Jaana Kristensen
Kalervo Järvelin
University of Tampere, Finland

The Effectiveness of a Searching Thesaurus in Free-Text Searching in a Full-Text Database

Kristensen, J., Järvelin, K.: **The effectiveness of a searching thesaurus in free-text searching in a full-text database.** Int.Classif.17(1990)No.2, p.77-84, 30 refs.

Synonyms, antonyms, quasi-synonyms and homonyms are typical features of natural language which cause difficulties in free-text searching in full-text databases. Searching thesauri which supply synonyms, antonyms and quasi-synonyms for given natural language terms have been suggested as a search aid to overcome these problems. To test these suggestions, a small searching thesaurus was constructed for Finnish newspaper articles on economic issues and a set of test searches were conducted and analyzed in a large Finnish newspaper article database. The original queries were elicited from journalists who also evaluated the relevance of the search results. Each query was searched in three distinct modes: basic search, synonym search and related term search. The basic searches contained only terms included in the journalists' original query statement. The logic of the basic search was developed by the researchers. In the synonym searches, the terms of the basic search were extended by disjunctions of the synonyms given by the searching thesaurus without modifying the overall logic of the basic search. In the related term searches, the disjunctions in the synonym searches were further extended by disjunctions of related terms (quasi synonyms) given by the searching thesaurus again without modifying the overall logic of the synonym search. The searches in each mode were analyzed w.r.t. relative recall and precision by setting the relative recall of the third mode searches to 100 %. The average relative recall in the first mode searches was ca. 45 % and in the second ca. 82 %. The average precision values were for the modes one, two and three were 51 %, 41 % and 33 %, respectively. The substitution of synonyms shows a considerable increase in recall with a negligible decline in precision. (Author)

1. Introduction

The evolution of information technology has made the electronic storage of very large text collections possible. Such collections, text databases (or full-text databases), often contain in a computer readable form the whole text, or at least large excerpts, of documents collected for some purpose. Text databases (briefly TDBs) may serve the daily routines of an organization, distribution of business information , or the communication of scientific information. TDBs are more relevant to the end-users than the conventional bibliographic databases because the former contain source data rather than surrogates. [4] [12] [13] [29]

Information retrieval in publicly available TDBs is based to a great degree on the same methods as the retrieval from bibliographic databases. Differences emerge in the importance of some operators (e.g. the proximity operator) and the applicability of free text searching. We shall use the term controlled vocabulary searching when the search terms are picked from a controlled vocabulary. The term free text searching is used when the search terms are picked from any available source (e.g. the user's mind). The term bibliographic retrieval denotes retrieval based on the bibliographic data fields; either controlled vocabulary searching or free text searching, or both may be used. (Obviously, in most of the cases, controlled vocabulary searching is applied on the index term field.) The term full-text retrieval denotes retrieval based on the document texts; in principle, either controlled vocabulary searching or free text searching, or both may be used (although the former is not quite relevant). Both bibliographic retrieval and full-text retrieval are possible in TDBs provided that the necessary data fields are available.

The relative effectiveness and associated potentials and problems (e.g. indexing costs) of controlled vocabulary searching and free text searching have been discussed in the literature (e.g. [1] [2] [5] [6] [7] [8] [18] [19] [26] [27] [30]). There does not seem to be one single best method -- controlled vocabulary searching and free text searching rather augment each other [6] [30]. The strengths and weaknesses of each method are well-known (e.g. [25]). Relatively recent empirical investigations on the performance of full-text retrieval in large operational TDBs are the evaluations by TENOPIR [30] and BLAIR and MARON [1] [2]. The general argument by Tenopir has been that full-text retrieval (through free text searching) improves recall at the cost of deteriorating precision. Blair and Maron argue that even the recall tends to be low and that automatic indexing methods are not very promising. The differences in the findings can be explained at least in part by differing approaches, different types of texts and evaluation criteria [21] [28]. SALTON challenges the results against automatic indexing [22]. Tools aiding in the identification of relevant search terms for free text searching have not been evaluated in these studies.

Free text searching requires that sufficient and appropriate synonyms, quasi-synonyms and antonyms are identified for the concepts describing the search topic. LANCASTER pointed out that any means, incl. existing controlled vocabularies, can be used as aids for term identification [15]. Later the use of searching vocabularies or searching thesauri have been suggested for this purpose [16] [20]. A searching thesaurus is not used in the indexing of the database contents. It is used only in the retrieval phase. In contrast to conventional thesauri, it does not seek to standardize the term choices by the searcher but rather provides alternatives (synonyms, quasi-synonyms, antonyms and related terms) to the terms the searcher has in his/her mind. Thus a searching thesaurus aims at providing some of the benefits of controlled vocabularies without the cost of indexing.

In this study, we evaluate the effectiveness of a searching thesaurus in terms of its effects on therecall and precision of search results. In particular, we want to find out its usability as an end-user tool in information retrieval. We do not know of earlier empirical, statistical tests on the effectiveness of searching thesauri. Our approach can be characterized in terms of sources of the test environment, the test thesaurus, the test searches and search result evaluation:

(a) The test environment is a TDB containing some 34,000 Finnish newspaper articles operated under the BASIS information retrieval and data management system.

(b) A small test thesaurus was constructed for the study on the basis of terms occurring in newspaper articles on economic issues. Thus the test thesaurus is literary warrant.

(c) Original queries (on economic issues) were elicited from journalists. Each query was searched in three distinct modes: basic search, synonym search and related term search. The basic searches contained only terms included in the journalists' original query statement. The logic of the basic search was developed by the researcher. In the synonym (and related term) searches, the terms of the basic search were extended by (inclusive) disjunctions of the synonyms (and the related terms) given by the searching thesaurus without modifying the overall logic of the basic search.

(d) Each journalist also evaluated the relevance of each article retrieved by the searches for his/her query statement in three classes (non-relevant, maybe relevant, relevant).

The chapter organization of this paper is as follows. Materials and methods are described in detail in Chapter 2; the findings are reported in Chapter 3; a discussion is given in Chapter 4; and the conclusions follow in Chapter 5.

2. Methods and Materials

2.1. The test environment

The test environment is a TDB containing Finnish newspaper articles operated under the BASIS [17] information retrieval and data management system. The database contains the articles published in the newspaper Satakunnan Kansa (published in Pori, Finland) from May 31, 1987 to September 7, 1987, altogether some 34,000 articles. The database is large enough for providing realistic findings (cf. [1]). In the basic index of the database, the terms were given in their basic form only irrespective of their inflected occurrences in the text. This was achieved by preprocessing the basic index entries by the morphological analysis program MORFO for the Finnish language [11] [14].

Although Finnish is a highly inflectional language, our study is not language bound, in principle. Synonymy, quasi-synonymy and other alternative forms of expression are features of all natural languages, not just Finnish. Only if it can be shown, that there are significant differences in the number of synonyms etc. and in the frequency of their use between different languages, our results are language bound to some minor degree. If so, our results are still indicative on the effects and value of searching thesauri.

2.2. The test searching thesaurus

A newspaper text is different from scientific and technical text both in vocabulary and in structure. There was no suitable documentation language available to serve as the basis for a searching thesaurus. The vocabulary and structure of Finnish subject heading lists and general classifications (e.g. the UDC) were unsuitable. Therefore a small test thesaurus was constructed for the study on the basis of terms occurring in newspaper articles on economic issues. Thus the test thesaurus has a literary warrant. As a further advantage, it has also a user warrant because journalists were the sole sources of both the articles in the database and the test searches.

The test searching thesaurus was organized by following the guidelines for thesaurus construction [10]. The structure of the thesaurus is quite conventional. However, there is no distinction between descriptors and entry terms: all terms belonging to the same equivalence class are treated equally because they were all equally possible search terms. The relationships between the terms were the usual ones: broader term, narrower term, synonym, related term. The thesaurus contained 180 different concepts represented by altogether 328 different terms. Some test thesaurus excerpts are given in the Appendix.

2.3. The test searches

The original queries (on economic issues) were elicited from journalists working on economic issues. The research setting was explained to them and they were asked to state queries they would execute in a text retrieval environment. Five journalists participated in the test and altogether 30 realistic queries were collected from them. The journalists' own words were the most compact expressions of their needs in free-text terms. Each separate concept a journalist had in his (her) mind was expressed by the terms the journalist used. This was used as the basis for the basic mode searches which contained only terms included in the journalists' original query statement.

The logic of the basic search was developed by the researcher. The terms representing different concepts were connected by the conjunction operator ('and'). When a concept was represented by a phrase, the words in the phrase were connected by a proximity operator restricting the word occurrences to a single paragraph. For example, if a journalist asks for articles on plans for nuclear energy in the paper industry, the concepts were nuclear energy and paper industry, and the basic search could be *nuclear energy* AND *paper industry*. (NB. the proximity operator for the phrase 'nuclear energy' is not given explicitly). The set of articles retrieved by this formulation is denoted below by B.

The problems due to synomyms, quasi-synonyms and other flexibility of natural language in full text retrieval are well known. They tend to appear also in indexing experiments. For example, IIVONEN has investigated inter-indexer consistency in free-text and controlled vocabulary indexing. She found out that different indexers, while being inconsistent in their term choices, tend to choose terms related by equivalence or associative relations [9]. The same is probably true of journalists writing newspaper articles: while writing on the same topic, different journalists use different terms which are likely to be related by equivalence or associative relations. To improve recall in full text searching, alternative forms of expression for each concept must be accounted for. We studied two extension types of the basic searches: synonym extensions and related term extensions. This was the rationale for the two additional search modes for each query: the synonym search and the related term search, which were also developed by the researcher following the guidance by the searching thesaurus.

In the synonym searches, the terms of the basic search were extended by (inclusive) disjunctions of the synonyms given by the searching thesaurus without modifying the overall logic of the basic search. Continuing the example above, the synonym search could be

(nuclear energy OR nuclear power OR atomic energy OR atomic power) AND (paper industry OR papermills)

assuming that the synonyms were provided by the searching thesaurus. This search is still an intersection of two concepts, i.e. the overall logic remains the same as in the basic search, but now each concept is represented by an inclusive disjunction for terms. The set of articles retrieved by this formulation is denoted below by S. Obviously, the result of a synonym search contains the result of the corresponding basic search, i.e. $S \supseteq B$.

In the related term searches, the disjunctions in the synonym searches were further extended by (inclusive) disjunctions of related terms (quasi synonyms) given by the searching thesaurus again without modifying the overall logic of the synonym search. Continuing the example above, the related term search could be

(nuclear energy OR nuclear power OR atomic energy OR atomic power OR nuclear power plants OR nuclear power stations OR nuclear energy production OR atomic energy production) AND (paper industry OR paper mills OR paper OR cellulose OR carton OR cardboard)

again assuming that the related terms were were provided by the searching thesaurus. The set of articles retrieved by this formulation is denoted below by R. Obviously, the result of a related term search contains the result of the corresponding synonym search, i.e. $R \supseteq S$.

Thus the results of the synonym searches contained the results of the basic searches; the results of the related term searches contained the results of the synonym searches, i.e. $R \supseteq S \supseteq B$. The full results of the related term searches were printed. The access numbers of the articles in the results of the basic and synomyn searches were recorded. For some queries there were either no synonyms or no related terms provided by the test searching thesaurus. The corresponding search mode was skipped in such cases. The synonym extension was possible for 23 original queries and the related term extension for 21 queries. There were 18 queries for which it was possible to form searches in all three search modes.

In the data analysis, also the net effect of query extension types recognized. The net effect of a synonym search is the set of articles retrieved by the synonym search and not retrieved by the corresponding basic search. In the above example, the net effect of the synonym search is $S - B$. The net effect of a related term search is the set of articles retrieved by the related term search and not retrieved by the corresponding synonym search. In the above example, the net effect of the related term search is $R - S$. By differentiating the net results, it was possible to analyze the quality of the search extensions in detail.

2.4. Search result evaluation

The researcher removed obviously irrelevent articles from the results of the related term searches. Thereafter each journalist also evaluated the relevance of the remaining articles retrieved by the related term search for his/her query statement. (NB: The articles removed by the researcher are counted as irrelevant in the statistics). The relevance evaluation consisted of three classes: irrelevant, maybe relevant, relevant. Only articles in the last class were counted as relevant in the recall and precision calculations. One or more relevant articles were retrieved for 26 different original queries, i.e. there were four queries for which none of the search modes retrieved any relevant articles. These four original queries were excluded from the analysis. Altogether 70 searches were conducted for the 26 original queries. The 70 searches yielded altogether 248 relevant and 493 irrelevant articles.

The search results were evaluated for recall and precision. The difficulty of estimating the recall of searches in operational databases of realistic size is well known. In our setting it was not necessary to estimate the absolute recall of the searches. Because we were interested in the comparison of the different search modes rather than their absolute performance and because the results from each related term search contained the results from the other search modes for the same original query, it was possible to compute relative recall figures for these searches. The relative recall of the related term search was set at 100 % for each original query. When the related term extension was not possible, the relative recall of the synonym search was set at 100 %.

The recall figures of the other search modes were then counted on the basis of their result sizes. For example, continuing the sample query on nuclear energy and paper industry, recall the search result sets **B**, **S**, and **R**. Let $\mathbf{R}^+$ be the set of relevant articles in **R**. Then the sets of relevant articles in the basic and synonym searches are $\mathbf{B} \cap \mathbf{R}^+$ and $\mathbf{S} \cap \mathbf{R}^+$, respectively and the number of relevant articles in each search $|\mathbf{B} \cap \mathbf{R}^+|$, $|\mathbf{S} \cap \mathbf{R}^+|$, and $|\mathbf{R}^+|$. Thus the relative recall of the basic search is $(|\mathbf{B} \cap \mathbf{R}^+| / |\mathbf{R}^+|) * 100$ % and the relative recall of the synonym search is $(|\mathbf{S} \cap \mathbf{R}^+| / |\mathbf{R}^+|) * 100$ %. We obtained therefore no data on how many articles were missed by all the searches but we got data on the relative performance of the search modes. In other words, we are able to compare the effectiveness of the test searching thesaurus as the source of suitable synonyms and related terms.

The false drops were also analyzed in order to find out their types and causes. We classified the false drops into those caused by a) article text, b) search strategy, and c) the test thesaurus. This classification is conventional (e.g. [15] [16]). The class for 'indexing errors' is non-relevant in our setting.

False drops caused by article text may be due to:
a1) search terms not being related in the text,
a2) incorrect relationship between the search terms,
a3) homonyms,
a4) the text dealing with the search topic only marginally or in a wrong way.

False drops caused by search strategy may be due to :
b1) failing to account for some concept in the original query,
b2) representing some concept by a too broad term,
b3) representing some concept by an incorrect term.

Conceptual correctness of the test thesaurus affects the number of false drops. Incorrect or spurious term relationships in the thesaurus cause false drops.

3. Findings

3.1. Overall recall and precision effects

The test searching thesaurus was relatively small in the conceptual coverage. Therefore the thesaurus could not give terms for extending the basic mode search either by synonyms or by related terms for some original queries. Recall and precision are therefore evaluated separately for all the 26 original queries (with 26 basic searches, 23 synonym searches and 21 related term searches) and for the subset of those 18 original queries for which all three search modes were available.

The results for all 26 original queries are given in Table 1. These findings indicate that the precision of the basic search was about 50% and it deteriorated by 10% in the synonym searches and by some 17% in the related term searches. As expected, the relative recall figures improved when the precision figures deteriorated. The basic mode searches retrieved, on the average, only 45% of the articles considered relevant in the related terms searches. The inclusion of synonyms improved recall by 36% (at the cost of 10% decline in precision) and the inclusion of related terms by additional 18% (at the cost of further 17% decline in precision). These effects in relative recall and precision are statistically significant both between the basic search and the synonym search and between the basic search and the related term search (see Table 4 in Section 3.3 below for significance tests).

The total number of articles retrieved for all 26 original queries varied from 1 to 51 per search. The averages of total number of articles retrieved and relevant articles retrieved were 9,3 and 4,0 for the basic search mode, 21,3 and 8,9 for the synonym search and 28,0 and 8,6 for the related term search. (NB: the average of the precision figures of individual searches is different from the figure obtained by dividing the average number of relevant articles retrieved by average number of articles retrieved and multiplying this by 100; e.g. in the basic mode average precision = 50,6 ≠ 100 x 4,0 / 9,3 = 43,0).

Table 1. Average percentages of relative recall and precision for the three search modes : the basic search, the synonym search and the related term search (all 26 original queries).

	Basic Search (N=26)	Average % Synonym Search (N=23)	Related Term Search (N=21)
Relative Recall	45,2	81,8	100,0
Precision	50,6	40,6	33,1

Table 2. Average percentages of relative recall and precision in the subset of 18 queries for which all the three search modes : the basic search, the synonym search and the related term search were evaluated.

	Basic Search	Average % Synonym Search	Related Term Search
Relative Recall	44,4	76,8	100,0
Precision	48,5	40,2	35,0

The analysis of the subset of those 18 original queries for which all three search modes (see Table 2) shows similar results. The increase in the relative recall seems here to be slightly better than in the set of 26 original queries. Precision seems to deteriorate slightly faster in this subset than among all original queries. The averages of total number of articles retrieved and relevant articles retrieved were 9,2 and 4,1 for the basic search mode, 18,8 and 7,6 for the synonym search and 28,4 and 9,1 for the related term search.

3.2 The net effect of query extension types

The net effect of query extension types is shown in Table 3. The net effect of the synonym searches was such that they brought almost one third of the total relevant articles and almost every third article retrieved was judged relevant. The net effect of the related term searches was such that they brought almost one quarter of the total relevant articles but only less than one out of six articles retrieved was judged relevant. The averages of total number of articles retrieved and relevant articles retrieved were 9,2 and 4,1 for the basic search mode, 9,5 and 3,5 for the synonym extension and 9,7 and 1,5 for the related term extension.

Although each extension type retrieved roughly equally many new articles, the number of relevant articles retrieved deteriorated considerably from the basic search to the related term extension. This is clear also from the precision figures. (NB: Here, also, the precision figures are averages of precision figures of individual searches while the counts are sums of total (relevant) articles divided by the number of searches (18). Therefore precision ≠ 100 x relevant articles / total articles).

3.3 Statistical significance of the findings

Statistical significance of the findings are reported in Table 4. Significance is tested both with the t-test and the

Wilcoxon test [24]. Both test types have been used for the testing significance of recall and precision effects in retrieval experiments [23]. For each case of comparison in Table 4, the number of searches involved is given. There were e.g. 23 searches for which both the basic search and synonym search was conducted. Therefore Table 4 gives the significance of the relative recall and precision differences for these 23 searches.

Table 4 indicates that the improvements in relative recall from the basic search to synonym search and related term search are statistically significant with a very small risk level (p < 0,001). Also the deterioration of the precision figures from the basic search to synonym search and related term search are statistically significant (p < 0,01). With this data, the deterioration of the relative recall in the net results of query extension types are not statistically significant. According to the Wilcoxon test, however, the deterioration of the relative recall from the basic search to the related term extension is almost significant (p < 0,05). In contrast, even this small data set shows a statistically very significant (p < 0,001) deterioration of precision from the basic search to the related term extension. The difference in precision between the basic search and the synonym extension is almost significant.

Table 4. The statistical significance of the differences of averages of recall and precision in the three search modes : the basic search, the synonym search and the related term search, and the two search extension types.

Search pairs compared	Test statistic t-test	Wilcoxon-test
RELATIVE RECALL		
Basic Search vs. Synonym Search (N=23)	4,963***	3,530***
Basic Search vs. Related Term Search (N=21)	9,459***	4,218***
Basic Search vs. Synonym Extension (N=18)	1,049	..,[a]
Basic Search vs. Related Term Extension (N=18)	1,665	1,787*
PRECISION		
Basic Search vs. Synonym Search (N=23)	2,771**	2,875**
Basic Search vs. Related Term Search (N=21)	2,676**	2,381**
Basic Search vs. Synonym Extension (N=18)	2,274*	2,275*
Basic Search vs. Related Term Extension (N=18)	4,177***	3,087***

*** p < 0,001 ** p < 0,01 * p < 0,05

[a] The number of non-equal pair compared is less than or equal to 15. The probability of the differences in relative recall is 0,138. This is not statistically significant.

Table 3. Average counts or total articles and relevant articles retrieved and average percentages of proportional recall and precision in the net results the basic search and the search extensions : the synonym extension and the related term extension. (N = 18)

Extension Type	Averages: Counts: Total art.	Relevant Art.	Percentages: Relative Recall	Precision
Basic Search	9,2	4,1	44,4	48,5
Synonym Extension	9,5	3,5	32,4	30,7
Related Term Extension	9,7	1,5	23,2	12,9

3.4. Analysis of false drops

The 493 irrelevant articles found by all the searches were analyzed as to the reasons of irrelevance. The results of this analysis are tabulated in Table 5. The most common sources for irrelevance were search terms not being related in the article and search topic being discussed only marginally in the article. These accounted for 80% of false drops. The use of proximity operators would have reduced the former type of false drops effectively. In this study, word proximity operators were used only for precoordinated phrases of the test thesaurus. The Boolean intersection was used between terms representing different concepts. The latter type of false drops could be reduced by setting term frequency thresholds for the search terms: articles with search term frequencies below the threshold would be rejected. In this study, term frequency thresholds were not tested.

Homonymy did not cause many false drops. However, the results showed instances of false drops due to partial homonymy which is very difficult to avoid in Finnish language. Partial homonymy means that terms with entirely different meanings have some inflected forms that coincide. For example, the expressions “ei lainkaan” [not at all] and “ei millèèn lailla” [in no way] match two inflected form instances of the word “laki” [law]. If one is searching for law issues, one has to account for inflection either by searching truncated terms “laki- or lai- or laei-” to cover inflection or by recognizing the basic forms of the text words for a basic form basic index and searching then for the basic form “laki”. The disjunct “lai-” in the former case matches the irrelevant strings. In the latter case, the basic form algorithm recognizes the noun “laki” as one possible basic form for the strings “lain- kaan” and “lailla”. By searching only for “laki-” one would certainly miss essential articles on the topic. The only

Table 5. The number of original queries and the number of different searches affected by false drops and the number of false drops for different sources of false drops.

Source of false drop	No. of queries	No. of searches	No. of articles
Terms not related	21	40	221
Incorrect term relationships in document	16	22	52
Homonymy	2	5	19
Topic marginal	23	48	174
Search strategy	2	2	11
Incorrect term relationships in thesaurus	3	3	16

remedy to these problems is a linquistically intelligent, context sensitive basic form recognition program either in the storage or in the search phase.

4. Discussion

We have tested the effectiveness of a searching thesaurus in free-text searching of a Finnish full-text database containing newspaper articles. The purpose of such a thesaurus is a) to improve search recall by giving synonyms and other alternative terms for the original search terms (or at least to relieve the searcher's burden in recognizing and keying in all possible synonyms) and b) to enable systematic broadening and narrowing of a search by presenting conceptual relationships.

The test results indicate (see Tables 1 and 2) an essential and statistically significant improvement in recall with neglible but statistically significant decline in precision due to the test thesaurus. This is especially true of the performance of the synonym searches. The findings suggest that augmenting original queries by synonyms provided by a searching thesaurus yields almost twice as many relevant documents as the original queries with a neglible (10%) decline in precision. The effect of the related term searches is more questionable but, if high recall is necessary, an important possibility. These possibilities may not be relevant to many needs of journalists because they may often be satisfied with only a few articles. Nevertheless, they are essential possibilities of search modification in general, e.g. for different kinds of clientele.

The analysis of the net effects of the query extension types (Table 3) corroborate these conclusions. The synonym extension retrieves an essential share of the relevant articles with a precision figure which is often tolerable. The related term extension still adds nearly one fifth of the relevant articles but now the precision is rather poor (13%). This means that the related term extension is worthwhile only when high recall is important. The figures also suggest that the output should be ranked by extension type.

The analysis of the false drops show that word proximity operators and word frequence thresholds could be used to reduce irrelevant results due to search terms not being related in the articles or due to marginal treatment of the search terms in the articles. Partial homonymy in inflectional languages like Finnish is a difficult problem to solve but does not seem to cause many false drops.

Untested possibilities in our study were: extensive use of proximity operations, word frequency thresholds, hierarchical (downward) expansion of abstract terms to cover a broad class of terms, and the efficiency of scalable synonymity. The last possibility means attaching to each synonym relation an index (e.g. in the range 0.0...1.0) describing the degree of the synonymity relation. This could be used in the search phase by allowing only suchsynonyms of original search terms which have a synonymity relation to the original term exceeding a given limit. A study testing these possibilities is under way.

The type of a searching thesaurus we have used gives terms and simple phrases for the terms the user has in his mind. This is a simplistic approach especially for abstract concepts. Concept recognition procedures described in [3] may provide an approach to augmenting searching thesauri in providing assistance for searching abstract concepts. Another issue is related to the users and updating of searching thesauri: should they be tools for a general audience, a professional group, or should they be personalized, user specific semantic tools? In the latter case, the users themselves would be responsible for updating their personal thesauri which therefore might differ even among persons working in the same organization and using the same database.

The user interface to the thesaurus needs also consideration. In this study, the researcher simulated, in effect, the operation of an automatic tool aiding the journalists in searching. When such a thesaurus is provided to the users as a routine semantic tool, it may be available upon request, automatically (i.e. queries are extended automatically through the thesaurus) or semiautomatically (i.e. the thesaurus suggests further terms but the user must approve or reject them).

Although Finnish is a highly inflectional language our results are not language dependent is this respect. We have considered semantic, not morphological terms relationships. Even the searches were conducted in an English-like manner due to the use of the morphological programs yielding basic term forms in the construction of the basic index of the database.

5. Conclusions

We have tested the effectiveness of a searching thesaurus in free-text searching of a Finnish full-text database containing newspaper articles. The test thesaurus improved search recall by giving synonyms and other related terms for the original search terms. The synonyms given by the thesaurus improved the relative search recall, on the average, from 45% to 82% while augmenting the searches also by the related terms given by the thesaurus further improved the relative search recall, on the average, from 82% to 100% (the maximal relative recall of all the searches). The average precision values were for the modes one, two and three were 51%, 41% and 33%, respectively. These findings were statistically significant and are not bound to the properties of the Finnish language.

The substitution of synonyms given by the thesaurus to the original search terms shows a considerable increase in recall with a negligible decline in precision. Further query augmentation by related terms seems appropriate, if high recall is required. The analysis of the net effects of the query extension types showed an essential improvement in recall (nearly 40% of all relevant articles) with a relatively high precision (31%) for the synonym extension. The related term extension showed further improvement in recall (nearly one fifth of all relevant articles) with a relatively poor precision (13%). These findings suggest that the search results be ranked by the search extension type.

In conclusion, searching thesauri seem to be effective tools for improving search results in free-text searching of full-text databases. The findings and the untested possibilities for search development through searching thesauri suggest the following topics to be elaborated in further studies: searching thesauri for different types for full-text databases and different user populations, extensive use of proximity operations, word frequency tresholds, hierarchical expansion of abstract terms to cover a broad class of terms, the efficiency of scalable synonymity relations, and concept recognition procedures for text processing.

Reference

[1] Blair, D.C.: Full text retrieval: Evaluation and implications. Int. Classif. 13 (1986) No.1, p. 18-23.

[2] Blair, D.C., Maron, M.E.: An evaluation of retrieval effectiveness for a full-text document retrieval system. Comm.ACM 28(1985)No.3, p.289-299.

[3] Croft, W.B.: User-Specified Domain Knowledge for Document Retrieval. In: Rabitti, F. (Ed.): ACM Conf. Res.& Developm.in Inform. Retrieval, Pisa, Italy, Sept.8-10,1986, p.201-206.

[4] Directory of Online Databases 9(1988)No.1. New York, NY: Cuadra-Elsevier 1988.

[5] Dubois, C.P.R.: Free text vs controlled vocabulary: a reassessment. Online Rev. 11(1987)No.4, p.243-253.

[6] Fugman, R.: The complementarity of natural and indexing languages. Int. Classif. 9(1982)No.3, p.140-144.

[7] Hersey, D.F., Foster, W.R., Stalder, E.W., Carlson, W.T.: Free text word retrieval and scientists indexing: Performance, profiles and costs. J. Document. 27(1971)No.3, p.167-183.

[8] Henzler, R.G.: Free or controlled vocabularies: Some statistical user-oriented evaluations of biomedical information systems. Int. Classif. 5(1987)No.1, p.21-26.

[9] Iivonen, M.: Indeksointituloksen riippuvuus indeksointiympèristÜste [The dependence of indexing results on the indexing environment]. Tampere, Finland: Univ.of Tampere, Dept. of Libr. & Inform. Sci., Report 26, 1989. 200 p.

[10 ISO 2788. Documentation – Guidelines for the Establishment and Development of Monolingual Thesauri. Paris: Int. Org.for Standardisation 1986. 65 p.

[11] Jèppinen, H., Ylilammi, M.: Associative model of morphological analysis: An empirical inquiry. Comput. Linguistics 12(1986)No.4, p.257-272.

[12] Kimberley, R. (Ed.): Integrating text with non-text. Proc. Inst.Inform.Scientists Text Retrieval '85 Conf. London, UK: Taylor Graham 1986. 120 p.

[13] Kimberley, R. Hamilton, C.D., Smith, C.H. (Eds.): Text Retrieval in Context. Proc.Inst.Inform. Scientists Text Retrieval '84 Conference. London, UK: Taylor Graham 1985. 152 p.

[14] Lehtola, A., Honkela, T., HyÜtyniemi, H. and Jèppinen, H.: Task Oriented Knowledge Representation Languages for NLP-Systems. In: Ras, Z. W., Saitta, L. (Eds.): Methodologies for Intelligent Systems 3. Proc. 3rd Int. Symposium on Methodologies for Intelligent Systems. Turin, Italy, 12-15 Oct., 1988. New York, NY: North Holland 1988. p.250-259.

[15] Lancaster, F.W.: Vocabulary Control for Information Retrieval. Washington: Inform. Resources Press 1972.

[16] Lancaster, F.W.: Vocabulary Control for Information Retrieval. (2nd ed.) Arlington, VA: Inform.Resources Press 1986.

[17] McDonald, M.: BASIS – Innovation in Data Management. In: Chen, C.C., Hernon, P. (Eds.): Numeric Databases. Norwood: Ablex 1984. p.219-236.

[18] Meunier, J.-G., Bertrand-Gastaldy, S., Lebel, H.: A call for enhanced representation of content as a means of improving online Full-Text Retrieval. Int. Classif. 14(1987)No. 1, p.2-10.

[19] Perez, E.: Text enhancement: controlled vocabulary vs. free text. Spec.Libr.s 73(1982)No.3, p.183-192.

[20] Piternick, A.: Searching vocabularies: A developing category of online searching tools. Online Review 8(1984)No.5, p.441-449.

[21] Ro, J.S.: An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval: 1. On the Effectiveness of Full-Text Retrieval. J. Amer.Soc. Inform.Sci.39(1988)No.2, p.73-78.

[22] Salton. G.: Another look at automatic text-retrieval systems. Comm.ACM 29(1986)No.7, p.648-656.

[23] Salton, G., McGill, M.J.: Introduction to modern information Retrieval. New York, NY: McGraw-Hill 1983. 448 p.

[24] Siegel, S., Castellan, N. J.: Nonparametric statistics for the behavioral sciences. New York, NY: McGraw-Hill 1988.

[25] Soergel, D.: Organizing Information: Principles of database and retrieval systems. New York, NY: Academic Press 1985. 450 p.

[26] Swanson, D.R.: Searching natural language by computer. Science 132(1960)No.3424, p.1099-1104.

[27] Swanson, D.R.: Historical Note: Information Retrieval and the Future of an Illusion. J.Amer.Soc.Inform.Sci. 39(1988)No.4, p.92-98.

[28] Svenonius, E.: Unanswered questions in the design of controlled vocabularies. J.Amer.Soc.Inform.Sci. 37(1986) No.4, p.331-340.

[29] Tenopir, C.: Full text databases. In : Williams, M.E. (Ed.): Ann.Rev.Inform.Sci. & Technol., Vol 19. White Plains, NY: Knowledge Industry Publ. 1984, p.215-246.

[30] Tenopir, C.: Full text database retrieval performance. Online Rev. 9(1985)2, p.149-164.

APPENDIX

Sample Extracts from the Test Searching Thesaurus

The extracts below are not accurate translations fron the Finnish test thesaurus. Rather, they illustrate the test thesaurus. Differences in synonyms and their number between Finnish and English are taken into account. The translations given below may not always match most common terms and phrases in English newspapers. In many cases, Finnish journalists use both Finnish terms and terms borrowed from English for a single concept in their texts. Therefore the extracts below are not as rich with synonyms as the Finnish version of the thesaurus is.

The thesaurus was restricted to economic articles of the newspaper and followed the literary warrant principle.

The thesaurus relations are reciprocal although this is not shown in the extracts.

Legend

SYN	synonymous terms
BT	broader terms
NT	narrower terms
RT	related terms
SEE	see the desrciption at the referred main term entry
...	Finnish terms with no known English equivalents

SAMPLE EXTRACTS

annual report
SYN annual review, annual overview, activity report,...
RT profit, balance sheet,...

annual review
SEE annual report

atomic energy
SEE nuclear energy

atomic energy production
SYN nuclear energy production
RT nuclear energy, nuclear power, atomic energy, atomic power, energy, nuclear power plants, nuclear power stations

automatic data processing
SYN EDP, electronic data processing
RT computer, computing, microcomputer, information technology

balance sheet
SYN profit,...
RT annual report, annual review, annual overview, activity report,...

factory
SYN industrial establishment, mill, production stablishment
RT industry

import
BT trade
RT cast-trade, soviet-trade, west-trade, eec-trade, foreign trade, export, order, shipment

industry
NT apparel industry, clothing industry, dockyard industry, forest industry, process industry, ship industry, shoe industry, textile industry, export industry,...
RT company, factory, firm, industrial establishment, mill,...

information technology
RT EDP, automatic data processing, electronic data processing, computer, computing, microcomputer

markets
NT home markets, world market, internal markets, foreign markets, export markets, common markets
RT trade, marketing, purchases, sales

nuclear energy
SYN nuclear power, atomic energy, atomic power
BT energy
RT nuclear power plants, nuclear power stations, nuclear energy production, atomic energy production

paper industry
SYN paper mills
BT chemical forest industry
RT paper, cellulose, carton, cardboard

textile
SYN apparel, suit, clothes, clothing
RT apparel industry, shoe industry, textile industry

textile industry
SYN apparel industry, clothing industry,...
BT industry
RT apparel, suit, clothes, clothing, textile

trade
NT bilateral trade, international trade, compensation trade, world trade, foreign trade, import, export...
RT markets, marketing, purchases, sales, orders, shipments

Address: Prof.Kalervo Järvelin and Ms. Jaana Kristensen
Department of Library and Information Science. University of Tampere, P.O.Box 607, SF-33101 Tampere, Finland