The word association methodology - a gateway to work-task based retrieval

Marianne Lykke Nielsen

Department of Information Studies
Royal School and Library & Information Science
Aalborg, Denmark

Peter Ingwersen

Department of Information Studies
Royal School and Library & Information Science
Copenhagen, Denmark

Abstract

This paper presents the word association methodology as means to improve user interaction and access to information retrieval (IR) systems. The result of the three research projects are described in which the method has been used to catch the intuitive, colloquial vocabulary of end-users. The evaluation of the research projects have been concentrated on the users' perception and assessment of the associative semantic networks, and there exist no conclusive data about the value of user associations as means to improve the interactive search behaviour. A new framework for the evaluation and testing of associative relations is developed. The evaluation method is based on the concept of simulated work task situations.

1 Introduction

The word association test is a common method within psychology which has been used to reveal the private world of an individual. 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). These associations reveal the respondents' verbal memories, thought processes, emotional states, and personalities. Psychologists examine the nature and probabilities of the response words, and sometimes the amount of time it takes to respond.

Within information science only a few researchers have dealt with the word association tests. Kiss [23] used the method in order to develop an association thesaurus of the English language, a kind of Rogets Thesaurus. In 1987 the methodology was used to relate a set of 140 paintings of an art exhibition in Hamburger Kunsthalle. In a computer program the paintings were grouped in relation to common thematic indexing words and associations provided by different users of the system [3]. Pejtersen [29] used the word association method to structure descriptors from a controlled vocabulary. The descriptors were structured according to their degree of overlap of association response words. If the descriptors had many response words in common, they were related in the thesaurus. In another test by Pejtersen [29] user associations were used to identify the associative meaning of icons for a graphical user interface. Lykke and Skrubbeltrang [26, 27] used the test method having two purposes in mind. One aim was to use the method to generate entry words to a domain-specific in-house thesaurus; another purpose was to find related concepts to descriptors of the thesaurus according to the users' language use. Lindholm Kjær, Møller & Sognstrup [25] wanted to use the method to identify differences in the language use of two distinct user groups. Ornager [28] has used the methodology to find terms related to subject categories used to index digitised press photographs in order to improve the search facilities of an image database.

The common denominator of the different research projects is the wish to capture the users' active mental models and understanding of the work domain with the aim to integrate this knowledge into the systems' interfaces and searching tools. It is a well known problem for users of IR systems to enter the systems, capture the vocabulary of the systems, and find the right access and search words. The intention has been to make the IR systems more intuitively understandable and user-friendly by integrating the colloquial vocabulary of the users and by relating the terms according to the users' specific use of language. The objective is to adapt to the users' search behaviour [4].

The test results of the different research projects indicate that the association test method is a valuable method to identify the specific mental model and language use of a limited user group. However, the aims of the different research projects have been quite different, and the methodology has not been analysed and tested systematically. Thus, it is not possible to make conclusive recommendations regarding the value of the method. The purpose of this work is to analyse the potential of the method within IR and to develop an experimental design to test the hypotheses that thesauri, augmented by user associations, may improve the interactive user-system dialogue.

The second section of the paper presents briefly the word association methodology. Section 3 provides an overview of the typical retrieval problems in information retrieval in order to clarify the purpose of introducing user associations in IR. The following section describes three projects in which word associations have been used and tested as means to improve the users' interaction with the retrieval systems. Section 5 discusses the methodologies used to evaluate the value of word associations. The final section is a description of a proposed methodological framework developed to test the value of user associations as means to improve the quality of IR interaction.

2 The Word Association Method

The association test is a way to identify the associative meaning of a stimulus word or relationship between two stimulus words [12]. The response words create a cluster of associative representations of the stimulus word. High similarity of response words between two stimulus words indicate relationship between the analysed stimulus words. The clusters of response words can be considered 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 association test exposes the respondents' feeling of "what goes with what;" it is the structure of our situational, working memory which is revealed by the association test [23]. The hypothesis is that, given such relationships, they can be used to form a user-driven thesaurus structure.

Various methods need to be considered when planning a word association test. The respondents may be allowed to associate freely (free association test) or responses may be limited to semantic categories, to particular synonyms, to terms within a certain context, or to choose among alternatives (controlled association test). Priming is a way of manipulating the responses 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.

The presentation of stimulus words varies a lot, depending on the methodology. One distinguishes between several different test methods. 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. By using a continuous test method there is a risk that the respondents will associate their responses to the last given response word, which provides a chain or set of associations. Comparisons of the possible effect on responses of the various test methods do not provide the consistent results needed to elucidate what methodology to use for specific purposes [29]. In order to develop user-oriented conceptual search tools within a specific work domain, a controlled test seems preferable. 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 relations as possible to a concept, the continued test mode seems most suitable, since the discrete test method will provide a smaller number of relations. The continued mode seems preferable to the continuous method, if one wants 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. This may be carried out by means of the concept of simulated work tasks as proposed by Borlund and Ingwersen [7].

3 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 [20]:

- 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, the request in NL or the query according to experience, learning and feedback

Retrieval in query-based systems 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. The outcome of the retrieval depends on the set of terms which is provided to the retrieval system and the way in which they are applied to the database. In non query-based systems the inherent features of the objects, i.e., terms, descriptors, or image features, even if indirectly, act as query terms, and often it is some words that provide the very first access to the retrieval system.

It is a well-known problem for online searchers to recall from memory and find appropriate access and search words. 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. As the nature of the need is often intuitive and only implicitly recognised at the initial stage of a search, users often approach the system with a query formulated out of the first words that come to mind. The process of coming to know what one wants and to describe it is seldom straightforward. When humans are going to describe something, they often activate many more words than they need, words which occasionally pop into an utterance inappropriately. Labelling is not an easy task. The process of searching and selecting words is best described as an interactive activation where irrelevant words get suppressed, until finally one word wins out over the others activated [2]. Iivonen [19] recommends to see the selection of search 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 have to take into account the ways in which the topic is described and organised in the vocabulary of the database. Ideally, the need should evolve and shift over time according to feedback and new knowledge of the subject. Users often use broader, general terms to label their information need [4, 21], and as many databases try to index as specifically as possible, the searchers do not necessarily hit the words of the database.

User studies show that experienced users 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 [13-15, 19]. End-users, however, use only one or few words for searching, and they have often problems to enter the systems. On the other hand they use a surprisingly great variety of words to refer to the same thing [16, 18]. In fact, it is impossible to predict what specific terms or phrases they will use in formulating their requests. Therefore, the access points should be of great variety to facilitate the match between the user‘s terms and those of the database. It should be possible for searchers to access the database using their personal, colloquial words, often intuitively selected to describe the information need. Also because of the fact that people can recognise required information more easily than recall it. The average person can recall only a fraction of the range of terms that are used to represent a concept, but can take in all the variants in an instant, and make a quick decision about desired terms for a given search [5]. Traditional conceptual search aids, like classification schemes and thesauri, are not designed according to the needs of end-users [4]. Neither the content nor the layout seem to be useful for the tasks and problems related to information retrieval. The result is that end-users do not explore the subject searching aids like thesauri [9, 24, 30, 31]. They do not get sufficient support to go through the important interactive process of recognising and selecting appropriate words to describe the information need. These retrieval problems on the user side also prevail in non-boolean retrieval environments, for instance, when users attempt to accommodate to query modification features of interfaces [6].

The dominant reason for introducing the word association methodology in IR is the wish to take hold of the spontaneous and variant vocabulary of the users. The aim is to improve the users’ interaction with the IR system and their use of conceptual tools like thesauri by including their intuitive thinking in the IR systems. The presumption is that associations which represent the intuitive vocabulary of the users may function well as entry vocabulary and guide the user to the semantic network of the thesaurus. The idea is that the words, provided by the association test, respond to the words which pop up in the mind of users when they are labelling their information need. The associative terms may also help the users when they are perceiving and conceptualising the information need by linking words according to their colloquial and intuitive understanding. Research within language and communication show that logical relationships (hierarchical and synonym relationships) are only partly useful for the reasoning ability whereas co-ordinate and collocational relationships are important [2]. Co-ordinates are words which cluster together on the same level of detail, such as *salt* and *peper* and *butterfly* and *moth*. Two words are collocated when they are likely to be found together in connected speech, e.g., *colour* and *red*. Co-ordinates and collocational links are very common responses in word association tests [2, 23].

4 Empirical Findings

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, intuitively and empirically 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. Another possible application, not yet tested, is to apply the user associations to automatic and user-driven query expansion.

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

Figure 1: Empirical projects applying association tests

As it appears in Figure 1 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.

Within the fiction domain the methodology has been thoroughly investigated by Annelise Mark Pejtersen [29]. The free association method has been tested to explore whether the method can be applied to verify the underlying network for concepts and relations accessible to the users for retrieving words. The hypothesis of the work was that the network of word associations is comparable to semantic networks and thus could be usefully incorporated into IR systems to aid user-system dialogues. The network is expected to encompass the user's point of view. A free discrete continued association test methodology was used. Fifty respondents made associations to 223 stimulus words consisting of subject terms from a bibliographic database of novels. There were no attempt to prime the respondents. They were asked to record spontaneously the first word that came to mind when the stimulus word were displayed on the screen. The subsequent interviews after the test revealed that the respondents found the task easy despite the great number of stimulus words. The first response word usually came quickly, while difficulties with the third word sometimes occurred. Participants felt that their associations revealed something about their inner life, and they often found their own responses surprising. The following six stimulus words received the association 'May the first':

- Labour-movement
- Class distinction
- Workers
- Class struggle
- Resistance
- Upper class

The data collection of response words was checked in order to ensure the precision of the statistical analysis of the data. Linguistic deviations, for instance singular versus plural, adjective versus substantives, definite versus indefinite articles, ways of spelling etc. were eliminated before further analysis. Afterwards the relations between the subject terms used as stimulus words were determined from the extent of overlap of the associative responses to the terms. The measure of overlap is described in detail in [29]. The analysis of term relations revealed a frequent occurrence of conventional, hierarchical, generic, and part whole relationships in addition to non hierarchical relationships. Thus, the word association method captured conventional as well as more unexpected relations. However, the number of hierarchically related terms under each entry term in the thesaurus was very small compared to the number of pure associative.

The resulting thesaurus was tested using a variety of test methods. The evaluation was focused on the users' assessment of the associative semantic network. The testing was aimed to gain information about the test persons' perception of the new kind of relations and to explore the performance of the associative thesaurus when used as an aid in the users' identification and formulation of a need during a browsing search mode. The evaluation was performed as a set of laboratory experiments. First the test persons' selection of terms from a thesaurus, consisting of words derived from the association test and words from a conventional thesaurus, was analysed, qualitatively and statistically, to explore the subjects' preferences for association terms compared to conventional thesaurus terms. The test showed that the final need formulation contained 62% associatively derived terms and 32% conventional term on average. In 91% of the cases the percentage of associative terms was highest. Secondly, the value of degree of term relatedness between terms in the thesaurus was evaluated by the test persons. This experiment confirmed that terms derived from word associations are perceived to be meaningfully related.

In the third experiment the deeper understanding of how and when the thesaurus was helpful/not helpful was explored. The test persons were asked to give concurrent verbalisation protocols and think aloud about their information need and their motivation for choice or rejection of terms in order to contribute with data that might enlighten their reasons for choice. Terms evaluated most positively were terms related by the word association method. In the last experiment the degree of inspiration and usefulness of the associative thesaurus was compared to a traditional thesaurus after the formulation of a boolean search query and again after the retrieved books had been displayed and evaluated. This was done by analysing the formulated queries and by asking the test persons' opinion. The conclusion was that the associative thesaurus performs as well as the conventional thesaurus with respect to the output of a search.

The project of Lykke and Skrubbeltrang [26, 27] was done in the Research Centre of an international food company. The main purpose of the project 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 involving nine users of the information system. The stimulus words were presented to the respondents in written as well as in oral form, and the respondents had 1 ½ min to associate, preferable with three responses. The continuous test method was chosen, because the pilot project showed that the test persons found it easier and more comfortable to associate three relations at a time. They had no problem keeping on track and continuing to associate the original stimulus word. The respondents were instructed to associate in relation to their work domain and work tasks avoiding general, broad concepts within the domain, e.g., *milk* and *cheese*. Every respondent wrote down his personal responses. In the test the respondents made associations to 24 stimulus words, taken from the system's controlled vocabulary.

The test was successful; the respondents could easily make the associations. Before further analysis the form of the response words was standardised linguistically. Afterwards a word database was created. The stimulus words (descriptors from the system's controlled vocabulary) were recorded as the title of the record, the response words were registered in the subject field, and the respondents in the author field. This database was used in the statistical analysis, calculating the frequency and overlap between the different types of words. The clusters of response words were compared with clusters of indexing terms. The clusters of indexing terms were extracted by a ZOOM-like algorithm from the following record fields: controlled terms, assigned uncontrolled terms, titles, and abstracts. The overlap between association words and indexing terms was calculated as the proportion between the total amount of unique association words to a stimulus word and the amount of unique association words, already known to the system. The comparison was made to see whether the termino-logy 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 areas 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%. The overlap was calculated as the amount of unique associations which have been associated by more than one of the nine respondents. 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. The test did not reveal a different use of synonyms, but did find that there was a different understanding of relationships for the concepts used as stimulus words. The test also showed that two of the respondents, recently coming from a similar, competing company, used another vocabulary. Some concepts within the work domain have a sort of nickname, and the new employees did not use the local nickname. They used their "old" local vocabulary.

A qualitative analysis of the response words showed that the response words, generally, were strongly related to the stimulus words and to the subject domain. The qualitative analysis was carried out by two subject experts, both having a degree in Dairy Engineering. One of the experts was working within the work domain whereas the other expert came from a company within the industry of food production. The test revealed that, compared to a controlled vocabulary within the subject field (Food Science and Technology Thesaurus), the method identified a greater amount of associative terms to the stimulus words, terms of a more specific level and terms from other hierarchies. Moreover, the test revealed that the respondents' way of relating concepts also depends on their personal focus and work tasks. The qualitative analysis also showed that more specific stimulus word within the work domain provided a more well-defined set of highly related response words, whereas more general stimulus word provided a set of less interrelated words. Some relationships, identified by the association test, are shown in Figure 2.

Stimulus word:	***Brine***	***UHT***	***Whey***

Response word:	*cheese making*	*aseptic packaging*	*b-lactoglobulin*
	cheese salting	*bacterial spores*	*cheese*
	chilled water	*chocolate milk*	*desalination*
	feta	*cooked flavour*	*filtration*
	microfiltration	*cream*	*hyperfiltration*
	salt	*desserts*	*lactose*
	salt slurry	*heat treatment*	*permeate*
	taste	*holding time*	*reparation of protein*
		infusion	*skim whey*
		juice	*traditional cheese*
		milk products	*UF cheese*
		protein denaturation	*ultrafiltration*
		recombined products	*whey protein*
		sterile products	*WPC*
		sterilization	
		whole milk	

Figure 2: Examples of associations to three selected stimulus words [27]

In the third 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 [25]. In the free test 41 students made associations to 18 stimulus words; 48 students made associations to 18 stimulus words in the controlled, primed test. The tests were carried out as continuous tests; the respondents had 1 minute to give 2 associations. The stimulus words were selected from a local, controlled list of descriptors. The relevance of the response word was evaluated according to the degree of relationship between the stimulus word and the response word. The response words and their relationship to the stimulus words were evaluated according to encyclopedias, dictionaries, and textbooks, and divided into four groups:

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

The categorisation was checked by three subject experts (teachers of the Business School), working within the selected subject areas. The comparison showed that the controlled test revealed a higher degree of relevant response words than the free test method. 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 stronger 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. These results confirm the results from the project of Lykke and Skrubbeltrang 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 topicality and interest provided response words of strong relevance.

Word associations have also been used to identify the associative meaning of images. Pictures are more ambiguous and multidimensional in their connotations than words, and the hypothesis of the project has been that the associations that they may produce may be useful in supporting the users' recognition and formulation of information needs [29]. The association methodology was used in different ways. It is possible to distinguish between tests which use the picture as stimulus words in order to generate a word-based associative meaning of the picture, and tests where designers respond to stimulus words by drawing a pictorial associative response.

Pejtersen used the two different approaches in conjunction. First, three artists were asked to give pictorial associative responses to a set of stimulus words organised in topical clusters. Some stimulus words gave immediately pictorial associations, while other words did not give immediate associations. To confirm the validity of the associative responses between pictures and words a group of test persons were asked to choose among alternatives of words responses. The pictures derived from the first pictorial test were used as stimulus signs. The terms chosen to relate to the pictures were later used to generate a picture association thesaurus. In the search process the keywords related to a picture are searched by default simultaneously, when the picture is selected by the user as ‘search entry’. The usefulness of the pictorial thesaurus as a searching aid was evaluated in a real-life work domain context in a public library. The evaluation consisted of an online questionnaire, which appeared after the end of each search. The questionnaire contained questions about the understanding of the pictures and the usefulness of the thesaurus in need recognition and formulation. The testing showed that the users liked the thesaurus because it was useful for getting good ideas, and books they retrieved by selection of a picture were considered relevant to these ideas [29].

In Figure 3 the main results of the projects are summed up.

Results concerning the use of language	**Results concerning the methodology**
• the test reveals the terminology of the respondents, e.g., the specificity and form (the use of abbreviations, spelling, plural/singular and compound terms) • the test reveals useful, related terms • the test reveals mostly associative relationships, indicating the specific focus of the respondents • the test reveals some synonyms; the more respondents, the more distinct synonyms • the test reveals some hierarchical relationships • the test reveals sub-languages within an organisation • the response words have a low overlap with words extracted from the titles and abstracts of documents within the work domain	• controlled, primed test methods give more relevant, specific responses than a free test • respondents, working in a specific limited work domain, associate easily with three responses • the association test provide situational user- and domain-oriented relations • broad stimulus words result in a lower overlap than more specific stimulus words. The set of response words is not so well-defined as for more specific stimulus words • stimulus words, concerning the main subject of the domain, result in a higher overlap and stronger relevance of response words • stimulus words in relation to new research result in higher overlap and stronger relevance of response words • stimulus words from more peripheral research result in lower overlap

Figure 3: Test results

Analysing the test results of the research projects, the conclusion is that the association test is a valuable method to identify the use of language of a specific, limited user group. The method elucidates the terminology of the users regarding word form, use of abbreviations and compound terms, and identifies the users’ intuitive, subjective way of relating terms. The relationships, identified by the test, are generally of strong relevance. Compared to indexer thesauri the relations are more specific, reach across hierarchies, and express the special focus of the work domain. 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

The outcome of the test depends to a great extent on the respondents: their knowledge and focus, capability to associate, and their representation. The described projects have been limited to small groups of users and to well-defined subject domains. To use the association test in a wider context of a less well-defined user group and subject domain, it might be difficult to define and find a suitable group of respondents being representative of the future user group with regard to knowledge and focus. Another problem is connected to the amount of resources needed to carry out the test. In the described projects only a limited number of stimulus words have been used. Even though the test methodology may be considered as a semi-automatic and economic method, it may cause problems when the amount of stimulus words increases as regards the time needed for making the associations and in relation to the resources needed for the qualitative analysis.

Regarding currency, there is always a 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 devalue the value of the relationships. This depends on the scope of the thesaurus. Relations representing the actual work situation ought to have a high value in an in-house search thesaurus. Another important problem is that the method is based on linguistic units out of context. The test does neither 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. The identified relationships are not well considered (spur of the moment) nor based on an explicit understanding of the subject domain. This fact both solidifies the method and devalues it. It solidifies the method, because it makes it possible to catch and integrate the more spontaneous, colloquial and subjective terms and relations in the thesaurus structure. It devalues the method because of the possible diversity of understanding that causes unreliability in the test result. 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 checked by asking the respondents to write down their understanding of the stimulus words. Another solution, proposed by Aitchison [2], is to present the stimulus in context.

5 Discussion of the Methodological Framework of the Evaluations

The evaluation of the research projects have mainly been concentrated on testing the characteristics of the response words and relationships. The relevance and value of the associative relations have been evaluated by panels of subject experts related to the work domain and compared to clusters of indexing terms and to records from conventional thesauri within the subject field [25-27]. The users' assessment and perception of the new relations have been tested by Pejtersen [29] in a laboratory setting. However, it has not been evaluated systematically, if the associative terms and relations are means to improve the users' interactive search behaviour. It has not been evaluated

- how the associations affect the search behaviour and interaction
- if the associations increase the end-users' understanding and use of conceptual searching tools

The objective of using the word association methodology in the construction of conceptual searching tools is twofold. One important purpose is to test whether the associations improve the finding of conceptual starting points; another aim is to see whether the inclusion of a spontaneous, user-oriented, and domain-specific vocabulary increases the use and understanding of the conceptual help from IR systems. In order to make some conclusive recommendations regarding the value of the word associations and their influence on the quality of interaction the effect of this new kind of relations on the search process and on user behaviour should be examined. The effectiveness of the associations should be evaluated by assessing the provision of assistance in the users' performance of work context-situated problems. Such an evaluation may only be meaningful, if tested in relation to the specific work domain: to real tasks and to real users interacting with real-world systems. It should be observed how the associations affect the searchers perception, choices, and transactions. Qualitative as well as quantitative measures may be used. The traditional performance measures, recall and precision, are quantitative performance measures and may not be satisfactorily indicative of efficiency, usability and acceptability of the word associations. Performance should also be tested in relation to behavioural aspects and more qualitative dimensions of the search process.

Recent experiments indicate that it has long been recognised that evaluation of IR systems may be carried out as close as possible to real-life [7, 8, 17, 22]. It is recommended to develop an experimental setting which is characterised by the following features:

- anchored in work task based situations
- use of real-life information needs
- observation of real users' search behaviour and interaction
- performance measures related to tasks and sub-tasks of the search process
- performance measured by dynamic situational relevance assessments

Borlund & Ingwersen [7] have introduced and tested the concept of simulated work task situations and simulated information needs. A simulated work task situation describes the information need scenario by providing information as to: (1) the source of the need; (2) the environment of the situation; (3) the problem which has to be solved, and (4) make the test person understanding the objective of the search. This test methodology provides a context which ensures a degree of freedom for each individual to react and respond in relation to his or her interpretation of the given indicative request for that particular situation. At the same time the evaluation method is controllable and provides data concerning performance and search behaviour which may be compared.

6 Experimental Setting

With the aim to test the affect of word associations on the quality of interaction an experimental methodology has been developed. The evaluation framework involves real users from a specific work domain as test persons and applies the concept of simulated work task situations and simulated information needs to ensure realism and to enable comparison of the test result. The concept of simulated information needs makes it possible to test distinct types of information needs of the work domain. It may be fruitful to compare information needs of different levels of conceptual complexity.

The testing will be carried out within the framework of a Danish company working within the pharmaceutical industry. The test setting will be the local document management system, registering reports, guidelines, articles, test results, and confidential documents. The subject of the documents is indexed with the use of a corporate thesaurus. Twenty researchers of the company will be involved as test persons. The test persons will have a varied knowledge of IR and the subject area. None of the test persons will have any experience with neither the document management system nor the thesaurus at the time of the test. The system as well as the thesaurus are under construction. Each test person will carry out 4 search tasks; two search tasks for each type of thesaurus. The simulated information needs will be based on real task situations and real needs. The search problems will be relevant for all of the test persons. The simulated information need will be given in permuted order in order to avoid any effect that the order might have on the result, because the test persons gain new IR and subject knowledge during the search.

The testing will be carried out as a comparison of two different thesauri:

- one augmented by entry words and relations, identified by an association test
- another augmented by entry words and relations, identified by an automatic NL clustering algorithm

The design and functionality will be identical for the two test thesauri to ensure that it is the content of the thesaurus which is the main factor to be evaluated. Automatic clustering of NL terms is a well-known approach to develop search thesaurus [1, 11, 20]. It will be interesting to compare the performance of these two types of thesaurus, especially because the comparison of words associations and indexing terms, made by Lykke and Skrubbeltrang, showed a low overlap [26, 27].

The experimental setting will consist of the following basic components:

1) An introductory orientation session
   - a brief introduction to the experiment
   - demo and training of the IR system and the thesauri

The test persons will not get to know that they are using two different thesaurus. They will be instructed that their task is to find as many relevant documents that they feel they need to satisfy the information need.

2) A pre-search investigation
   - a brief questionnaire to identify searcher skills and experiences

   a pre-search interview to describe the searcher's perception of the work task situation and his plans for handling the situation.

The test persons will be asked to explain how they interpret the problem situation and the information need. They must also explain their choice of access terms. The interviews will be taped.

3) A search sessions in which the searcher interacts with the system and the thesauri
   - observation (via audio-video recording) and automatic logging of the search process with the aim to get information about user behaviour

Search behaviour will be gathered through audio-video recording and automatic logging. The search behaviour will be analysed according to the search tasks: a) analysis and conceptualisation of information need; b) location and choice of access points; c) search strategy; d) modification of search terms and search strategy. The search tasks will be divided into sub-tasks. The task of choosing access points, e.g., may be divided into considerations regarding: ambiguities with the understanding of a term; variant forms; levels of abstraction; different viewpoints; non topical search criteria; etc. The division into sub-tasks will be based on general IR models, modified according to the results of a user study identifying the search behaviour of the specific user group.

4) A post search investigation
   - a post-search questionnaire to reveal searcher's perception and satisfaction with the interaction and the search result

User satisfaction will be measured through a questionnaire in which the user evaluates the effectiveness and usability of the thesauri: the access points; the content of the thesaurus; the organisation of the content; the display of the content; the received support; the influence of the thesaurus. The satisfaction will be measured by the use of semantic differentials and Likert scales. These measurements have proven to be effective and accurate methods for acquiring, validating and analysing subjective data [8, 22]. In order to obtain objective, non-emotional information and avoid the "interview effect", it will not be possible to ask the test persons directly for their opinion of the two different kinds of test thesauri. The test design will be placebo-like and carried out as a blind test. The performance of the system will be measured by precision. The calculation of precision will be based on the individual user's interpretation of the simulated information need and based on situational relevance criteria, but the judgement will be controlled according to the simulated, but known work task situation.

7 Conclusion

Critical and important problems of end-user searching are connected to access and to the finding and choice of appropriate starting points for the search. In order to solve these typical retrieval problems it has been suggested to adapt the vocabulary of the IR systems to the mental models of the users and to the dominant work tasks of the work domain. In several research projects the word association methodology has been used to capture the intuitive, colloquial vocabulary of the end-users. The methodology has proved to be a valuable method to identify the use of language of a specific, limited user group. Furthermore, experiments show that word associations may also be useful in relation to pictorial information objects. A picture association thesaurus within the domain of fiction, consisting of pictures and their associated words, has proved to be a valuable tool for obtaining good ideas and retrieving relevant novels.

Until now, however, the evaluation projects have not focused on the effect of word associations on user behaviour and user interaction. In order to evaluate the value of associative relations it is suggested to study the quality and enhancement of user interaction. User associations are strongly connected to the specific work domain - to the users and their work tasks - from which they originate. It is proposed that such a user and domain-dependent system may only be meaningfully evaluated in real-world settings, where the researcher may observe the behaviour of real users, undertaking real tasks. A methodological framework based on the concept of simulated work task situations is developed, and it is hoped that future work will make it possible to present some conclusive recommendations regarding the value and methodology of word associations.

References

1. Aitchison, J, Gilchrist, A & Bawden, D. Thesaurus construction and use: a practical manual. London : Aslib, 1997
2. Aitchison, J. Words in the mind: an introduction to the mental lexicon. Oxford : Blackwell, 1994
3. Art in network. Hamburg Kunsthalle. Exhibition catalogue.
4. Bates, M J. Subject access in online catalogs: a design model. Journal of The American Society for Information Science, 37 (6), 1986, 357 - 376.
5. 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.
6. Beaulieu, M. Experiments on interfaces to support query expansion. Journal of Documentation, 53 (1), 1997, 8 - 19
7. Borlund, P and Ingwersen, P. The development of method for the evaluation of interactive information retrieval systems. Journal of Documentation, (53) 3, 1997, 225 - 250.
8. Brajnik, G, Mizzaro, S & Tasso, C. Evaluating user interfaces to information retrieval systems: a case study on user support. In: Frei, H.P., Harman, D., Schäuble, P. Wilkinson, R. eds. Proceedings of the 19th ACM Sigir Conference on Research and Development of Information Retrieval. Zurich, 1996, Konstanz: Hartung-Gorre, 1996, 128 - 136.
9. Efthimiadis, E N. A study of end-user behavior in searching cd-rom bibliographic databases. In: Proceedings of the 15th National Online Meeting, New York, May 10 - 12, 1994, 113 - 120.
10. Campbell, I and van Rijsbergen, K. The Ostensive Model of developing information needs. In: Ingwersen, P and Pors, N O, eds. Proceedings CoLIS 2 Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, October 13 - 16, 1996, 251 - 268.
11. Chen, H, Yim, T, Fye, D & Schatz, B. Automatic thesaurus generation for an electronic community system. Journal of the American Society for Information Science, 46 (3), 1995, 175 - 193
12. Deese, J. Form class and determinants of association. Journal of verbal learning and verbal behavior, 2, 1962, 79 - 84.
13. Fidel, R. Searchers' selection of search keys I. The selection routine. Journal of the American Society for Information Science, 42 (7), 1991, 490 - 500.
14. 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.
15. Fidel, R. Searchers' selection of search keys III. Searching styles. Journal of the American Society for Information Science, 42 (7), 1991, 515 - 527.

16. 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.

17. Hersh, W, Pentecost, J & Hickam, D. A task-oriented approach to information retrieval evaluation. Journal of the American Society for Information Science, 47 (1), 1996, 50 - 56

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

19. 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.

20. Ingwersen, P. Information retrieval interaction. London : Taylor Graham, 1992

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

22. Jose, J M, Furner, J & Harper, D J. Spatial querying for image retrieval: a user-oriented evaluation. In: Wilkinson, R., Croft, B. & van Rijsbergen, K. ed. Proceedings of the 21th ACM Sigir Conference on Research and Development of Information Retrieval. Melbourne, 1998. Australia: York Press, 1998, 232 - 240.

23. 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.

24. Lancaster et al. Identifying Barriers to Effective Subject Access in Library Catalogs. Library Resources and Technical Services, 35 (4), 1991, 377 - 391.

25. 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.

26. Lykke, M and Skrubbeltrang, C. Indekseringsstrategi og søgethesaurus for UMI-informationssystemet (Development of an indexing strategy and a searching thesaurus for the UMI systems). Aalborg: Royal School of Librarianship, 1992.

27. 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.

28. 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.

29. Pejtersen, A Mark. Interfaces based on associative semantics for browsing in information retrieval. Roskilde : Risø Laboratory, 1991.

30. 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.

31. Wallace, P M. How do patrons search the online catalog when no one's looking? Transaction log analysis and implications for bibliographic instruction and system design. Reference Quarterly, 33 (2), 1993, 239 - 252