USE OF KEYPHRASE EXTRACTION SOFTWARE FOR CREATION OF AN AEC/FM THESAURUS

SUBMITTED: November 1999
REVISED: February 2000
PUBLISHED: March 2000 at http://itcon.org/2000/2/
EDITOR: B-C. Bjoerk

Branka Kosovac, PhD candidate,
University of British Columbia;
email:branka@civil.ubc.ca

Dana J. Vanier, PhD,
National Research Council Canada;
email:Dana.Vanier@nrc.ca

Thomas M. Froese, PhD,
University of British Columbia;
email:tfroese@civil.ubc.ca

SUMMARY: *The paper describes a method used to collect terms needed for the development of a thesaurus in the roofing domain. This work is part of a larger effort to investigate the potential of thesauri as an aid in product modeling and as a tool for information management in model-based systems. Extractor, a software module that extracts keyphrases from documents, was used for collecting candidate thesaurus terms from Internet sources. The principal advantage of the Internet as a source of candidate terms is that it reflects the language that is actually used in communications concerning buildings and that it covers the widest range of different views on the domain. The advantage of using Extractor or similar software is that it allows processing huge text corpora available on the Internet while eliminating irrelevant terms. The methodology used was found to be highly useful, although it was not sufficient by itself for constructing a thesaurus for the architecture, engineering, construction and facilities management industries, as considerable human intervention was required. Some possibilities for customizing the software and for partially automating a thesaurus construction process are suggested.*

KEYWORDS: *thesauri, Internet, automatic indexing software, thesaurus construction.*

1. INTRODUCTION

A thesaurus is a set of terms that are used in a specific domain of knowledge, "formally organized so that a priori relationships between concepts are made explicit" (Aitchinson and Gilchrist, 1987). Originally intended for indexing and retrieving documents, thesauri have increasingly been seen as knowledge bases and used beyond the domain of librarianship (Kosovac, 1998). The overall objective of this research direction is to investigate the potential of thesauri to assist the development and use of product models in the architecture, engineering, construction, and facilities management (AEC/FM) industries. It is believed by some (Vanier, 1994) that thesauri can help solve many of the semantic problems hindering the creation of robust product models for the industry and the efficient use of model-based systems. The goal of this project was the development of a prototype thesaurus in the roofing domain, and more specifically, for low slope roofs. One of the purposes was to explore the use of the Internet as a source of candidate thesaurus terms. The paper describes this particular aspect of the project — the method used to collect terms that are used in the low-slope roofing domain.

Extractor (Extractor 2000) is a machine-learning-based software module, developed by the Interactive Information Group of the National Research Council Canada, that scans an electronic document and extracts keyphrases best describing the document's subject matter. In this project, the authors used Extractor 2.0 as a support tool for collecting and selecting terms to be included in the proposed thesaurus.

In the process described, Extractor was used for a specific task and under given circumstances. Time and resource constraints did not allow full exploitation of Extractor's capabilities. These constraints also precluded

testing corpora (document collections) large enough to provide statistically valid results. Therefore, the work described is of explorative nature and cannot be considered as a study that evaluates the performance of Extractor. However, the patterns noticed in the analysis of the results can point to possible use of the software for related purposes and suggest possibilities for further research and development.

1.1 Problem description and general approach

Pragmatic analyses of the domain, available thesauri, and intended applications of the proposed thesaurus indicated that it should follow the format of the Canadian Thesaurus of Construction Science and Technology (TC/CS), a thorough, comprehensive, yet "dated" construction thesaurus (TC/CS 1978). One approach for developing a compatible microthesaurus is as follows:

- extract selected terms from the TC/CS to form a sectorial thesaurus,
- update the sectorial thesaurus to reflect current technology and terminology, and
- develop a micro-hierarchy of narrower concepts, if required.

Because of the "datedness" of the TC/CS (1978), it was decided that the following "bottom up" approach was more suitable:

- collect terms relevant to the field,
- select terms to be included in the thesaurus according to its intended purpose,
- check the terms against the existing TC/CS thesaurus,
- organize the terms into hierarchies following the TC/CS guidelines.

Therefore, the first phase of our work involved collecting candidate thesaurus terms. Typical sources for the initial collection of terms include:

- terminological sources in standardized form: existing thesauri, dictionaries, glossaries, classification schedules, encyclopaedias, lexicons, journal indices, back-of-the-book indices, term lists, treatises on terminology of a subject field;
- literature scanning;
- question (i.e., user query) scanning;
- users', subject experts', and compilers' knowledge (Aitchinson and Gilchrist, 1987).

Literature and question scanning play a crucial role for the usability of a thesaurus while the other sources serve mostly to clarify the meanings of terms, facilitate their arrangement into hierarchies, and fill gaps.

1.2 Use of information technology in thesaurus construction

Along with their use for thesaurus-management, computers have long been used for collecting thesaurus terms (Gilchrist 1971, Lancaster 1986, Aitchinson and Gilchrist, 1987). Their main use has been for automatic scanning of either titles and abstracts or full text documents from large textual databases and ranking the derived terms by frequency of occurrence. The explosive growth of the Internet has provided huge and diverse machine-readable corpora that have been used for a variety of purposes in computational linguistics and natural language processing. However, little work has been reported to date where Internet documents were used for extracting candidate thesaurus terms.

Since the mid-eighties, there have been considerable research efforts related to automatic extraction of semantic relationships and automatic construction of structured thesauri, e.g. Chodorow et al. (1985), Fox et al (1988). Despite some claims of successful R&D projects, such as MindNet (Richardson et al., 1998), wide use of full automation in thesaurus construction is still unrealistic. One simple but important reason is that the procedure requires special corpora — in almost all projects machine-readable dictionaries have been used. As a comprehensive, authoritative and up-to-date source for a certain subject domain is rarely available, it can be expected that computers will still be used only as support tools in this part of the process for a time to come.

1.3 Use of thesauri in product model-based systems

While there are many uses for technical thesauri, this project is aimed at exploring their application to product models and model-based integrated systems for AEC/FM: systems that communicate and work with object-oriented, semantically rich representations of project information. This paper focuses on the creation rather than the use of thesauri, but a few observations about the potential role for thesauri in model-based systems can be made.

Thesauri address the general task of mapping related concepts, a task that arises frequently in model-based systems. They provide formalisms and approaches for mapping concepts, such as the addition of relative specificity of relationships (e.g., one term might be described as related to, but more narrowly defined than, another term). Thesauri can be used to map not only related words and phrases, but to map a wider range of representational elements such as semantic models, heterogeneous forms of computer-based data, etc. Roles in which thesauri may be useful in model-based systems include the following:

- Basic terminology: Data models must use specific, unambiguous terms for each concept represented in the model. Real world language, and even the terminology used within the set of all AEC/FM computer applications, is much less precise, and a wide variety of terms may commonly be used to refer to the same concept. Thesauri can help map the terminology used by different applications and users to the terminology used within a shared product model.
- Schema: Information sharing in distributed model-based systems relies on common, standard data models (schemas). In practice, however, many different schema exist that must be mapped to each other. Examples include the mapping of application schemas to common shared schemas, accommodation of different views and perspectives representing various disciplines or users, and the generation or interpretation of new semantics from a project database. While research and development is progressing in the area of mechanisms for mapping schema (such as the Express-X language), several aspects of this mapping relate to thesauri issues and techniques.
- Heterogeneous representations: Within distributed model-based AEC/FM systems, information about a single physical component of a building may be represented in numerous forms, such as textual and numerical properties of a software object, data values dynamically calculated on-demand, 2D and 3D geometry in one or more formats (e.g., surface models, solid models, etc.), digital multi-media information such as photographs, hyperlinks to external references, and so on. Techniques must be developed to inter-relate all of these data forms at varying levels of granularity and to manage the heterogeneous data sets. Again, while this is beyond the scope of traditional thesauri applications, the underlying requirements are very similar and thesauri-based approaches are likely to provide useful contributions to the required capabilities.

Further work in this research direction, including the use of the developed thesaurus, will be reported in subsequent papers.

2. PROJECT

2.1 Approach

The (TC/CS 1978) is a huge thesaurus (approximately 15,000 terms) covering a wide subject field. Searching it for all terms relevant to the subdomain would be an expert-labour-intensive process of following links throughout different hierarchies and numerous general terms, and deciding which terms are relevant and current. Possibilities for automating this task are minimal. Another problem is that the TC/CS is rather outdated, especially considering the significant changes in the field of low slope roofing that occurred during the 1980s, with the introduction of new materials and types of roofing systems.

On the other hand, the TC/CS has a thoroughly elaborated structure and well defined inter-term relationships that facilitate the addition of new terms, given that their exact meaning is known. Furthermore, it is available in electronic form on the World Wide Web (http://www.cisti.nrc.ca/irc/thesaurus/.) thus allowing easy searching for known terms. For these reasons the "bottom up" approach suggested earlier seemed to be the most appropriate approach for this work.

As one of the primary intended uses of the proposed thesaurus is indexing and retrieval of Internet/intranet sources, the most useful source of terms would be corpora available on the Internet. The Extractor 2.0 documentation pointed to the suitability of the software for performing "literature scanning" of Internet sources as it integrates HTML and e-mail filters and permits processing of large corpora by extracting only relevant terms.

2.2 Preparatory tests

Since Version 2.0 of Extractor had been recently released at the time of this work (Version 5.1 has now been released), and since there were no similar previous efforts to build upon, the development of the methodology required some initial tests to roughly assess available sources and Extractor's behaviour. The tests were done on documents retrieved by general search services, such as Altavista (http://www.altavista.com) and the services listed below, and on documents from selected Web sites. The documents were processed by Extractor 2.0 varying the setting for the number of keyphrases to be extracted. After analysing Extractor's output, the results and the limited resources necessitated some modifications to the initially considered strategy.

First, the query initially used to test harvesting corpora using automatically generated queries that combine synonyms of the top term (*summum genus*) and its immediate narrower terms was as follows:

("low-slope roof" OR "flat roof*") AND ("built up" OR BUR OR "multi ply" OR "single ply")*

(where BUR is an abbreviation for built-up roofs). This query did not provide optimum recall and precision within some search services as processing a sample large enough to compensate for the deficiencies was unrealistic. The query was thus modified to the following:

("flat roof" OR "low slope") AND ("built up" OR BUR OR roofing OR membrane*)*

This better reflected the content and the language of the relevant documents. Although the new query did not eliminate all the "noise", nor did it ensure absolute recall, the retrieved sets of documents seemed acceptable for the purpose.

Second, the initial strategy was to process documents using the lowest setting for the number of keyphrases, identify terms that should be added as stop phrases (i.e., terms such as "roofs" or "roofing" that appeared as keyphrases in most of the documents preventing extraction of more specific terms), cluster documents based on the rest of the keyphrases, process them with the highest setting for the number of keyphrases, and repeat the process until the desired level of specificity was achieved. However, the inability to frequently customize the software by adding new stop phrases and the labourious task of processing the same documents more than once made this strategy unfeasible. Using the new release with the maximum setting for the number of keyphrases derived and simply removing the most frequent terms proved to yield satisfactory results.

It was observed that long documents, which tended to abound with very specific terms, yielded only very general terms in the Extractor output. Although keyphrases derived by Extractor reflected well the subject of the documents, they did not include specific terms that would be more useful for the purpose. Efforts to automatically divide long texts into meaningful sub-documents did not prove feasible with most of the Web documents. It was, therefore, done only on a small number of scientific papers that tended to be well organized and have a better HTML structure, thus allowing easy division by searching for heading tags.

2.3 Methodology

The immediate goal of the work is to extract relevant terms from Internet documents—not to evaluate Extractor 2.0. However, Extractor's performance has been continuously evaluated after each step in order to re-design the methodology or even give up the use of Extractor 2.0 if it would not produce the desired output. Though many of the tasks had to be performed manually, the methodology was structured in a fashion that would lend itself to automating the process or at least the use of clerical instead of expert labour.

2.3.1 Corpora

The following collections were selected to serve as the initial document set:

- Documents retrieved by general search services

The first 15 documents retrieved by advanced search with each of five major general search services (AltaVista, Excite, HotBot, Infoseek, and Lycos) were used, 75 documents altogether. Although most of the search services perform better with other search options, for consistency, the following Boolean query or the closest allowable option was used in each search:

("flat roof" OR low-slope) AND (built-up OR BUR OR roofing OR membrane*)*

The services were searched in alphabetical order, taking care to avoid duplicates. Where relevant documents from a certain site were grouped together, only the first one was used in order to avoid language of one author and frequent appearance of the same corporate names and trademarks.

- IRC Roofing Resources collection (Roofing Resources 1998)

This collection was included as it represents the core of the collection to be searched by the future thesaurus. Files bigger than 20 KB (arbitrarily established limit) were divided by headings to form 40 documents for the extraction of keyphrases.

- Relevant documents retrieved at FacilitiesNet (FacilitiesNet 1998)

The criterion for the inclusion of this site was its high content of documents relevant to the facilities-management aspect of flat roofs. This aspect is represented only in a minor portion of documents available on the World Wide Web but is important to the wider context of this work. Sixty-one relevant documents were retrieved and processed.

- Collection of selected articles

This collection was compiled by following links from various lists of relevant sources but without aspirations to be comprehensive, exhaustive, nor of highest quality. It has been included to allow comparison of the results with those achieved by automatic harvesting of sources using general search services.

Messages from newsgroups archives were not processed separately but a small number of this type of document was included in AltaVista hits.

2.3.2 Procedure

Each document was processed by Extractor 2.0 with the number of keyphrases set to maximum. . The extracted terms were gathered in a list. The following information was recorded for each term in the Extractor list:

- relevance factor number (provided by Extractor),
- document from which the term was extracted,

and for each document:

- collection ID,
- size of the file.

After processing each set of documents, the results were analyzed, compared with other sets, and the sets were integrated and processed together. The final set of keyphrases was reviewed and compared to the (TC/CS 1978) and existing glossaries.

2.3.3 Criteria

Single- and multiple-word terms were not treated separately, as the list was of modest length. Singular and plural forms of the same term were counted together but terms that may have different meanings when used in plural (e.g. roofing--roofings) were specially marked [PL]. The terms were first searched in TC/CS for exact matches [=]. Qualifiers from the thesaurus (i.e. auxiliary terms used to make the meaning of the term unambiguous) were ignored in this step. Terms identified as general terms in the thesaurus, i.e. terms that do not have a fixed hierarchical level in TC/CS but can be associated with terms of varying degrees of specificity (e.g. "life" or "requirements") were additionally marked [GT]. So were those that had a further developed hierarchy of narrower and/or part terms in TC/CS [+].

After the identification of exact matches, terms from the list were also searched for occurrence as:

- [PH] phrases from the thesaurus
- [Q] qualifiers in the thesaurus.
- [~] close matches, meaning terms having the same stem.

Analysis of the remaining terms identified:

- [A] acronyms;
- [N] proper names;
- [*] phrases ill defined by Extractor (meaning that they could not stand alone either as terms in the thesaurus—which includes only nouns and noun phrases—or as meaningful keywords for a document—e.g. "requiring", "single", or "install"); and
- [$] matches that have the same form but different meaning in the thesaurus.

The appendix A contains a marked list of terms that were extracted as keyphrases from more than two documents.

3. RESULTS AND DISCUSSION

Terms with extremely high frequency of occurrence, which should have been made stop-phrases, were identified early in the process. These terms were the same for each set of documents and also could have been identified by processing the list itself in Extractor 2.0 with the number of keyphrases set to the minimum..

The results from the general search services were then compared with those from selected collections. There were no significant differences noted in the relevance of extracted keywords that would justify the laborious task of searching, evaluating, and selecting sources. The quantity and diversity of documents that can be easily retrieved by general search services can successfully compensate for the quality of selection. As the first 20 documents from the compiled collection described in section 2.3.1 did not bring new terms to the list, this collection was not further processed and it is not included in the final results. However, the documents have been saved for later comparison of scholarly and natural language terms and for exploration of specific sub-areas.

The final list consists of 1054 terms (2423 occurrences) extracted from 176 documents. Almost half of the terms extracted were single words (49 %). They accounted for almost all of the top 4% most frequent terms with only two exceptions that would normally be included in stop-phrases. The usual practice of treating such terms separately was not followed as most of the terms were also identified as single word terms in the TC/CS.

A huge number of single occurrences of a term (78%) can be explained by the insufficient size of the sample. In order to extract relevant terms from this group, they were searched for component words and stems that could also be found in other terms. Terms found in this way were ranked higher in the list as more relevant. Rough scanning of the remaining single-occurrence terms found very few terms that were relevant to the field. For that reason, these terms were excluded from further processing. It is important to note, however, that this group contained a substantially smaller percentage of terms marked as ill-defined, significantly contributing to the overall average of only 1% ill-defined terms.

3.1 Comparison with the existing thesaurus

Checking the final list of terms against the TC/CS revealed a high relevance of the terms extracted by Extractor 2.0 to the field. Among 130 terms that appeared more than twice, 56% had exact matches in the thesaurus, 12% were found only in phrases, and 6% were marked as close matches. This breakdown is illustrated in Fig. 1. Terms that occurred twice had even more exact matches (63%) in the thesaurus but were less frequently found in phrases as they included more multi-word terms.

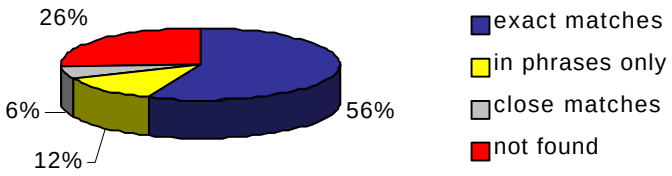

FIG.1: List of terms with more than two occurrences compared with TC/CS

Fig. 2 shows the breakdown of terms with more than two occurrences that had no matches or close-matches found in the existing thesaurus — 12% (or 3% of all the terms extracted) were proper names, 8% (2% of all the terms) were acronyms, and 19% (or 5% of all the terms) were marked as ill defined.

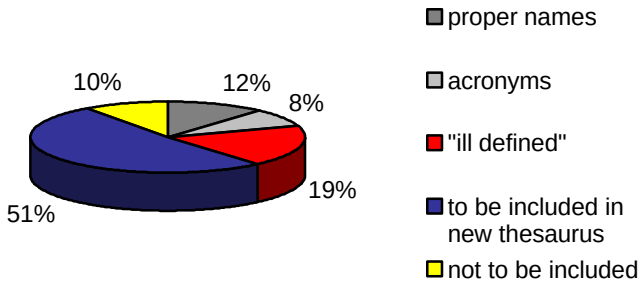

FIG. 2: Breakdown of terms not found in TC/CS

The majority of mismatches, however, do not indicate irrelevance of the terms but more often the outdatedness of the TC/CS. The frequent occurrence of the term "membranes" for example, and the phrases containing "membranes" that are not found in TC/CS reflects changes in the field of low slope roofing and its terminology. The noise-making terms, i.e. extracted terms that do not belong to the low-slope roofing domain, come from specific kinds of documents, mostly glossaries and book catalogues. Such documents can be easily excluded from the beginning by modifying the initial query.

The relatively high coincidence of terms may, on the other hand, indicate a lack of more specific terms that would be required for developing a microthesaurus. Whether this is the case can be established only later in the process.

The matches were found in all semantic classes and in various hierarchies, showing a broad coverage of the domain. Completeness of coverage, however, is yet another problem that cannot be properly evaluated at this point but only after organizing terms into hierarchies (Petersen 1990).

3.2 Comparison with the glossary

The terms were also matched against relevant (i.e. low-slope-roofs related) terms extracted from a roofing glossary (Biegel 1989). The coincidence was significantly lower; only 31 % of the glossary terms were found in the Extractor's output. The unmatched terms were mostly very specific terms (e.g. "back-nailing" or "cutoff"). Since such narrow concepts usually do not represent a document topic, it cannot be expected that keyphrase generation software will retrieve the terms denoting them.

4. CONCLUSIONS AND RECOMMENDATIONS

4.1 Observations on performance of Extractor

Extractor 2.0 appears to be a suitable tool for collecting thesaurus terms from the Internet. It can be used on any PC running Windows 95/98/NT, or on Linux or Solaris systems. Although its use in this project required extensive manual work, it is estimated to be more efficient than both manual and automatic full-text literature scanning. The principal advantage is that it allows scanning and handling of a significantly larger number of documents, thus providing better coverage of the field and its terminology. In addition, Extractor can be easily embedded in other software. This feature, which was not used in this project, allows further automation of many of the described tasks.

The number of phrases marked as ill-defined (see the explanation in section 2.3.3) was 1% of all the extracted terms. Since the total automation of the thesaurus constructing process is not considered and since ill-defined phrases cause no serious consequences, this percentage can be considered negligible. Therefore, increasing the number of keyphrases even above the maximum that was possible in version 2.0 (in order to retrieve more specific terms) would probably be safe. In most cases the lack of more specific terms in the Extractor output will not represent a deficiency; terms too specific to be the subject of a document are rarely included in a thesaurus. In addition, their presence might make one of the most important decisions in thesaurus constructing — where to stop — even more difficult. However, if constructing a microthesaurus, or if for any other reason more specific terms are needed, these terms may be obtained by processing larger corpora or by narrowing the searches for the Internet documents to be processed.

4.2 Observations on the Internet as a source of thesaurus terms

The Internet represents an extremely useful source of thesaurus terms. It provides huge corpora covering numerous aspects of a domain and different vocabularies—from the highly scholarly to the most informal. Internet documents reflect the language that is current, actually used in the field, and most likely to be used in queries. The results also showed that documents randomly harvested using general search services could provide equally valuable terms as controlled subject collections. The collected documents can be further analysed to complement the results of the described methodology.

4.3 Implications for the further work on the thesaurus

After establishing relationships between terms and organizing them into the hierarchies of the existing thesaurus, it can be expected that certain areas would need further development. Upon identification of such areas, gaps will first be filled with the following:

- terms derived by Extractor from new sets of documents retrieved by more specific terms,
- terms manually extracted from documents already retrieved and judged as relevant to the subfield according to keyphrases derived by Extractor.

The use of these two sources is prioritized for the reasons listed in section 4.2. However, these sources cannot ensure comprehensive coverage of the domain. Therefore, manual extraction of terms from alternative sources will probably be needed. Encyclopaedic and textbook type documents, roofing manuals, various kinds of term lists, and architectural details' labels are expected to best serve the purpose.

4.4 Possibilities for automating the process

Some of the tasks that could be fully or partly automated in applications used for similar purposes include the following:

- automatic retrieval of relevant documents from the Internet and their processing with Extractor,
- periodical processing of the list by Extractor for finding terms with significantly high occurrence and making them stop-phrases,
- ranking terms by frequency of occurrence,
- exclusion of terms that occur only once in large corpora,

- automatic exclusion of geographic names ,
- grouping of terms containing the same words or stems,
- grouping of terms by co-occurrence in documents,
- suggesting inter-term relationships by co-occurrence and syntax.

5. FINAL NOTES

The work described was carried out in February 1998. In the meantime, a 336-terms pilot thesaurus has been developed (http://www.nrc.ca/irc/thesaurus/roofing). As these terms represent only a very small portion of the domain and of the terms collected in this process, final conclusions on the usefulness of the methodology cannot be drawn yet.

Since the time of the study, several new versions of Extractor have been released. Anyone interested in implementing the method described in this paper should review new features of the software by visiting the Extractor Web site or by contacting the software's authors.

Original data and full results of the study are available from the authors of this paper.

6. ACKNOWLEDGEMENTS

The authors wish to thank the Institute for Information Technology of the NRCC for the use of Extractor 2.0, and more specifically, to thank the software's author, Dr. Peter Turney, for his support and encouragement. The authors also wish to acknowledge the Canadian Institute for Scientific and Technical Information and the Institute for Research in Construction, both of the NRCC, for their financial contributions to this research. The authors thank both Dr. Ferrers Clark and Mr. Scott Mellon from these respective organizations for their support and encouragement during the course of the research. The authors also acknowledge the work of Prof. Colin Davidson and the IF Group in the development of the TC/CS Thesaurus.

7. REFERENCES

Aitchinson, J. and Gilchrist, A. (1987). *Thesaurus Construction*, Aslib, London, UK.

Biegel, S. (1989). Roofing Materials, *Encyclopedia of Architecture, Design, Engineering & Construction*. Vol. 4, American Institute of Architects, 314-319.

Chodorow, M., Byrd R. and Heidorn G. (1985). Extracting semantic hierarchies from a large on-line dictionary, *Proceedings of the 23rd Annual Meeting of the ACL*, 299-304.

Extractor (2000). National Research Council of Canada, Interactive Information Group, Ottawa, Canada. Available from: http://extractor.iit.nrc.ca. [Accessed February 7, 2000]

FacilitiesNet (1998). Trade Press Publishing, Milwaukee, WI, USA. Available from: http://www.facilitiesnet.com. [Accessed November 27, 1999]

Fox, E.A., Nutter J. T., Ahlswede T., Evens M. and Markowitz J. (1988) . Building a large thesaurus for information retrieval, 2nd Conference on Applied Natural Language Processing, Association for Computational Linguistics, (Ballard B., ed.), Bell Communications Research, Morristown, NJ, 101 -108.

Gilchrist, A. (1971). The Thesaurus in Retrieval, Aslib, London.

Kosovac, B (1998). *Internet/Intranet and Thesauri*, Canadian Institute for Scientific and Technical Information, Internal Report, National Research Council Canada, Ottawa, Canada. Available from: http://www.nrc.ca/irc/thesaurus/roofing/report_b.html [Accessed November 27, 1999]

Lancaster, F.W. (1986). *Vocabulary Control for Information Retrieval*, Information Resources Press, Arlington, VA, USA.

Petersen, T. (1990). Developing a new thesaurus for art and architecture, *Library Trends*, Vol. 38, No. 4, 644-658.

Richardson, S. D., Dolan W. B. and Vanderwende L. (1998) . MindNet: acquiring and structuring semantic information from text. Microsoft Research Technical Publications (MSR-TR-98-23). Available from ftp://ftp.research.microsoft.com/pub/tr/tr-98-23.doc [Accessed January 31, 2000]

Roofing Resources (1998). National Research Council of Canada, Institute for Research in Construction, Ottawa, Canada. Available from: http://www.nrc.ca/irc/roofing/roofing.html. [Accessed November 27, 1999]

TC/CS (1978). *Canadian Thesaurus of Construction Science and Technology*, Department of Industry, Trade and Commerce, Government of Canada, Ottawa. Available from: http://www.cisti.nrc.ca/irc/thesaurus/. [Accessed November 27, 1999]

Vanier D.J. (1994). Canadian thesaurus of construction science and technology: A hypercard stack, *Proceedings of the Joint CIB Workshops on Computers and Information in Construction (Montreal, Que., Canada)*, (*CIB Proceedings*, Vol. 165), 559-564.

APPENDIX A

TABLE 1: List of terms extracted as keyphrases from more than two documents Sorted by frequency of occurrence indicated by the number in the first column.

	TERM	**=**	**~**	**GT**	**+**	**PH**	**Q**	**A**	**N**	*****	**PL**	**$**	**NOTE**
158	roofing(s)	=									PL	$	
139	roof(s)	=											
74	membrane(s)					PH	Q						
50	materials	=											
32	installation	=				PH					PL		installation(activity)
32	products	=				PH							products(agents)
27	performance	=				PH							
27	water	=				PH							
26	design	=				PH					PL		
25	insulation					PH							
25	roofing system(s)												
24	requirements	=		GT		PH							
22	asphalt	=											
19	deck(s)					PH							
19	repair		~										repairing
17	maintenance	=	~		+								maintenance(restoring)
17	manufacturer	=											
16	flashing(s)	=			+								
14	coating(s)	=									PL		coating(process)
13	fasteners					PH	Q						
13	inspection	=		GT									
13	joint(s)	=			+								joints(junctions)
13	specifications	=		GT		PH							
12	BUR							A					built up roofings found
12	costs	=											
11	construction												
11	felts	=				PH						$	
11	structures	=											structures(buildings) structures(construction) structures(non building)
11	waterproofing					PH							
11	wind	=				PH							
10	flat roof(s)	=											
10	properties	=		GT							PL		property(quality)
9	components	=					Q						
9	moisture					PH							
9	sheets	=			+								sheets(shape)
9	slope(s)	=					Q						
9	standards	=				PH							
8	industry	=				PH							
8	install									*			
8	projects			GT		PH							
8	replacement					PH							replacement value
8	resistance	=				PH							non-descriptor
8	walls	=				PH							
8	warranties	=				PH							

7	air	=											
7	ballast	=										$	ballast(gravel)
7	building owners	=											
7	flexible membrane												
7	market(s)	=											
7	modified bitumen(s)												
7	panel(s)	=											
7	seams					PH							
6	applications	=				PH						$	
6	built-up roofing	=											
6	consideration												
6	drain(s)					PH							
6	evaluation	=		GT									
6	installing									*			
6	life	=		GT		PH							
6	load(s)	=				PH							
6	protected membranes												
6	roofing contractor(s)												contractors
6	shingles	=											out of scope
6	substrate												
6	temperature	=		GT									
6	vapour barrier(s)	=			+								
5	assemblies	=											assemblies(components)
5	base flashing(s)	=										$	
5	Bitumen(s)	=											
5	built-up roof		~										built up roofings
5	condition(s)			GT							PL		
5	drainage	=				PH							
5	EPDM		~										epdm rubber
5	experience	=										$	ND for "skill"
5	facilities					PH	Q						
5	leak		~										leakage
5	low slope												
5	pressure					PH							
5	preventative maintenance												but *prevention* & *maintenance*
5	protection	=		GT									
5	roof decks		~										roof deckings
5	single									*			
5	thermosets		~										thermosetting resin (ND)
4	built-up									*			
4	coal(-)tar	=											
4	damage	=											
4	flow	=											flow(fluids)
4	layer					PH						$	
4	movement					PH							ND for *motion*
4	recommendations	=		GT									
4	reinforcing		~			PH						$	
4	requiring		~									$	
4	Roofing Contractors Association								N				
4	roofing membranes												
4	thermal insulation	=											
4	vapour retarders												
3	Aduron								N				
3	architect(s)	=											
3	barrier					PH						$	ND
3	base coat(s)											$	
3	building envelope												
3	Canada								N				
3	coating system												but "coating(process)"
3	condensation	=				PH							
3	contract(s)	=											
3	control	=		GT									
3	counter(-)flashing(s)												
3	cracks	=											cracks(fissures) ND
3	differential movement(s)	=											ND
3	edge(s)	=		GT		PH							edge(boundary)
3	equipment	=											
3	flat					PH				*			

3	foot					PH							
3	gravel	=											
3	installer					PH							
3	liner												
3	measurement(s)	=		GT		PH							
3	metal roof		~										metal roofings
3	minimum					PH							
3	National Research Council								N				
3	owner(s)	=				PH							
3	parapet					PH							
3	polymer	=											ND for "polymeric materials"
3	reinforcement					PH							
3	review												irrelevant
3	SPF							A					
3	steel	=											
3	surfacing	=											
3	waterproofing membrane		~										waterproof membranes (ND)

PL terms may have a different meaning when used in plural
= exact match found in TC/CS
~ close match found in TC/CS
GT term represents a general term in TC/CS, i.e. term that does not have a fixed hierarchical level but can be associated with terms of varying degrees of specificity
\+ term has further developed hierarchy of narrower and part terms in TC/CS
PH word found as a part of phrases included as terms in TC/CS
Q term used as a qualifier in TC/CS
A acronym
N proper name
* term "ill-defined", i.e. cannot stand alone as a thesaurus term
$ matches that have the same form but different meaning
ND non-descriptor, i.e. term pointing to a descriptor (authorized term) used as a thesaurus entry