User Associations - Do They Have Any Value in a Corporate Thesaurus?

Marianne Lykke Nielsen
Royal School of Library & Information Science

Abstract:

In the paper demands to corporate thesauri are discussed. For more than thirty years the thesaurus has been a valuable tool in information retrieval (IR). Still the thesaurus is regarded as an important tool. However, the role of the thesaurus is changing. Future thesauri will be used more in retrieval than in indexing. Future corporate thesauri should be developed according to the specific needs of the work context. The word association methodology is a method which reveals the language use of the respondents. The method and some practical projects are described in order to evaluate the value of user associations in the construction of thesaurus. Some methodological problems are discussed regarding choice of respondents, choice of stimulus words, currency, and meaning of the test result.

Keywords:

Corporate thesauri ; word association methodology ; thesaurus construction ; methodología de associaciones ; construcción de thesauri ; thesauri

Introduction

The information landscape of most enterprises is multifarious. Information is retrieved from a wide range of internal as well as external information resources: document management systems, local databases, external resources, the Internet etc. The searching is carried out by librarians and end-users, each group characterised by a different search behaviour. The thesaurus has played an important role in the retrieval process for a long time, guiding the indexer as well as the searcher, and providing precision and security in the searching. In recent years the role of thesaurus in IR has been challenged by the possibilities of full-text retrieval, weighed search techniques, non-boolean query systems, ranking techniques, and relevance feedback. User studies also show that subject searching aids like thesauri are rarely used by the end-users which, in return, more frequently than formerly make their own searches. Today end-user searching is widespread in many enterprises. It characterises end-user searching that end-users do not fully take advantage of the different types of conceptual search facilities. Typically, they are not aware that there may exist ambiguities and variants in the understanding and naming of a concept, and they do not consider the problem when they access the IR system. Another constraint for the thesaurus is the fact that it is a costly tool. It requires time and high skill to use the thesaurus and especially to construct and maintain it. All these facts are important reasons to rethink the concept of thesaurus and reconsider the content, function and construction methods for the thesaurus. It is imperative to develop a tool, which corresponds to the need of the end-users because they constitute an important part of information retrieval.

This paper will focus on and explore the requirements which may be put on future corporate thesauri. The word association methodology will be analysed in order to evaluate whether this methodology may improve the usability of the thesaurus, especially in relation to end-user searching. The first section of the paper introduces briefly the corporate thesaurus. Section 3 provides an overview of the typical retrieval problems in IR in order to clarify the need of retrieval tools like the thesaurus and to outline the tasks in which the thesaurus may help the searcher. Section 4 presents the word association methodology and its role in the construction of thesauri. The following section describes two projects in which word associations have been used and tested as means to improve the users' interaction with the retrieval systems. Section 6 discusses some methodological problems which have to be considered, applying the word associations in the construction of a corporate thesaurus. The conclusion points to future research projects.

What is a Corporate Thesaurus?

The corporate thesaurus is a thesaurus which is developed according to the information landscape and conceptual needs of a particular company. Like other thesauri the function of the corporate thesaurus is to provide a vocabulary to facilitate information retrieval. A thesaurus is a tool which guides the indexer, selecting terms for indexing. During searches the thesaurus is a tool which may help the searcher to explore and perceive the different aspects of the topic, conceptualise the information need and locate appropriate access and search terms. Today the function of the thesaurus is changing, and future thesauri will probably be used more as retrieval tools than indexing tools. The

knowledge of the thesaurus may be used directly and consciously by the user himself, or may be applied automatically by the system [20].

Retrieval Problems in IR

Information retrieval is an iterative process consisting of six main tasks which are strongly connected and in practice are carried out interactively:

- perception of a work task or interest/problem situation
- analysing and conceptualising the information need
- locating and choosing the appropriate sources and access points
- searching
- evaluating the search result
- modifying the information need or the query according to experience, learning, and feedback

In corporations retrieval is carried out in a variety of different information systems. Some systems are query-based systems, others are not query-based. In both types of systems the retrieval is based on a match between terms specified by the users to represent the information need and terms appearing in the database to represent the information objects, and the outcome of the retrieval depends on the set of terms which is used to access the retrieval system. In non-query-based systems the retrieval may be based on other features, but it is often some words that provide the very first access to the retrieval system.

In a typical subject retrieval process the user is supposed to describe and find a way to something that he or she does not know, and it is a well-known problem for online searchers to recall from memory and find appropriate access and search words. As the nature of the need is often intuitive and only implicitly recognised at the initial stage of a search, the result is that users often approach the system with a query formulated out of the first words that come to mind. They use broader, general terms to describe their information need [2, 13], and as many databases try to index as specifically as possible, the searchers do not necessarily hit the words of the database. The process to perceive and conceptualise the information need is seldom straightforward. Kuhlthau [15] characterises the task of perception and conceptualisation as a exploratory stage and describes it as the most difficult stage of the retrieval process. Feelings of confusion, uncertainty and doubt are often dominating this stage. Thoughts centre on becoming oriented and sufficiently informed about the topics to form a focus or a personal point of view. At this stage an inability to express precisely what information is needed makes communication between the user and the system awkward. Iivonen [12] recommends to see the selection of access terms as a meeting place of different discourses. The selection process is not only a translation process, but a situation where the users should try to identify and understand the different ways of talking and thinking about a certain topic. The users must take into account the different ways in which the topic may be described in the vocabulary of the database. Ideally, the need should evolve and shift over time according to feedback and new knowledge of the subject.

User studies show that experienced users like librarians understand the need for variety in the vocabulary when they wish to do a thorough search. They use a variety of sources to find search alternatives [5-7,12]. End-users, however, use only one or few words for searching. They use a surprisingly great variety of words to refer to the same thing [8, 11]. In fact, it is impossible to predict what specific terms or phrasings they will use in formulating their requests. If the system does not provide a huge set of access points, end-users may have difficulties to get into the systems and find the information objects they desire [9]. Therefore, the amount and variety of access points for a concept are crucial. It should be possible for searchers to access the database using their own vocabularies and naming of concepts. Access points are important because people, in general, can recognise required information more easily than recall it.

Thus, the users may need two kinds of conceptual help: access points and conceptual knowledge to explore and understand the subject field. In figure 1 the different sub-tasks of the retrieval process are outlined in which the thesaurus should support the searcher. Looking at the figure, it seems clear that the well-known basic semantic relationships provide the searcher with valuable information. Referring to equivalent, broader, narrower and associatively related terms certainly may help the searcher to realise the history, the ambiguities, the variants, et cetera of the terms. The "good old" knowledge of a traditional thesaurus is still relevant, but the organisation and terminology should reflect the practices of the work or subject domain in which the thesaurus is going to be used. When the thesaurus is used for multi-database searching, the function as a "meeting place of discourses" is even more important. A varied presentation of different understandings and approaches to a concept is essential.

The Word Association Methodology

The word association test is a method which has been used by psychologists to reveal the intuitive, unconscious knowledge of an individual. It is an accepted method to examine the respondents' verbal memories, thought processes, emotional states and personalities. In its simplest form a series of disconnected words (stimulus words) are projected orally or in writing to the respondents who must respond with the first word which comes to mind (response words).

Sub-tasks related to the perception of the work task/problem situation
• interpret the aspects, attributes and approaches of the work task/problem situation
• delimit the work task/problem situation
• describe the work task/problem situation
Sub-tasks related to the conceptualisation of the information need
• place the information need in appropriate subject fields
• define aspects, attributes and approaches of the information need
• get information about the aspects, attributes, and approaches
Sub-tasks related to the location and choice of access points
• consider that ambiguities or variants may exist with the understanding and classification of a term
• take into account that there may exist a topical, social, and cultural history of the term
• consider that a concept can be referred to by a large set of variant forms: synonyms, spelling forms, popular and scientific forms, nicknames, names in different languages etc.
• consider proper names as search terms
• consider that a concept may be referred to by different controlled terms in different databases
• consider that a subject can be described on different levels of abstraction
• consider that a subject may be described from different aspects/facets
• consider that a subject may be described from different viewpoints
• consider the value of expanding a search to related subjects
• consider search criteria other than subject terms
• evaluate the value of search words
Sub-tasks related to the modification of the information need or query
• evaluate the search result
• evaluate the used search terms
• evaluate alternative search terms appearing in the retrieved documents
• consider other aspects, attributes and approaches to the topic/work task
• consider new search terms

Figure 1: Sub-tasks in information retrieval

The respondents may be allowed to associate freely (free association test) or responses may be limited to certain semantic categories, to particular synonyms, to terms within a certain context or to choose among alternatives (controlled association test). The responses may be manipulated by priming the respondents. This is done through verbal instruction and through the setting of the physical equipment of the simulated (work) context. Explicit information about the purpose of the test and context of the stimulus words as well as visual impressions communicated by the physical surroundings is expected to prime the respondents' mental models of the work domain and thus influence their associative responses. Priming is normally used in controlled tests.

Different methods exist to present the stimulus words. In a discrete test a stimulus word is presented once and the respondent must associate one response. In a continued test the same stimulus word is displayed a number of times and the respondents must continue to give new responses to the original stimulus word. In a continuous test the stimulus word is used as a starting point for a chain of responses. It is displayed only one time to the respondents. Comparisons of the possible effect on responses of the various test methods do not provide the consistent results needed to elucidate which methodology to use for specific purposes [22].

Within IR the association test is considered as a way to identify the associative meaning of a stimulus word - one can say that the response words create a cluster of meaning or associative representation of the stimulus word. The cluster of response words is regarded as an indication of the respondent's (un)conscious understanding of the relevance and relationship between the stimulus concept and the concepts revealed by the test. The associations expose the respondent's feeling of "what goes with what" [4]. It is the structure of our situational, working memory which is revealed by the method [14].

One of the main reasons for introducing the word association methodology in thesaurus construction is the need of a huge and varied access vocabulary. It is a difficult task to get hold of all the different expressions and names for a given concept. By the word association method it should be possible to capture the spontaneous, associative and variant vocabulary of the users, and the method may be an efficient way to collect a variety of the different names. Another important objective is to use the method to capture associatively related terms showing domain-specific aspects and approaches to

the stimulus words. The associative relationship is of great importance for the perception and conceptualisation of the information need, because it shows the context of the analysed concept and gives reference to a variety of aspects connected to the concept.

Within thesaurus construction it is a well-known problem to define and identify the associative relationship. The ISO standard for thesaurus construction recommends to found the associative relationships according to the frames of reference shared by the users of the system [1]. Research in recent years has also shown the importance of incorporating domain-specific semantic knowledge related to the specific work task situation or to a specific subject domain into the thesauri [3, 10, 23]. A number of differences exists in the character of information needs and the vocabulary to formulate the needs across intellectual domains, between formal and informal meanings and structures of concepts, and between individuals. A corporate thesaurus should reflect the different meanings and names relevant in the particular context of the corporation. If the content of the thesaurus is not relevant and meaningful for the user and the problem situation, the thesaurus will not be used [5-7]. The association test has proved to be a gateway to terms related to the stimulus word according to the frame of the respondents and their work context.

Empirical Findings

Until now only a few IR researchers have dealt with the word association method [18, 19, 21, 22]. The common denominator of the research projects has been the wish to capture and integrate the users' active mental models and understanding of the vocabulary in the information systems. The intention has been to make the IR systems more intuitively understandable and user friendly by integrating the individual, colloquial vocabulary of the users.

The empirical projects can be divided into four groups, characterised by the way the user associations are used. One approach is to **structure** a set of stimulus words according to their associative relatedness, identified by the overlap of common response words. Another approach considers the response words as related terms and uses the methodology to identify connotative, empirical derived **relations** to the stimulus words. A third application is to generate an associative **lead-in vocabulary** to the stimulus words. The last approach sees the associations as representations of the users' mental model of their domain and uses the methodology to capture the specific understanding and **use of language** in a certain work domain. Having the objective of this paper in mind, only the pro-jects, in which the word associations have been used to identify associative relations, entry terms and the language use of a specific user group will be described.

	Type of application of association tests			
	Structure	**Relations**	**Lead-in-vocabulary**	**Language use**
Textual information objects:				
Fiction [22]	X			
Food technology [19]		X	X	X
Marketing [18]		X	X	X
Computer Science [18]		X	X	X
Pictorial information objects:				
Fiction [22]		X		
News media [21]			X	

Figure 2: Empirical projects applying association tests

As it appears in figure 2 the methodology has been used within very different domains: fiction, food technology, computer science, marketing, and the news media as well as in relation to different kinds of information objects.

The main purpose of the project of Lykke and Skrubbeltrang [19] was to test whether the association test is suitable for identification of:

- the language use of a specific user group: their terminology (choice of words, form, use of abbreviations and compound terms etc.) and their way of relating concepts
- associatively related terms to the stimulus words
- synonyms and near-synonyms to selected concepts in order to generate a large system of lead-in-terms

The association test was carried out as a controlled, primed test in the research centre of an international food company. In spite of control and priming the respondents could easily make the associations. Afterwards the form of the response words was standardised linguistically, and a word database was created. This database was used to calculate the frequency and overlap between different types of words. The clusters of response words were compared with clusters of indexing terms. The

cluster of indexing terms was extracted by a ZOOM-like algorithm from the record fields of controlled terms, assigned uncontrolled terms, titles, and abstracts. The comparison was made to see whether the terminology of the authors and indexers were similar to that of the users/respondents. There was a surprisingly low degree of overlap; on average 31%. Within the key subjects of the company the overlap was better, on average 49%. Stimulus words belonging to actual research areas of the company also resulted in a better overlap compared to words belonging to research of an older date. The overlap between the test persons was also low, on average 22%. Even though the users belonged to the same work domain and more or less had the same professional backgrounds, the test revealed an extremely varied use of language. Because of the low overlap a thorough qualitative analysis of the response words was carried out by a group of subject experts. The analysis showed that the response words generally were of high relevance to the stimulus words and to the subject domain. The test revealed that the respondents' way of relating concepts depends on their individual work tasks and personal focus. It was also revealed that two of the respondents, recently coming from a similar, competing company, used their own specific vocabulary. Furthermore, the analysis showed that more specific stimulus word within the work domain provided a more well-defined set of highly related response words compared to more general stimulus word which provided a set of less interrelated words. The test also showed that compared to a controlled vocabulary within the subject field (Food Science and Technology Abstract) the method identified a greater amount of associatively related terms to the stimulus words, terms of a more specific level and terms from other hierarchies.

In another project at a business school the controlled test method was compared with the free association method to test the presumption that the controlled test method provides response words of stronger relevance to the stimulus words [18]. The stimulus words were selected from a local, controlled list of descriptors. The comparison showed that the controlled test revealed a higher degree of relevant response words than the free test method. The relevance of the response words were tested by a qualitative analysis which divided the response words into four groups:

- response words of strong relevance
- response words of weaker relevance
- response words of remote relevance
- response words of no relevance (noise).

Of the response words provided by the controlled method, 45% were of high relevance, whereas 24% of the response words provided by the free test method were of high relevance. Another purpose of the test was to see whether two distinct user groups made different associations to the same stimulus words. A group of marketing students were tested and compared to a group of students of computer technology. The comparison showed a different use of language according to word form as well as relationships. The response words associated by the computer students were of a more precise, specific level compared to the response words of the marketing students. Generally, the response words were also of higher relevance to the stimulus words. An explanation could be that the computer studies are more well-defined than the marketing studies. Another reason could be that the computer students had studied one year longer than the marketing students participating in the test.

Analysing the test results of the research projects, the conclusion is that the association test is a valuable method to get hold of the language use of a specific, limited user group working within the same context. The method elucidates the terminology of the users and identifies the users' intuitive, subjective way of relating terms. To sum up, the word association methodology is a useful method to catch:

- a varied and complex entry vocabulary
- a large set of variant forms for each concept
- a structure reflecting the work domain
- user-oriented, colloquial terms

Methodological Considerations

In order to develop a corporate thesaurus, it seems relevant to use a controlled test. Searchers working within a certain framework do not associate freely when searching information. They make their considerations and associations in relation to the work domain and frame of interest; the respondents are already primed - so to speak. To capture as many variants and relations as possible to a concept (the stimulus word) the continued test mode seems most suitable. The discrete test method will probably provide a smaller number of relations. When applying the continued modus compared to the continuous method it is more likely that the respondents keep on track and continue to give responses to the original stimulus word. Priming could be valuable to keep the respondents on track and to generate terms related to a particular context.

The outcome of the test depends to a great extent on the respondents' knowledge and focus and on the character of the stimulus words. The results confirm that the association test, not surprisingly, will provide a better result concerning relevance and overlap the more knowledge the respondents possess about the subject domain. The relevance of the stimulus words for the respondents will also affect the relevance of the response words. Both tests showed that stimulus words of high interest to the respondents provided response words of high relevance.

Regarding respondents a mix of people with a current and thorough knowledge about the subject domain and the work task situation and persons with less knowledge should be chosen. The subject experts will provide the most relevant related terms, but the set of entry terms will be more varied and colloquial, if persons with less knowledge also make associations. In order to avoid words of remote relevance and relation, one solution may be to ask the persons with less or cursorily knowledge, for instance the corporate librarians, only to identify possible synonyms and near-synonyms, another solution would be to analyse the test result according to the respondents' degree of subject knowledge.

The choice of stimulus words is the difficult part. Ideally, all thesaurus terms should be tested. The amount of thesaurus terms will determine, if this is possible. However, the fact is that terms of a specific level, for instance *direct marketing* compared to *marketing*, provide the best result. The test results also show that terms representing very important and actual topics have a positive effect on the test result. Therefore, a practice could be to test only specific terms, for instance terms from the lower levels of a hierarchy, and terms belonging to the core or new interests of the company. Another approach is to use Eleanor Rosch's concept of basic-level categories and test the terms, which represent the cognitive basic within the work domain. Rosch [16] has observed that categories are not merely organised in hierarchies from the most general to the most specific, but are also organised so that some terms form the basic level from which the generalisation proceeds upward and downward in the hierarchy [16]. We use these basic terms when we are perceiving and learning. They may serve well as access terms and as starting points for a conceptual model of a given topic. The identification of the basic level terms could cause problems, but they may be identified by subject experts from the same sources which are normally used in thesaurus construction: dictionaries, thesauri, textbooks, written user requests etc.

As to currency, there is always the risk that the relationships, identified by the users turn out to be strongly subjective as well as very related to the present situation. The identified relationships might only have a short and very specific relevance which may devaluate the value of the relationships. This depends on the scope of the thesaurus. In a corporate thesaurus tailored according to the needs of the company, relations representing the actual work situations ought to have a high status, especially in the search.

Another important problem is that the method is based on linguistic units out of context. The test does not directly reveal the meaning and the respondent's understanding of the stimulus words or the response words. The association test is based on individuals' intuitive and subjective associations, and the identified relationships are not well considered nor based on an explicit understanding of the subject domain. The results, so far, do not provide evidence of incorrect interpretation of the stimulus concepts, but incorrect interpretation may occur. A pilot test at the Royal School of Library and Information Science has shown that the problem may be solved by asking the respondents to write down their understanding of the stimulus words. By such a description it is possible to check the respondents' perception of the stimulus words. The control is time-consuming and will affect the time needed to carry out the test. In the pilot project the time consume increased considerably. However, a spin off of the control is the set of related words which may be extracted from the description.

Conclusion

The corporate thesaurus is a thesaurus, developed according to the problem situations and information needs of a particular corporation. The word association methodology has proved to be a valuable method to identify a set of terms related according to the mental model of the employees within the work domain. It seems reasonable to enrich the corporate thesaurus by user associations representing the vocabulary of the work domain. Until now the evaluation projects have mainly been concentrated on testing the characteristics of the word associations. The relevance and value of the associative relations have been evaluated by panels of subject experts. It has not been evaluated systematically, if the associative terms are means to improve the users' interactive search behaviour. In future research it should be tested if the associations increase the end-users' understanding and use of thesauri which was the overall objective of applying the method in thesaurus construction.

References

[1] Aitchison, J, Gilchrist, A & Bawden, D. Thesaurus construction and use: a practical manual. London : Aslib, 1997. 212 p.

[2] Bates, M J. Subject access in online catalogs: a design model. Journal of The American Society for Information Science, 37 (6), 1986, 357 - 376.

[3] Bates, M J. Indexing and access for digital libraries and the Internet: human, database, and domain factors. Journal of the American Society for Information Science, 49 (13), 1998, 1185 - 1205.

[4] Deese, J. Form class and determinants of association. Journal of verbal learning and verbal behavior, 2, 1962, 79 - 84.

[5] Fidel, R. Searchers' selection of search keys I. The selection routine. Journal of the American Society for Information Science, 42 (7), 1991, 490 - 500.

[6] Fidel, R. Searchers' selection of search keys II. Controlled vocabulary or free -text searching. Journal of the American Society for Information Science, 42 (7), 1991, 501 - 514.

[7] Fidel, R. Searchers' selection of search keys III. Searching styles. Journal of the American Society for Information Science, 42 (7), 1991, 515 - 527.

[8] Furnas, G W, Landauer, T K, Gomez, L M & Dumais, S T. The vocabulary problem in human-system communications. Communications of the ACM, 30 (11), 1987, 964 - 971.

[9] Gomez, L M, Lochbaum, C C & Landauer, T K. All the right words: finding what you want as a function of richness of indexing vocabulary. Journal of the American Society for Information Science, 41 (8), 1990, 547 - 559

[10] Hjørland, B & Albrechtsen, H. Toward a new horizon in information science: domain-analysis. Journal of te American Society for Information Science, 46, 1995, 400 - 425

[11] Iivonen, M. Consistency in the selection of search concepts and search terms. Information Processing & Management, 31 (2), 1995, 173 - 190.

[12] Iivonen, M. Selection of search term as a meeting place of different discourses. In: Green, R. ed. Proceedings of the Fourth International ISKO Conference, 15 - 18 July, 1996, 224 - 230.

[13] Ingwersen, P. Search procedures in the library analysed from the cognitive point of view. Journal of Documentation, 38, 1982, 165 - 191.

[14] Kiss, G R. An associative Thesaurus of English: structural analysis of a large relevance network. In: Kennedy, A. And Wilkes, A. (eds), Studies in long term memory (pp. 103- 121). London: Wiley, 1975.

[15] Kuhlthau, Carol C. A principle of uncertainty for information seeking. In: Journal of Documentation, 49 (4), 1997, 339 - 355.

[16] Lakoff, G. Women, fire, and dangerous things. What categories reveal about the mind. Chicago: The University of Chicago, 1987

[17] Lancaster et al. Identifying Barriers to Effective Subject Access in Library Catalogs. LRTS, 35 (4), 1991, 377 - 391.

[18] Lindholm Kjær, S, Møller, H & Sognstrup, H. Ordassociationstest i teori og praksis [Word association test in theory and practice]. Aalborg: Royal School of Librarianship, 1994.

[19] Lykke Nielsen, M. The word association test in the methodology of thesaurus construction. In: Efthimiadis, E. N. ed. Proceedings of the 8th ASIS SIG/CR Classification Research Workshop. Held at the 60th ASIS Annual Meeting, Washington, D.C. November 1-6, 1997, 43 - 58.

[20] Milstead, J L. Invisible thesauri: the year 2000. In: Online & cdrom, 19 (2), 1995, 93 - 94

[21] Ornager, S. Image retrieval: Theoretical analysis and empirical user studies on accessing information in images. In: Proceedings of the 60th ASIS Annual Meeting. Washington, DC, November 1-6, 1997, 202-214.

[22] Pejtersen, A Mark. Interfaces based on associative semantics for browsing in information retrieval. Risø Laboratory, Roskilde, Denmark, 1991.

[23] Siegfried, S, Bates, M J, Wilde, D N. A profile of end-user searching behavior by humanities scholars: the Getty online searching project report no. 2. Journal of the American Society for Information Science, 44 (5), 1993, 273 - 291.