondents as stimuli, providing a total of 2885 associative responses. Twenty stimuli words, of high relevance for the work domain, were presented to the full set of respondents, the rest of the stimuli were only presented to a single work task group.

Four expert groups representing the main business areas were established to support the thesaurus manager in identifying the relevant topical areas, in choosing preferred terms and evaluating the structure as regards specificity and relationships. The members were selected according to work domain experience and decision competence concerning vocabulary control. Formal group meetings were held regularly, and experts were consulted ad hoc by e-mail or phone.
Sub-processes of thesaurus construction: | Issues related to thesaurus construction:
---|---
Collection | Topical coverage  
| Domain orientation  
Formation | Grammatical forms  
| Selection of preferred terms  
| Definition  
Structuring | Type of relationships  
| Specificity  
| Synonym variations

Table 3.1. Evaluation criteria

The evaluation was carried out in three steps. First a quantitative comparison was made measuring the overlap as regards relationships between the literary-based thesaurus records and the 100 records based on the word association tests. The aim of this analysis was to evaluate whether relationships identified by the two methods differ. The second evaluation was qualitative, analysing the nature of the terms and relationships discovered by the word association tests, hereby investigating the differences between the literary scanning and the word association method. In the third step the conceptual information provided by the expert groups was compared to the other methods. The comparison was founded on the evaluation criteria presented in Table 3.1.

4. Discussion of empirical findings

Concerning collection the literary scanning of external sources provided a broad, general set of terms requiring subsequent selection of domain-relevant concepts. The scanning of internal documentation provided a more domain-oriented subset of concepts. 54% of the literary-based term relationships established in the first version of corporate thesaurus were also elicited by the word association tests (see Table 4.2), indicating that the two methodologies collect a high amount of similar conceptual information. However, the word association method elicited more up-to-date conceptual information reflecting the latest research and present focus of the corporation. Many relevant concepts were provided by the word association tests, confirmed by the fact that 92% of the new, unique terms (397 terms), discovered by the word association method, were later included in the corporate thesaurus. Neither of the methods mentioned provided a complete, exhaustive picture, which became clear when expert groups were revising the thesaurus. Expert groups have an important role controlling whether the thesaurus covers the viewpoint of the corporation. The word association method also provided a considerable amount of useless information. In the present project 31% of the word associations to the 20 stimuli qualitatively evaluated, were considered useless: word associations not possible to verify, broad concepts out of context, too specific concepts, synonym variations in other languages or curious translations, and phrases. In a professional environment this kind of terms is not useful as retrieval synonyms.
Concerning formation, the word association tests as well as the expert group evaluation revealed a pattern of grammatical forms for individual concepts, but it was not possible to identify a general pattern, and thesaurus construction guidelines have been used for the corporate thesaurus. As regards choice of preferred terms, the literary-based thesaurus followed international vocabularies selected by the expert groups. However, the later expert group evaluation revealed that the literary sources were sometimes obsolete as regards current naming. An analysis of word associations showed that the frequent word associations (associated by three or more respondents) tend to follow the preferred terms found by literary warrant, on average 58%, indicating that choice of preferred terms might be based on word associations, especially because the word associations furthermore reflected up-to-date language use. Definitions are an important part of a corporate thesaurus supporting learning and knowledge exchange. In a highly regulated environment like the present, it is natural to base definitions on authoritative and scientific literary sources. However, the word association tests revealed domain-specific knowledge, making it possible to work up more domain-oriented definitions compared to the generic, “neutral” definitions found in the literature. From the word associations, for example, given to the stimuli interactions, the thesaurus manager became aware of the connection between interactions and cytochrome P-450. The word associations inspired her to investigate further the relationship between interactions and cytochrome P-450. The investigation had as result that the concept cytochrome p-450 has been included in the corporate thesaurus, and that the definitions have been revised according to the findings (see Table 4.1).

Concerning structuring the hierarchical relations was best found by the literary scanning. International sources, like the MESH thesaurus, provided exhaustive and well-defined hierarchies for many of the concepts covered by the corporate thesaurus. For other concepts no appropriate sources existed, and therefore expert groups played an important role in the hierarchical structuring. Some domain-specific concepts, e.g., corporate drug products, require hierarchies tailored to the context and may only be developed by domain experts. Concerning degree of specificity of hierarchies, the preliminary domain study identified the most important concept categories to be represented with a high degree of specificity. The expert groups also provided useful information about specificity. Neither the word association tests nor the literary scanning provided such information. The literary sources did not reveal a high amount of associative relations, whereas the majority of the word associations consisted of related terms. The qualitative analysis showed that the total number of BT was 10%, the number of NT is 8%, the number of synonyms 5%, and the number of RT 46%. The relationships identified by the word association tests covered 54% of the relations established from the literature scanning (see Table 4.2 for coverage of literary-based relationships).
Table 4.1. Revision of definitions according to findings of word association tests.

<table>
<thead>
<tr>
<th>Type of relationships</th>
<th>Coverage %</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT - Broader terms</td>
<td>68</td>
</tr>
<tr>
<td>NT - Narrower terms</td>
<td>52</td>
</tr>
<tr>
<td>RT - Related terms</td>
<td>52</td>
</tr>
<tr>
<td>SYN - Synonyms</td>
<td>45</td>
</tr>
<tr>
<td>ALL - All relationships</td>
<td>54</td>
</tr>
</tbody>
</table>

Table 4.2. Literary-based relationships covered by the word associations for the 100 stimuli.

These findings show that a high amount of similar conceptual information is collected by the two methods compared. In addition, a large amount of unique associative relations not discovered by the literary scanning was elicited by the word association method. These relationships played an important role as mediator of up-to-date, domain-oriented knowledge. Both literary scanning and the word association method identified a number of equivalence relationships. The expert groups were used to revise the synonyms according to the domain-specific language use.

### 5. Conclusion and future work

The evaluation of the three thesaurus construction methods has clearly shown that they complement each other having different forces. Our work shows that a construction process should always be based on a domain study clarifying the
purpose and role of the thesaurus. The main concept categories will be identified by the domain study, and the collection process should be initiated by carrying out word association tests based on stimuli words covering the main concept categories. The stimuli words should preferably be selected from up-to-date documentation, strongly related to the topical scope of the thesaurus. Expert groups are recommended to assist the thesaurus compiler in selecting literary sources on which the subsequent evaluation of relevance and validity of the word associations should be based. One may say that the word associations act as "entrance" to the current corporate languages in play. The word association tests will not cover exhaustively the topics, and must be supplemented with conceptual information from literary scanning of internal and external sources.

All the methods evaluated require considerable human intellectual resources. The challenge is to combine the methods in such a way that the advantages of each method are exploited optimally. By joining the methods it is possible to construct a corporate thesaurus, which, at the same time, map the diversity of the corporate language use, integrate corporate knowledge and refer to relevant retrieval synonyms.

Another important issue to be addressed is thesaurus maintenance: how to ensure that the corporate thesaurus at all times exposes vocabulary developments. We believe that word association tests, due to their ability to elicit up-to-date, domain-oriented knowledge, performed on a regular basis will be the ideal tool in the future maintenance of the corporate thesaurus.

Future research will concentrate on developing the word association methodology and create word association tests in electronic form to be sent out by e-mail. The research will furthermore focus on investigating the possibilities of collecting conceptual information other than word associations.

References


