Verification of bibliometric methods' applicability
for thesaurus construction

Jesper W. Schneider

Verification of bibliometric methods' applicability
for thesaurus construction

Jesper W. Schneider

CIP – Cataloguing in Publication

Schneider, Jesper W.

Verification of bibliometric methods' applicability
for thesaurus construction / Jesper W. Schneider. PhD dissertation.
Aalborg: Department of Information Studies, Royal School of
Library and Information Science, 2004. xiii, 356 p. plus
Appendix volume.

Available: <http://biblis.db.dk/uhtbin/hyperion.exe/db.jessch04>

ISBN: 87-7415-286-6

ISBN: 87-7415-286-6

© Copyright Jesper W. Schneider

All rights reserved

Verifikation af bibliometriske metoders anvendelighed i
forbindelse med tesauruskonstruktion

Jesper W. Schneider

In memory of
Jes Viborg Schneider
1934-2003

Acknowledgements

If you want to know how indispensable you are, then stick your finger in a water bowl, pull it up, and see how large a hole it leaves.

– Confucius

Many people have contributed to my PhD work, directly or indirectly, to you I am eternally grateful. You have been an inspiration, you have had patience with me, you have shown me confidence, and you have been supportive and cheerful in good times and bad times.

First, I want to thank my supervisors, Research Professor Peter Ingwersen and Associate Professor Pia Borlund. Peter, I owe the topic of the thesis to you, thanks to your incredible stream-of-consciousness. Besides, your interest in military history is certainly rewarding. Pia, while I owe the topic of the thesis to Peter, I owe you everything else! Thank you – you have been a tremendous supervisor and a dear friend. I am sorry that supervising me meant reading tons of pages, but, if it is a comfort, the red comments were *pearls* of knowledge and they taught me a lot.

Besides my supervisors, I am indebted to a number of colleagues and friends, who have helped and encouraged me before and during my PhD work. Especially Associate Professor Marianne Lykke Nielsen, Assistant Professor Erik Thorlund Jepsen, Head of Department Mona Madsen, and Assistant Professor Birger Larsen. Thank you Marianne, for encouraging me to take up teaching and eventually to apply for a PhD scholarship. Now the circle is completed, as you ended up being head of my assessment committee, for that I am grateful.

I also want to thank the people who took care of me during my study trips abroad. In Tampere, Professor Kalervo Järvelin introduced me to the sauna, a memorable experience indeed – *kiitos*! Likewise, thanks to Eija Airio and Ari Pirkola for helping me with *Connexor*. Thanks also to Rickard Danell, Fredrik Åström, and especially Professor Olle Persson at Inforsk in Umeå. Olle, it was such an inspiration studying at your department in Umeå. But most important, *Bibexcel* rules thanks to your skilful and imaginative inventions. This thesis would have been much different without the opportunity to use *coc*- and *ma*-files! And while we are at it, I want to thank Olle and Professor Sara von Ungern-Sternberg for their willingness to join my PhD assessment

committee together with Marianne. I know it was hard work. I appreciate your constructive critic.

I would also like to thank Research Professor Birger Hjørland for introducing me to the important and related work of Lorna Rees-Potter. Likewise, I would like to thank Professor Dagobert Soergel, for his advice on the use of statistical methods in connection with manual intellectual thesaurus construction.

I am also indebted to the administration and IT-department at the Royal School of Library and Information Science in Aalborg; thank you Lisbeth, Mimi, Dorte, Tove, Jette, Jette Irene, Liselotte, Frank, Bo and Lisbeth. Likewise, I want to thank Susanne Acevedo, our secretary at the Department for Information Studies. Also, I am very grateful for the support the school's library in Copenhagen has given me, and especially the kindness and support Librarian Karen Margrethe Ørnstrup has shown me in the last hectic days before the deadline when pdf-formatting was top of the agenda.

Family and friends have supported me enormously during the last three years – you have accepted my habitual mental and physical absences, my long working hours, and supported me in my grief. Thank you Susan, Jesper, Nicholine, Josephine, Lone, Per, Simon, Sólvor, Ole, Ole, Martin and Henrik. I appreciate that you are still around.

Finally, without the support from two outstanding women, this project would not have succeeded. Mom, I thank you and Dad for the trust you have always shown me, regardless of my occupational habits. This time we succeeded. Maria, I am forever in depth to you. It is incredible what you have come to terms with. Without your love and care – as well as superb housekeeping – I could not have done this PhD. I am very fortunate to be your husband.

Abstract

The present doctoral dissertation work concerns the development and exploration of a semi-automatic thesaurus construction approach based on bibliometric methods.

The main objective of the dissertation is to reintroduce, and further extend, the theoretical and methodological aspects of bibliometric methods to the research area of knowledge organization for the purpose of semi-automatic thesaurus construction.

Thesaurus construction approaches are typically separated into manual approaches and automatic approaches. Albeit, some form of manual thesaurus construction is mandatory due to the relational complexities, semantic ambiguities, and dynamics, inherent in languages. Manual construction and maintenance are complex and time consuming. It is therefore beneficial and necessary to combine manual approaches with automatic approaches due to the complementarity of the two approaches (e.g., Soergel, 1974; Anderson & Pérez-Caraballo, 2001a; 2001b). When automatic approaches are used as a tool for thesaurus constructors, and not as a mean in itself, then we speak of semi-automatic thesaurus construction (Soergel, 1974). This is the foundation for the approach explored in the present dissertation.

In order to pursue the main objective of semi-automatic thesaurus construction, a proposed bibliometric-based methodology of five components is explored as a supplement to manual intellectual thesaurus construction. The methodology is used as a framework for the investigation of the ability of bibliometric methods to identify candidate thesaurus terms and thesaural relationships, as well as to monitor potential terminological and conceptual changes within a specialty area. The bibliometric methods investigated include document co-citation analysis, citation context analysis, co-word analysis, and bibliometric ageing methods. The methodology is explored in a case study of periodontology, a specialty area within dentistry.

The main contributions of the dissertation work are 1) an overall verification of the applicability of co-citation analysis, citation context analysis, and co-word analysis for semi-automatic thesaurus construction; 2) demonstration of the ability of co-citation analysis and citation context analysis to specifically identify important candidate thesaurus terms among a number of potential noun phrases, such terms are important because they are contextual and agreed upon in the scientific community; and 3) demonstration of the ability of co-citation analysis and co-word analysis to detect thesaural relationships between terms that do not, or rarely, co-occur directly with each

other in citation contexts. These results are a direct consequence of the applied bibliometric based methodology.

Consequently, the research reported on in the present doctoral dissertation is a contribution to the development of ‘automatic methods’ as tools for manual intellectual thesaurus construction.

Resumé

Afhandlingens formål er at undersøge anvendeligheden af bibliometriske metoder som et supplement til manuel thesauruskonstruktion. Manuel thesauruskonstruktion er nødvendigt, men processen er kompleks og tidskrævende. Derfor er det hensigtsmæssigt at udvikle værktøjer, der kan supplere den manuelle konstruktionsproces.

Udgangspunktet for afhandlingsarbejdet har været at udvikle halvautomatiske thesauruskonstruktionsmetoder baseret på bibliometri. Til dette formål er der udviklet fem såkaldte komponenter, der hver især undersøger væsentlige aspekter vedrørende thesauruskonstruktion med udgangspunkt i forskellige bibliometriske metoder. De bibliometriske metoder omfatter blandt andet co-citationsanalyse, kontekstanalyse og co-ordsanalyse. Foruden de bibliometriske metoder, benyttes der også et avanceret lingvistisk værktøj i afhandlingen til identifikation af navneordsfraser.

De fem komponenter undersøger tre forhold vedrørende thesauruskonstruktion: 1) identifikation og udtrækning af potentielle thesaurustermer, 2) identifikation af mulige relationer mellem disse udtrukne termer, samt 3) identifikation af berebsmæssige ændringer over tid til brug ved vedligeholdelse af en thesaurus.

De halvautomatiske konstruktionsmetoder er undersøgt i et case studie indenfor tandlægevidenskab, nærmere bestemt parodontologi. Afhandlingens resultater knytter sig derfor til de bibliografiske, bibliometriske, og terminologiske egenskaber, dette domæne udviser.

Afhandlingens hovedbidrag omfatter en verifikation af anvendeligheden af bibliometriske metoder til frembringelse af potentielle thesaurustermer, og til en vis grad, identifikation af mulige relationer mellem disse udtrukne termer. Det er derimod ikke lykkedes tilfredsstillende at identificere begrebsmæssige ændringer over tid ud fra den undersøgte metode.

Afhandlingens resultater bidrager dermed til udviklingen af halvautomatiske konstruktionsmetoder som værktøjer eller supplement til den manuelle thesauruskonstruktion.

Table of contents:

1.	INTRODUCTION	1
1.1	BACKGROUND.....	2
1.2	OBJECTIVES OF THE DISSERTATION	4
1.3	RESEARCH ASSUMPTIONS	6
1.4	RESEARCH QUESTIONS	9
1.5	ORGANIZATION OF THE DISSERTATION	11
2	THESAURUS CONSTRUCTION	15
2.1	INFORMATION RETRIEVAL SYSTEMS	16
2.2	NATURAL VERSUS CONTROLLED LANGUAGES.....	18
2.2.1	<i>An example of a controlled index language</i>	<i>20</i>
2.3	THESAURI	22
2.3.1	<i>Approaches to traditional manual thesaurus construction.....</i>	<i>23</i>
2.3.2	<i>Main steps in thesaurus construction.....</i>	<i>25</i>
2.3.3	<i>Term selection</i>	<i>26</i>
2.3.4	<i>Concept analysis</i>	<i>28</i>
2.3.4.1	<i>Compound terms.....</i>	<i>30</i>
2.3.5	<i>Structuring of concepts and terms.....</i>	<i>33</i>
2.3.5.1	<i>Equivalence relationship</i>	<i>33</i>
2.3.5.2	<i>Hierarchical relationship</i>	<i>34</i>
2.3.5.3	<i>Associative relationship.....</i>	<i>36</i>
2.3.6	<i>Thesaurus displays</i>	<i>37</i>
2.4	SUMMARY.....	42
3.	VOCABULARY CONSTRUCTION: TERM SELECTION.....	45
3.1	LINGUISTIC AND LEXICAL PROPERTIES OF NATURAL LANGUAGE TEXTS.....	47
3.1.1	<i>The properties of noun phrases in indexing</i>	<i>49</i>
3.1.2	<i>Lexical cohesion.....</i>	<i>51</i>
3.2	THE USE OF DOCUMENT STRUCTURE TO IDENTIFY INDEX TERMS.....	52
3.3	DISTRIBUTION OF WORDS IN NATURAL LANGUAGE	54
3.3.1	<i>The empirical findings of Zipf and Luhn's notion of resolving power</i>	<i>55</i>
3.3.2	<i>The Poisson distribution.....</i>	<i>57</i>
3.3.3	<i>The 'burstiness' of word distributions.....</i>	<i>58</i>
3.4	STEP 1A: LEXICAL ANALYSIS	60
3.5	STEP 1B: THE VECTOR SPACE MODEL.....	62
3.6	STEP 1C: TERM WEIGHTING	66
3.6.1	<i>Term frequency within a single document.....</i>	<i>67</i>
3.6.2	<i>The distribution of terms across a document collection.....</i>	<i>68</i>
3.6.3	<i>Term discrimination value.....</i>	<i>69</i>
3.7	ALTERNATIVE APPROACHES FOR IDENTIFICATION OF INDEX TERMS	71
3.8	SUMMARY.....	74
4.	TERM ASSOCIATION AND VOCABULARY ORGANIZATION	77
4.1	STEP 2: TERM ASSOCIATIONS.....	78
4.1.1	<i>Co-occurrence based on grammatical relations</i>	<i>79</i>
4.1.2	<i>Co-occurrence of words within a text window</i>	<i>80</i>
4.1.3	<i>First and second order co-occurrences.....</i>	<i>80</i>
4.2	STEP 2: IDENTIFICATION OF PHRASES.....	81
4.2.1	<i>Statistical phrase identification.....</i>	<i>82</i>
4.2.2	<i>Syntactical phrase identification</i>	<i>83</i>
4.2.3	<i>Normalization and weighting of phrases.....</i>	<i>85</i>

4.3	PROXIMITY MEASURES.....	86
4.3.1	<i>The composition of proximity measures</i>	88
4.3.2	<i>Classes of proximity measures</i>	89
4.3.3	<i>Linearity of proximity measures</i>	93
4.3.4	<i>The geometric meaning of proximity measures</i>	93
4.4	STEP 3: VOCABULARY ORGANIZATION.....	96
4.4.1	<i>The characteristics of cluster analysis</i>	96
4.4.2	<i>Clustering methods and algorithms</i>	98
4.5	AUTOMATIC STATISTICAL THESAURUS CONSTRUCTION.....	101
4.6	AUTOMATIC SYNTACTIC-STATISTICAL THESAURUS CONSTRUCTION.....	106
4.7	SUMMARY.....	109
5.	BIBLIOMETRIC METHODS	113
5.1	THE CHARACTERISTICS OF BIBLIOMETRIC METHODS.....	115
5.1.1	<i>Citation analysis</i>	117
5.1.2	<i>References and citations</i>	119
5.2	CO-OCCURRENCE OF DOCUMENT ENTITIES.....	120
5.2.1	<i>Co-occurrence relations</i>	122
5.2.2	<i>The characteristics of 'entity groups'</i>	124
5.2.3	<i>'Core' and 'scatter' behaviour of 'entity groups'</i>	125
5.3	DOCUMENT CO-CITATION ANALYSIS.....	127
5.3.1	<i>Document co-citation analysis</i>	127
5.3.2	<i>Proximity measures and ordination techniques</i>	129
5.3.3	<i>Critique of co-citation analysis</i>	130
5.4	AGEING STUDIES OF LITERATURES AND DOCUMENTS.....	131
5.4.1	<i>Obsolescence of literatures</i>	132
5.4.2	<i>Synchronous and diachronous methods of obsolescence</i>	133
5.4.3	<i>Ageing patterns for individual documents</i>	134
5.5	CO-WORD ANALYSIS.....	135
5.5.1	<i>The asymmetric nature of co-word maps</i>	136
5.5.2	<i>Problems with co-word analysis</i>	137
5.6	CITATION CONTEXT ANALYSIS.....	139
5.6.1	<i>Classifying citations</i>	140
5.6.2	<i>Content analysis of citation contexts</i>	141
5.7	CITER MOTIVATIONS.....	144
5.7.1	<i>Reward or persuasion</i>	146
5.8	SUMMARY.....	150
6.	A SEMI-AUTOMATIC THESAURUS CONSTRUCTION APPROACH	153
6.1	THE EXPLORATIVE METHODOLOGY FOR SEMI-AUTOMATIC THESAURUS CONSTRUCTION.....	154
6.1.1	<i>First component: Creation of a sample text corpus</i>	156
6.1.2	<i>Second component: Creation of 'concept groups' by use of document co-citation analysis</i>	159
6.1.3	<i>Third component: Term selection and identification of 'concept symbols' by use of citation context analysis and shallow parsing</i>	162
6.1.4	<i>Fourth component: Construction of conceptual networks by use of co-word analysis</i>	165
6.1.5	<i>Fifth component: Investigation of terminological and conceptual changes by use of bibliometric methods</i>	169
6.1.6	<i>The rationale behind the selection of a specialty area for the case study</i>	171
7.	DATA ANALYSIS: CREATION OF SAMPLE TEXT CORPORA	175
7.1	CASE STUDY: PERIODONTOLOGY.....	176
7.1.1	<i>The etiology and pathogenesis of periodontal diseases</i>	176
7.1.2	<i>The treatment of periodontal diseases</i>	178
7.1.3	<i>Paradigms in periodontology</i>	178

7.2	CREATION OF ‘OVERLAPPING DOCUMENT SETS’ BY USE OF THE DATA SET ISOLATION METHOD	180
7.2.1	Cross-file searching	181
7.2.2	Isolation of overlapping documents	185
7.3	VALIDATION OF SAMPLE TEXT CORPORA BY USE OF DESCRIPTOR FREQUENCY PROFILES	191
7.3.1	Measures of corpus homogeneity – the problem of hypothesis testing	191
7.3.2	Descriptor frequency profiles	193
7.3.3	Spearman rank order correlation statistics	195
7.3.4	Identification of relative descriptor frequency differences	202
8.	DATA ANALYSIS: EXPLORATION OF COMPONENTS 2 TO 5 OF THE PROPOSED METHODOLOGY	205
8.1	SECOND COMPONENT: CREATION OF CONCEPT GROUPS (VOCABULARY ORGANIZATION)	208
8.1.1	Methodical steps of the second component	210
8.1.1.1	Document citation analysis	211
8.1.1.2	Document co-citation analysis	213
8.1.1.2.1	Mantel correlation statistic for test of monotonicity between two proximity matrices	217
8.1.1.2.2	Adjustment of the Jaccard proximity matrix	220
8.1.1.3	Document co-citation cluster analysis	221
8.1.1.3.1	Cophenetic correlation statistic for test of match between cluster result and original proximity values between co-cited pairs of references	224
8.1.1.4	Pathfinder network scaling	226
8.1.1.4.1	PFNET solution for the 2001 sample of cited references in periodontology	230
8.1.2	Summary and discussion of results	234
8.2	THIRD COMPONENT: TERM SELECTION (CITATION CONTEXT ANALYSIS)	236
8.2.1	Citation context selection	238
8.2.2	The notion of ‘consensus passages’	241
8.2.2.1	Identification of ‘consensus passages’	242
8.2.3	Extraction of noun phrases by use of shallow parsing	245
8.2.4	Identification of concept symbols	250
8.2.4.1	Summary of results from the concept symbol analysis	252
8.2.5	Evaluation and naming of concept groups	257
8.2.5.1	Summary of results from the evaluation and naming of concept groups	260
8.2.6	Selection of candidate thesaurus terms	265
8.2.6.1	Validation of selected candidate thesaurus terms	267
8.2.6.1.1	Representativeness of selected candidate thesaurus terms in the MeSH® vocabulary	268
8.2.6.1.2	Representativeness of selected candidate thesaurus terms in the Glossary of Periodontal Terms (2001)	272
8.2.7	Summary of results	275
8.3	FOURTH COMPONENT: CONCEPTUAL NETWORK (TERM ASSOCIATION)	278
8.3.1	Presentation of the noun phrases selected for the co-word analysis	281
8.3.2	Interpreting co-occurrence relationships	285
8.3.3	Creation of a conceptual network by use of co-word analysis: Special focus on primary candidate thesaurus terms	287
8.3.4	Detection of the strongest empirically thesaural relationships for the remaining noun phrases in the conceptual network	293
8.3.5	Visualizing the conceptual network as a PFNET	297
8.3.6	Summary of results	300
8.4	FIFTH COMPONENT: MONITORING OF TERMINOLOGICAL AND CONCEPTUAL CHANGES BY USE OF BIBLIOMETRIC METHODS	302
8.4.1	Concept profile analysis	305
8.4.2	Results of the retrospective analysis	311
8.4.3	Summary and discussion of results	312

9. SUMMARY AND CONCLUSIONS	315
9.1 SUMMARY OF DISSERTATION OBJECTIVES AND METHODOLOGICAL COMPONENTS	316
9.2 SUMMARY OF RESULTS	322
9.3 DISCUSSION AND RECOMMENDATIONS TO FUTURE WORK.....	327
10. REFERENCES.....	331

Appendices are found in part II of the dissertation:

Appendix 1:	Data and calculations used for validation of the sample text corpora	3
Appendix 2:	Top 64 cited references from the 2001 sample initially chosen for document co-citation analysis	5
Appendix 3:	Jaccard and cosine matrices of the top 64 cited references from the 2001 sample, initially chosen for document co-citation analysis	7
Appendix 4:	Reduced Jaccard matrix of 45 cited references	9
Appendix 5:	Top 64 cited references from the 2001 sample initially chosen for document co-citation analysis; the cited references marked in grey are the 19 highly cited references removed before the cluster analysis.	11
Appendix 6:	Clustering dendrogram of the 45 cited references	13
Appendix 7:	Cophenetic matrix	15
Appendix 8:	Network analysis; normed measures of centrality: degree, closeness, and betweenness for the 2001 sample	17
Appendix 9:	Concept symbol analysis	21
Appendix 10:	List of the 88 citing from which the sample of citation contexts is extracted	67
Appendix 11:	Histogram of the distribution of citation contexts per cited reference	69
Appendix 12:	Bibliographic description of individual cited references, their sample citation contexts, and extracted noun phrases	71
Appendix 13:	Concept group validation	73
Appendix 14:	Example of an 'enhanced document representation'	87
Appendix 15:	Candidate thesaurus terms	89
Appendix 16:	Candidate thesaurus term validation: MeSH® descriptors and Glossary terms	97
Appendix 17:	Four different matrices used in the co-word analysis	101
Appendix 18:	Network analysis; normed measures of centrality: degree, closeness, and betweenness for concept group 1 in the 2001 sample	105
Appendix 19:	Document co-citation analysis for the 1993 and 1997 samples	107

Appendix 20:	Frequently occurring noun phrases attached to cited references from the 1993, 1997, and 2001 samples concerning guided tissue regeneration	111
Appendix 21:	Consequences by applying a more elaborate search strategy	117

List of figures

FIGURE 2.1.	Example of the term ‘bibliometrics’ in the Thesaurus of ERIC descriptors.	38
FIGURE 2.2.	Example of the term ‘Guided Tissue Regeneration’ in the MeSH® tree structure.	39
FIGURE 2.3.	Example of the term ‘Guided Tissue Regeneration’ in the alphabetical display of MeSH®.	39
FIGURE 2.4.	‘Arrow graph’ by Rolling (1965), reproduced from Foskett (1981, p. 128).	41
FIGURE 3.1.	Illustration of Zipf’s findings and Luhn’s notion of resolving power.	56
FIGURE 3.2.	A simple vector space representation of three documents.	63
FIGURE 4.1.	The process of shallow parsing.	84
FIGURE 4.2.	Prototypical cluster types (Sparck Jones, 1971, p .56).	100
FIGURE 5.1.	Bibliometric methods divided according to the specific types of bibliometric analyses they support.	116
FIGURE 5.2.	Entity-relationship diagram of co-occurrence relations inspired by Morris and Yen (2004).	121
FIGURE 6.1.	‘Sample of overlapping set of citing documents’	157
FIGURE 6.2.	Direct link between primary and relative document entities in automatic thesaurus construction.	160
FIGURE 6.3.	The indirect approach to concept group creation.	161
FIGURE 6.4.	Term selection by use of citation context analysis and shallow parsing. The circles indicate how to read the successive steps illustrated in the figure. (1) Illustrates a citing paper in a research front. One reference in the bibliography is highlighted together with its corresponding citation context in the main document. (2) Illustrates the citation context surrounding the citation marker. (3) A concept symbol analysis identifies an agreed upon concept in the citation context. (4) Shows how a shallow noun phrases parser is able to extract noun phrases from the citation contexts. The dotted lines illustrate that the portfolio of noun phrases is attached to the concept symbol and to the concept group. The	164

	concept symbols and their portfolios of noun phrases define the common concept of the group and is the basis for term selection.	
FIGURE 6.5.	Conceptual network of a concept group.	166
FIGURE 7.1.	The influence of the file order on data set isolation. The file order determines where the overlapping document set is isolated.	186
FIGURE 7.2.	Ranked-log-frequency distribution for the 2001 population.	196
FIGURE 7.3.	Comparison of annual Spearman rank order correlation coefficients.	199
FIGURE 7.4.	Frequency of descriptors in populations and sample.	200
FIGURE 7.5.	Unique descriptors in population and sample.	201
FIGURE 8.1.	Methodical steps of the second component.	210
FIGURE 8.2.	Frequency distribution of citations in the 2001 sample.	212
FIGURE 8.3.	Link reduction procedure due to the triangle inequality. Two paths connect GOTFLOW_86 and NYMAN_82a. If $r = 8$ path 2 is longer than path 1, violating the triangle inequality, so it needs to be removed (W =dissimilarity weight).	229
FIGURE 8.4.	PFNET of the 2001 Jaccard matrix visualized with Pajek (Batagelj & Mrvar, 1998).	230
FIGURE 8.5.	'Artery' and 'veins' of the 2001 PFNET; node sizes indicate betweenness scores.	234
FIGURE 8.6.	Proportion of consensus passages to citation contexts distributed among document sections.	254
FIGURE 8.7.	Comparison of mean 'consensus scores' for individual document sections. TOTAL corresponds to mean for all document sections combined.	255
FIGURE 8.8.	PFNET representation with concept group names indicated. CG is an abbreviation for concept group. The dotted circle to the left indicate the semantic similarity of the concept groups in this part of the network. The dotted line to the right indicates that concept group 10 is semantically related to the concept groups in the left side of the network.	262
FIGURE 8.9.	Overlap score between selected primary candidate thesaurus terms and MeSH® descriptors.	269
FIGURE 8.10.	Overlap score between selected secondary candidate thesaurus terms and MeSH® descriptors.	270
FIGURE 8.11.	Overlap score between selected primary candidate thesaurus terms and important periodontal terms from Glossary of Periodontal Terms (2001).	273
FIGURE 8.12.	Overlap score between selected secondary candidate thesaurus terms and important periodontal terms from Glossary of Periodontal Terms (2001).	273

FIGURE 8.13.	PFNET representation of ‘first order association conceptual network’ for concept group 1. (CC = Relationship of ‘contextual relationship’; CC(H) = One-sided overlap, indicating a hierarchical relation; and DR = Definitional relationship with a threshold value over 0.80).	298
FIGURE 8.14.	Citation history and ageing pattern for NYMAN_82a, NYMAN_82b, and GOTTLow_86. ‘guided tissue regeneration’ is introduced as a MeSH® descriptor in 1992. The circles indicate the annual samples investigated in the present case study. PY+ specifies the number of years from publication of the reference to its ‘citation peak’.	307
FIGURE 8.15.	Similarity between the three concept profiles calculated by use of the Ochiai measure. The circles illustrate the concept groups and their attached noun phrases, which constitutes the concept profile. The numbers on the lines indicate the similarity score.	310

List of tables

TABLE 4.1.	Proximity measures.	90
TABLE 4.2.	Head-modifier relations (Ruge, 1999, p. 81).	108
TABLE 5.1.	Garfield’s 15 reasons for citing (Garfield, 1965, p. 189).	145
TABLE 6.1.	Matrix of the exploratory methodology.	172
TABLE 7.1.	Results from cross-file search of 2001 search string.	185
TABLE 7.2.	Removal and reverse duplicate removal of the 2001 cross-file search result.	187
TABLE 7.3.	The isolation of overlapping document sets in MEDLINE® and SCI®.	188
TABLE 7.4.	Results of the data set isolation method for four 1-year document sets - MEDLINE®.	190
TABLE 7.5.	Excerpt from the ranked results of MeSH® descriptors in the 2001 sample.	193
TABLE 7.6.	Dissolution of subject strings.	194
TABLE 7.7.	Results of the Spearman rank order correlation analyses.	198
TABLE 7.8.	Contingency table for descriptor frequencies.	202
TABLE 7.9.	Relative frequency similarities and differences between descriptors in the population and sample.	204
TABLE 8.1.	2×2 contingency table from Anderberg (1973, p. 83).	215
TABLE 8.2.	Binary formulas of the Jaccard and cosine proximity measures.	215

TABLE 8.3.	Mantel test of monotonicity: Jaccard versus cosine similarity matrix.	219
TABLE 8.4.	Cluster result of 2001 sample containing 45 cited references. Median ages of the publications in the clusters are indicated in the right column.	223
TABLE 8.5.	Cophenetic correlation test.	226
TABLE 8.6.	Central and connective nodes deduced by visual inspection of the 2001 periodontology intellectual base PFNET.	231
TABLE 8.7.	Top ranks on three normed centrality measures (degree, closeness, and betweenness) for the 2001 periodontology PFNET.	232
TABLE 8.8.	Citation contexts for LOE_63 and the calculation of the normed 'consensus score'.	244
TABLE 8.9.	List of content words that appear in two or more citation contexts to LOE_63.	244
TABLE 8.10.	Example of noun phrase parsing of citation contexts by use of the Connexor parser.	247
TABLE 8.11.	Result of the concept symbol analysis for the 45 highly cited references in the 2001 sample of overlapping documents in periodontology.	253
TABLE 8.12.	Name of concept groups and quantitative indication of semantic coherence. The semantic coherence value is the average similarity scores of individual portfolios of concept symbols within a concept group compared to the 'aggregate portfolio' of noun phrases for the entire group.	260
TABLE 8.13.	Chi-square test of proportional difference in relevant index terms between primary and secondary candidate thesaurus terms – based on match with MeSH® descriptors.	271
TABLE 8.14.	Chi-square test of proportional difference in important domain terms between primary and secondary candidate thesaurus terms – based on match with Glossary terms.	274
TABLE 8.15.	Subsumption of head noun variants. The parsed phrases are indicated to the left and their frequency counts in parenthesis. The result of the subsumption procedure is the factored phrases to the right.	282
TABLE 8.16.	The noun phrases selected for the conceptual network of concept group 1. Primary candidate thesaurus terms are indicated in grey.	283
TABLE 8.17.	Equivalence proximity matrix of first order association between the selected noun phrases in concept group 1. Only the lower-left half of the equivalence proximity matrix is shown since it is symmetric.	289
TABLE 8.18.	Second order co-occurrence analysis between binary 'association profiles' of enamel matrix protein and enamel matrix derivative. The two noun phrases are represented as binary vectors; their similarity score is indicated to the right.	290

TABLE 8.19.	Empirically derived first and second order relationships by use of co-word analyses. The matrix is similar to the equivalence matrix in TABLE 8.17., however the upper-right half in the present symmetric matrix indicates the strongest empirically derived thesaural relationships by a code.	294
TABLE 8.20.	Citation counts and index scores for the three cited references from the 2001 concept group concerning guided tissue regeneration. Citation counts are the actual number of citations in the samples of overlapping documents. The index score is the relative citation count for annual samples. The index score makes it possible to compare citation counts across the annual samples.	306
TABLE 8.21.	The 1993 and 1997 concept groups denoting ‘guided tissue regeneration’ and their cited references. The three cited references constituting the 2001 concept group is marked in bold.	308

1. Introduction

A thesaurus is a controlled vocabulary and is used for indexing and/or retrieval of documents¹ (Aitchison, Gilchrist & Bawden, 2000). In other words, a thesaurus is a knowledge organization system. Within library and information science, knowledge organization denotes classification, indexing, and cataloguing, applied to storage, access, and retrieval of documents in information retrieval (IR) systems (Anderson & Pérez-Carballo, 2001a). According to Anderson and Pérez-Carballo (2001b) all classification and indexing is derived from classing or clustering of objects, terms as well as documents, based on similarities of characteristics. Anderson and Pérez-Carballo (2001a) further point out that the term ‘clustering’ implies an automatic process that generates *a posteriori* clusters, in contrast to the term ‘classing’ that usually implies human judgement and requires pre-existing classes. Whether the focus is on actual indexing, classification, or the construction of knowledge organizing systems, the main issue is the same, that is, the description of document content in order to group items of similar characteristics (Anderson, 1997). Basically, there are two fundamental approaches to the description of document content and consequently to construction and maintenance of knowledge organization systems: 1) manual intellectual analysis, and 2) automatic algorithmic analysis (Anderson, 1997; Anderson & Pérez-Carballo, 2001a; 2001b; Lancaster, 2003). These two approaches are increasingly combined to benefit from the strengths of each approach, and to counterbalance their weaknesses as well (Anderson, 1997). Manual intellectual construction and maintenance of knowledge organization systems is recommended when dealing with languages due to the dynamic and complex nature of language (e.g., Blair, 1990; Aitchison, Gilchrist & Bawden, 2000). But, the manual intellectual approach is a resource demanding and costly process, which motivates to do research on less resource demanding construction and maintenance methods (e.g., Soergel, 1974; Aitchison, Gilchrist & Bawden, 2000; Anderson & Pérez-Carballo, 2001a; 2001b).

There are several ways to characterize knowledge organization systems. Hodge (2000) uses the complexity of the systems as the primary criterion for characterization. This results in three major groupings: term lists, which emphasize list of terms often with definitions, including for instance authority files and glossaries; classifications

¹ The term *documents* is used to denote all kinds of information objects or texts in a semiotic sense, i.e., organized sets of symbols chosen to represent a message, albeit the focus of dissertation concerns scientific text documents.

and categories, which emphasize the creation of subject sets, including for instance subject headings, classification schemes, and taxonomies; and relationship lists, which emphasize the connections between terms and concepts, including thesauri, semantic networks, and ontologies. Thesauri are the most complex systems as they ideally express different subject sets, conceptual structures, and relational connections between terms and concepts within and between these subject sets. Consequently, relationship lists are likewise the most demanding knowledge organizational systems to construct and maintain.

Motivated by the complementarity of manual and automatic indexing, the overall aim of the research project presented in this dissertation is to develop and verify a semi-automatic thesaurus construction approach. This means that the emphasis of the dissertation is on the development of a useful methodology and not on performance evaluation.

1.1 Background

Thesaurus construction requires the collection of a set of candidate terms (words and phrases) within a given subject area, terminological and semantic control of the identified terms, and finally creation of term classes and relationships between terms through relational structuring (Soergel, 1974; Aitchison Gilchrist, & Bawden, 2000). The classes cover restricted topics of specific scope, and collectively they cover the complete subject area in question. Thus, thesauri are fundamentally linguistic and conceptual in nature (Miller, 1997, p. 484). Structural, semantic and terminological problems are ever present, and manual intellectual construction work is necessary when dealing with these problems (Aitchison, Gilchrist, & Bawden 2000). This is usually done by a combination of document analysis and a group of experts who review the subject matter, suggest potential terms, and propose reasonable class arrangements (Lancaster, 2003). One major disadvantage inherent to the use of any thesaurus is the necessity to maintain it (Blair, 1990; Aitchison, Gilchrist & Bawden, 2000). New thesaurus classes of interest emerge and the thesaurus needs to accommodate for collection growth. Especially, in some disciplines where the vocabulary changes rapidly maintenance of the thesaurus is important. The traditional approach taken to meet the requirements of thesaurus construction and maintenance in a less resource demanding way is the automatic algorithmic approach (e.g., Salton & McGill, 1983).

The automatic approach to indexing and thesaurus construction automatically derives subject matter words from the author's text (Salton & McGill, 1983). Two approaches to automatic thesaurus construction exist: 1) a statistical approach based on co-occurrence analysis of terms in the text corpora, and 2) a linguistic approach, which combines syntactical parsing of natural language text with co-occurrence analysis of the parsed syntactical structures (Salton & McGill, 1983; Grefenstette, 1994; Kowalski & Maybury, 2000). The assumption behind the statistical approaches is that contextually related co-occurring terms (i.e., often appearing in the same sentence, paragraph, grammatical context, or document) are semantically related and hence should be grouped together in the same thesaurus cluster (Lancaster, 2003, p. 292). Automatic thesaurus construction, however, typically relies on simple and flawed term distribution assumptions (Kowalski & Maybury, 2000; Moens, 2000). Little attention is given to the role of schematic, discourse and thematic structures in documents, and their influence on term distributions (Katz, 1996; Bookstein, Klein & Raita, 1998).

Statistical based thesaurus construction usually detects terms and possible associative relationships (e.g., Kowalski & Maybury, 2000). But, several intrinsic problems exist in relation to automatic thesaurus construction. For example, it is difficult to identify synonyms by use of simple term co-occurrence analysis within in the same documents (e.g., Peat & Willett, 1991; Schütze & Pedersen, 1997; Lancaster, 2003). To detect the specific semantic nature of these terms, and their relationships, is usually beyond their scope. Recent research indicates that a 'local'² approach, which pays more attention to the schematic, discourse and thematic structures within documents may be valuable for identification of the semantic nature of and between terms (Moens, 2000; Bookstein et al., 2003). In addition, term association research, which uses natural language processing (NLP) techniques, shows that head-modifier relationships of noun phrases are useful for identification of term relationships (e.g., Ruge, 1992; Grefenstette, 1994; Godby, 2002). Even so, automatic thesaurus construction methods cannot function alone, if an elaborate structure and semantic term validity is desired (Lancaster, 2003).

The present work uses bibliometric methods, such as co-citation analysis, citation context analysis, and co-word analysis, as well as NLP techniques, such as syntactical parsing. The motivation behind the use of bibliometric methods stems from the notion that cited documents are analogous to 'subject headings' or 'concept symbols', and most importantly, that the relationships between them can be uncovered without depending on natural language terminology (Garfield, 1974; 1979a; Small, 1978;

² We define 'local' as smaller more focused text corpora, where document structure has an effect. Conversely, 'global' is large text corpora, equivalent to the 'bag-of-words' allegory in information retrieval where documents' structure is not utilized.

Rees-Potter, 1989). Co-citation analysis is used to identify and structure key documents in specialty areas. Citation context analysis is used to locate citation contexts within the local structure of the key documents. The citation contexts are the target for term extraction. Syntactical parsing of noun phrases is employed to extract coherent candidate terms and phrases, and co-word analysis is used to identify relational structures between the extracted terms. Neither of these methods are applied in the usually vigorous automatic manner. They are explored and utilized in conjunction with manual thesaurus construction methods. Thus, we introduce a semi-automatic approach to thesaurus construction, based on bibliometric methods, as a supplement to manual intellectual approaches.

The dissertation builds on former research in this area. Especially the research by Rees-Potter (1987; 1989) on semi-automatic thesaurus maintenance is a good example of the combination of manual intellectual and automatic algorithmic approaches. Furthermore, like us, she builds upon bibliometric methods. The objective of her study was twofold, that is, to identify conceptual changes in two disciplines over time, and to investigate bibliometric methods' ability to identify candidate thesaurus terms (Rees-Potter, 1989). The results indicate that highly cited and co-cited documents act as 'concept symbols', hence, verifying former research by (Small, 1978). Thus, citing documents' citation context can be investigated as sources of candidate thesaurus terms. However, the investigated citation contexts in the study of Rees-Potter (1987) primarily come from monographs due to domain specific conditions (Cronin, Snyder & Atkins, 1997). In addition, none of the employed methods by Rees-Potter (1987; 1989) accomplishes to identify conceptual changes over time. The initial intention of Rees-Potter was to implement her findings in a full text system for semi-automatic thesaurus construction and maintenance, but this has not been achieved. Nevertheless, her research indicates the value of bibliometric methods for the selection of candidate thesaurus terms, and as such, as an alternative to the traditional term co-occurrence methods. Some of the conditions and problems verified in her study are addressed in the present dissertation.

1.2 Objectives of the dissertation

Bibliometric methods may point to common topical characteristics of documents and their authors, and may be used to uncover, otherwise hidden knowledge structures of a discipline or subject area and its users (White & McCain, 1989; Borgman, 1990). The underlying hypothesis of the research project is that bibliometric methods can be used

as a supplement to the already established methods of thesaurus construction, because bibliometric methods may uncover conditions, patterns and relationships between documents and their concepts (e.g., Small, 1978; Rees-Potter, 1989). Based on the hypothesis, the overall objective of the research project is to reintroduce, and further extend, the theoretical and methodological aspects of bibliometric methods to the research area of knowledge organization for the purpose of thesaurus construction and maintenance. A non-exhaustive exploratory methodology based on bibliometric methods and supplemented with multivariate statistical techniques, network analysis, NLP, and term frequency analyses are presented as a combined semi-automatic approach to thesaurus construction and maintenance. The semi-automatic approach is to be seen as a supplement to the resource demanding approach of manual intellectual thesaurus construction. The supplement is to be understood as a bibliometrically created and document-oriented basis that may help manual construction. The supplement consists of candidate terms, concepts and concept groups. In addition, the supplement will indicate conceptual relations and structures, as well as possible conceptual changes over time. We do not attempt the ‘complete’ construction of a subject specific thesaurus. As pointed out above, we see the approach and its methods as a supplement to other thesaurus construction approaches. In our opinion, it is beneficial to combine methods because different construction methods emphasise different aspects. They have different strengths and weaknesses, and therefore may yield different results. Hence, we agree with Lykke Nielsen’s (2002, p. 35) ‘holistic approach’ to thesaurus construction, when she emphasizes the usefulness of different methods for different purposes.

We re-introduce and modify Rees-Potter’s former research approach (Rees-Potter, 1987; 1989). Her research is almost 20 years old and much has happened methodically since then, that is, in relation to bibliometrics and its related analytical techniques. The advancement in computer technology has been a prime mover in this respect. The important concept of information visualization has gained much attention recently within library and information science (e.g., Chen, 1999; 2003; Lin, White & Buzydlowski, 2003). It is accepted that information visualization can help a diverse community to gain overviews of patterns and trends, and to discover hidden semantic structures embedded in data matrices (Chen, 2003). Obviously, this is interesting for thesaurus construction approaches based on relational data matrices. In relation to the research project, information visualization relies on multivariate statistical techniques, network analyses and network algorithms such as Pathfinder network scaling (Schvaneveldt, 1990). Motivated by such methodical advances we expand and update the theoretical and methodological aspects of former bibliometric approaches to the

research area of knowledge organization – specifically for the purpose of thesaurus construction and maintenance. Further, our objective is to verify the appropriateness of visualization techniques when used in combination with bibliometric methods for the purpose of thesaurus construction and maintenance.

1.3 Research assumptions

The overall assumption behind the research project is that bibliometric methods can be used advantageously to cluster semantically related highly cited documents. The value is believed to come from the fact that documents are clustered from the distribution patterns of co-citations, and not, as traditionally done, by use of ‘global’ distribution patterns co-occurring terms. In this way related documents are collocated in smaller subject clusters, essentially as a citation network as described by Garfield and Small (1985). This is valuable in relation to thesaurus construction because subject clusters based on co-citations may contain term relationships that are not found in clusters created from traditional co-term analysis. Obviously, clusters based on citations are independent of language and changing terminology (Rees-Potter, 1989; Leydesdorff, 1997). We disperse with the traditional crude ‘global’ text corpora approach to automatic thesaurus construction (e.g., Belew, 2000; Kowalski & Maybury, 2000). In line with Anderson and Pérez-Carballo (2001b), and Blair and Kimbrough (2002) we initially structure the subject area, and cluster its text corpora into smaller, more manageable, ‘local segments’ of semantically related ‘exemplary documents’ in a citation network. This ‘local’ structure then functions as the basis for various statistical analyses, including co-word analyses, which in many ways resemble co-term analysis.

We assume that documents and their bibliographies represent concepts of individuals and groups (e.g., Garfield, 1970; Small, 1978). Semantic relationships exist between citing and cited documents (e.g., Cronin, 1994; White & Wang, 1997). Highly cited documents and highly co-cited pairs of documents reflect key concepts in a specialty or discipline (e.g., Small, 1978; 1980; Cozzens, 1985a). Likewise, the terminology used in the citing documents reflects the concepts represented by the highly cited and co-cited documents (e.g., Small, 1986). Consequently, highly cited documents act as ‘concept symbols’. To some extent, if a majority of people citing these documents use the same terms and phrases to denote the overall concepts, then there exist some kind of agreement or consensus on what concepts these document

denote and how they are expressed (Small, 1978). The usage agreement makes the document a ‘concept symbol’.

Citation context analysis is the principal method applied to identify potential ‘concept symbols’. However, citation context analysis is a labour intensive work of manually scanning full text documents to identify citation contexts (the sentences surrounding a reference) in order to select key concepts that capture major aspects of the cited documents (e.g., Small, 1978; 1986; Rees-Potter, 1987; 1989). To reduce the labour intensive work in connection with citation context analysis, we want to investigate whether a natural language parsing tool can yield comparable results in selecting key concepts from citation contexts. Therefore are citing documents’ citation contexts in the present work defined as *units of texts* and target for semi-automatic content analyses based on natural language parsing and co-occurrence analysis. The assumption is that isolation of semantic related documents in a ‘local structure, or citation network, and the selection of highly specific areas of the document structure as target for extraction (indicated by consensus usage), enables selection of candidate terms and phrases. Further, this makes it possible to generate ‘local’ conceptual networks from the extracted candidate terms that indicate various relationships among concepts and terms.

Rees-Potter (1987; 1989) verified that within two disciplines of the social sciences (i.e., economics and sociology) citation contexts can be used to identify candidate thesaurus terms. However, a high proportion of monographs appear to be a characteristic of these disciplines (Rees-Potter, 1989), and “[a]ny difference between citation context and citation terminological analysis of monographs and journal articles may be a point for further research” (Rees-Potter, 1987, pp. 29-30). On this basis, we want to verify if identification of candidate thesaurus terms can be satisfactorily accomplished in a scientific domain where scholarly knowledge is primarily mediated through journal articles, expecting that this medium influences the terminology used in the citation contexts. In addition, “[c]onsidering that the purpose of this project [Rees-Potter’s] is to use documents as concept symbols to produce appropriate terminology for specialty areas, perhaps the use of documents that would be concept symbols for narrower specialty areas would be more useful and cause fewer problems in specialty identification” (Rees-Potter, 1987, p. 77). Hence, we carry out our investigation on a lower disciplinary level, that is, *periodontology*, which is a specialty area within the discipline of dentistry, with dentistry belonging to the category of life sciences. Because we do not work with a ‘global’ text corpus it is necessary to create a reliable text corpus sample that reflects the subject aspects of the speciality area to be studied (i.e., the population). For this reason, the data set isolation method (Ingwersen &

Christensen, 1997), in combination with sampling validation methods used in corpus linguistics, is used to construct a manageable sample of documents. The data set isolation method enables the generation of an overlapping text corpus of document representations indexed in both MEDLINE® and Science Citation Index®(SCI®), the specialty area of periodontology is hereby covered from two different perspectives – Medical Subject Headings® (MeSH®) and citations.

The idea is that the transformation from a 'concept symbol' citation network to a conceptual network enables structural and relational investigations of links in such a network by use of visualization techniques. A concept symbol has a portfolio of terms attached to it. The terms in the portfolio show significant distributional behaviour, such as high frequency of occurrence, in the citing contexts of their attached concept symbol. The co-citation clusters contain semantically related documents that act as 'concept symbols'. The portfolios from each of the individual 'concept symbols' in the clusters are joined in a matrix³. This enables the investigation into thesaural relationships, such as equivalence, hierarchical and associative relationships, from the co-occurrence patterns inherent in the matrix (e.g., Soergel, 1974; Callon et al., 1983; Roberts, 1997).

It has been demonstrated that citation context analysis reflects concepts in a specialty area, as well as the history of concepts in a specialty area (e.g., Small & Greenlee, 1980). Therefore it seems reasonable to assume that the proposed approach, that is, to employ bibliometric methods to identify candidate terms, likewise can be used as a basis to uncover terminological changes within the given subject specialty area of periodontology over time. The present study thus uses the bibliometric methods of citation and co-citation context analysis and further extends this research to an application in the area of terminological analysis. A retrospective word profile analysis of several volumes of identical 'concept symbols', and their attached citing context *portfolios*, is applied to detect possible terminological change over time and terminological variation among individual 'concept symbols' and groups of 'concept symbols' (Braam, Moed & Van Raan, 1991a; 1991b). Terminological change may to some degree reflect conceptual change (Rees-Potter, 1989; Temmerman, 2000). An analysis of terminological change reflects, to some degree, how concepts are changing in a subject specialty area (Rees-Potter, 1987; 1989).

³ It should be noted that semantically related documents may express the same 'concept symbol', but it is unlikely that their portfolios are identical.

1.4 Research questions

The overall aim of the research project presented in this dissertation is to verify the applicability of bibliometric methods as a supplement to intellectual manual based construction and maintenance of thesauri. The dissertation builds on former research in this area (Rees-Potter, 1987; 1989). We reintroduce, modify, and expand former research approaches in that we evolve an exploratory methodology, introduced as a semi-automatic approach to thesaurus construction and maintenance. We seek to verify the overall aim by posing three research questions that each sets out to explore vital characteristics and methodical aspects required in thesaurus construction and maintenance.

The basic characteristic of thesaurus construction is the identification of candidate thesaurus terms. This leads to the first research question:

1. Is it possible, by use of bibliometric methods, to detect candidate thesaurus terms in a specialty area within the life sciences given its disciplinary, publication, and terminological conditions?

The first question will clarify whether candidate thesaurus terms can be identified at a lower domain level than hitherto done, and in a scientific domain where scholarly knowledge is primarily mediated through journal articles. The focus of Rees-Potter's (1987; 1989) research was the disciplinary level of economics and sociology and most of the identified concept symbols were books. Our focus is a specialty area (periodontology) within the discipline of dentistry. Furthermore, Rees-Potter (1987; 1989) suggests that books may denote and behave differently from journal articles in the framework of citation context analysis. For this reason, our emphasis is on a scientific domain where journal articles are the principal written medium for knowledge diffusion. The selected candidate thesaurus terms are evaluated by use of an inferential statistical validation procedure. The procedure consists in matching the selected terms with possible corresponding terms in two acknowledged authority lists, which are the MeSH® vocabulary and *Glossary of Periodontal Terms* (2001).

A unique and important characteristic of a thesaurus is its ability to depict three types of relationships – equivalence, hierarchical and associative relationships. This leads to the second research question:

2. To what extent can the applied bibliometric methods, outlined in the exploratory methodology of the semi-automatic approach, help identify equivalence,

hierarchical and associative relationships between individual terms and concepts within the concept groups?

The second question investigates to what degree the applied methodology is able to help identify and uncover synonyms, related terms, and structure between individual terms and concepts. This will clarify whether it is possible to identify, and to what extent, thesaural relationships by use of the tools applied in the suggested methodology. A combination of, first, co-citation analysis, followed by co-word analyses, and finally visualization techniques, are applied to the 'concept symbols' and their *portfolio* of noun phrases in order to make a codification of relations in the conceptual network. In order to evaluate the results, we analyze the statistically derived relations between noun phrases and concepts by comparing them with similar relations in the MeSH® vocabulary, a domain specific classification scheme, and characterizations of the involved phrases in textbooks of periodontology, including the *Glossary of Periodontal Terms* (2001). Thus, it is possible to interpret and verify the types of relationships.

Thesauri are time consuming and can be difficult to maintain and update. Hence, there is a need to develop a methodology by which thesauri can be more easily updated to account for change and variation in the terminology of a subject area. This leads to the third and last question:

3. To what extent can the applied bibliometric methods, outlined in the exploratory methodology of the semi-automatic approach, monitor and identify terminological and conceptual changes in a given subject specialty area over a given time period?

The third question investigates the dynamical aspects of the proposed methodology. To be able to evaluate how well our methodology reflects dynamical changes in language, we have incorporated quantitative research methods used to map dynamical aspects of science and technology (Braam, Moed & Van Raan, 1991a; 1991b). We assume that terminological changes within a subject area, to some degree also reflect conceptual change within that area (Rees-Potter, 1989; Temmerman, 2000). Consequently, a retrospective analysis examines a number of selected concept profiles (i.e. the profiles represent the most persistent noun phrases for a semantically related group of 'concept symbols') for different time periods to detect possible terminological changes within the profiles over time. Moreover, a bibliometric study of the ageing patterns of the related 'concept symbols' is performed to substantiate the results from the quantitative concept profile analysis. Even though the evaluation is retrospective, it

may indicate whether changes in concept profiles and the citation history of highly cited documents can be used as an indication of conceptual change in a subject area?

Together, the three quantitative analyses give answers to the three research questions, which will indicate the overall value of the applied exploratory methodology and verify the applicability of using bibliometric methods as a *supplement* to intellectual manual based construction and maintenance of thesauri. However, the research questions may individually point to prosperous methodical advances independent of each other.

1.5 Organization of the dissertation

The doctoral dissertation presents a methodology for semi-automatic construction and maintenance of thesauri, essentially building on former research within manual intellectual thesaurus construction, automatic text analysis, and bibliometrics. The methodology is delineated through illustrative empirical cases. In effect, this brings about a natural organization of the dissertation in two parts, a theoretical part, that comprise the Chapters 2 to 6, and an empirical part that comprises the Chapters 7 and 8. Chapter 9 comprise summaries and conclusions. The organization of the dissertation is provided below with a brief representative description of each chapter. Introductory reviews are by no means exhaustive, they serve as background information, in order to place the proposed methodology in a suitable context.

Chapter 2 provides an introduction to the purpose and construction of the traditional manual built thesaurus. The focus is on IR, as the major area of thesaural use. The chapter also describes the complementarity of natural language and controlled vocabularies. The chapter closes with an introduction to the main steps of manual intellectual thesaurus construction. The latter is not an exhaustive description; instead, it serves as an introduction, where the focus is on aspects of special interest to the dissertation.

The process of automatic thesaurus construction is basically an extension of automatic indexing. Consequently, we divide the treatment of automatic thesaurus construction between term selection methods (Chapter 3) and term association methods (Chapter 4). Chapter 3 gives an introduction to methods of automatic indexing and the assumptions behind these methods. The focus is on the different distributional characteristics of natural language terms, and how this is used for selecting and weighting of index terms. Furthermore, the principal representational

model used in automatic indexing, the vector space model, is thoroughly outlined. The chapter closes with a look upon alternative principles for identification and selection of index terms; here the focus is on the schematic, discourse and thematic structures of documents.

Chapter 4 introduces the notion of phrases as index terms, term associations, and how they are applied for automatic thesaurus construction. The chapter outlines the two principal approaches used for automatic identification of phrases and term associations, which is statistical term co-occurrence analysis, and syntactical parsing of mainly noun phrases. Attention is given to methods both from within IR and within computational linguistics. Furthermore, the chapter presents a range of different attempts at automatic thesaurus construction in order to illustrate the methodology, its strength and weaknesses.

Chapter 5 presents the main bibliometric methods applied in the dissertation. This is co-citation analysis, citation context analysis, ageing studies and co-word analysis. The major part of the chapter introduces and discusses these methods. Special emphasis is given to the notion of 'visualization of literatures'. This is the subject area of co-citation and co-word analysis, and how these methods can be used to structure the subject literature of a specialty area. Further, we delineate the main assumptions and controversies behind the application of bibliometric methods in general.

Chapter 6 serves as a link between the theoretical Chapters 2 to 5 and the succeeding empirical Chapters 7 and 8. Chapter 6 presents the proposed semi-automatic thesaurus construction methodology. The rationale and assumptions behind the components of the methodology, as well as, its application are given based on the preceding literature review of founding research in the foregoing Chapters 2 to 5.

Chapter 7 serves two functions: 1) it describes the specialty area of periodontology that is used for the empirical case, and 2) it presents and validates the data sampling method. The data set isolation method is used to generate the document sample of *periodontology* papers from MEDLINE® and SCI®, thus establishing the text corpus used for the semi-automatic thesaurus construction methodology. The document sample is statistically validated by use of parallel corpus sampling validation measures, used in corpus linguistics.

Chapter 8 describes the application and validation of the proposed methodology's individual components in a data analysis of the generated document sample. Essentially, the three research questions are pursued, that is, the ability of the proposed methodology to: 1) identify candidate thesaurus terms, 2) to help identify equivalence, hierarchical and associative relationships, 3) and its ability to monitor and identify terminological and conceptual changes within the given specialty area of

periodontology over a given time period. Positive results from the quantitative evaluations indicate the usefulness of bibliometric methods for semi-automatic thesaurus construction.

Chapter 9 summarizes the findings of the dissertation, and draws the conclusions in relation to the three research questions. In addition, the significance and the implications of the results of both the theoretical and empirical findings are discussed, and recommendations for future work are outlined.

2 Thesaurus construction

As introduced in Chapter 1, the present work concerns the development and exploration of a semi-automatic approach to construction and maintenance of thesauri for textual documents. We will not discuss indexing of other media types. Within library and information science, knowledge organization is concerned with the study of cataloguing, classification, and indexing applied to documents in IR systems (Anderson & Perez Carballo, 2001a; 2001b). From the perspective of library and information science, a thesaurus is an indexing language (a knowledge organizing system). Indexing languages serve as ‘language tools’ and subject aids for indexing and retrieval of documents, as well as devices for understanding of conceptual structures of specific subject areas (Soergel, 1974; Aitchison, Gilchrist & Bawden, 2001). As mentioned in Chapter 1, two overall approaches exist to the process of indexing, and to the construction of indexing languages, i.e. manual and automatic approaches (Anderson & Perez Carballo, 2001a; 2001b). The two approaches are essentially complimentary (Anderson & Perez Carballo, 2001a; 2001b; Lancaster, 2003). The purpose of this chapter is to show that a comprehensive thesaurus design demands a high degree of manual intellectual construction work, and that manual construction work is a time consuming and costly process (e.g., Aitchison, Gilchrist & Bawden, 2000; Lancaster, 2003).

The aim of the present chapter is therefore to outline the setting of thesauri and the main characteristics of manual thesaurus construction. In order to comprehend the complexities inherent in the construction of thesauri, we find it important to describe the setting in which thesauri are used, as well as the special characteristics that pertain to them. Section 2.1 outlines the functions of an IR system, as the primary setting for a thesaurus. Manually constructed thesauri belong to the category of controlled indexing languages. Section 2.2 outlines the most important differences between natural language and controlled indexing languages. The emphasis is on the need for both languages in the context of indexing. Section 2.3 defines the traditional thesaurus, and outlines its primary and secondary functions. Section 2.3 is divided into sub-sections that outline the main aspects of manual thesaurus construction. The chapter is not an exhaustive review of manual thesaurus construction. However, it serves as an introduction to common concepts, definitions, characteristics, and facts in relation to thesauri as perceived and applied in the dissertation.

2.1 Information retrieval systems

From the perspective of computer science, IR deals with the processes of representation, storage, retrieval, and presentation of information (e.g. van Rijsbergen, 1979; Belew, 2000; Kowlaski & Maybury, 2000). In this context, the term information is used in a broad sense and denotes the traditional 'meaning of text' stored in documents⁴ (Kowlaski & Maybury, 2000). IR systems represent implementations of these processes. The goal of IR research is to develop systems that retrieve information that is relevant to an information need (e.g. van Rijsbergen, 1979; Belew, 2000; Kowlaski & Maybury, 2000). Consequently, IR research is concerned with the development and refinement of efficient models and algorithms for representation, storage, retrieval, and presentation of information (e.g., Frakes & Baeza-Yates, 1992; Grossman & Frieder, 1998). Although the term IR does not refer to information representation, storage, or presentation, these ideas are implicit in any comprehensive definition of IR (e.g. van Rijsbergen, 1979; Belew, 2000; Kowlaski & Maybury, 2000). Also implicit in today's conception of IR is the idea of user interaction (Ingwersen, 1992; 1996). To date, most IR research have focused on the representation, storage, retrieval, and presentation of text-based content due to the ease with which textual symbols can be represented and manipulated within computers (e.g., Frakes & Baeza-Yates, 1992; Grossman & Frieder, 1998; Kowalski & Maybury, 2000). However, in recent years, research on multimedia IR has grown rapidly (e.g. Downie, 2003; Smeaton, 2004). It should be noted, that IR systems need not be computer based, but computers are extremely useful in processing large quantities of data and information, and at speeding up this process (e.g., Grossman & Frieder, 1998).

The basic operations of IR systems are: 1) document analysis, which is the representation and storage of documents in the system, and 2) matching, which is the retrieval and presentation of documents relevant to the information need.

Each IR system provides access to a set of documents. A document represents an item of information in the IR system. Documents may range in length from short passages of text to complete monographic works (Lancaster, 2003). The document

⁴ We do not elaborate on the concept of *information*; however, Buckland (1991) identified information by its context of use as a process, a thing, or knowledge. Most relevant to this work is Buckland's definition of information-as-thing, represented by something tangible such as a document or a text that may impart information.

might consist of the ‘complete content’ of the information item, in which case it is considered a full text document representation (Kowlaski & Maybury, 2000). Alternatively, if the full text of the document is not available, as is frequently the case in bibliographic information retrieval systems, a surrogate document representation is stored instead (Kowlaski & Maybury, 2000). A document representation is a formal description of a document for the purpose of retrieval, identification, and relevance judgement (e.g., Soergel, 1985; Lancaster, 2003). The document representation usually consists of a bibliographical description that registers formal data concerning a document, and a subject description, encapsulating aspects of the document’s content (e.g., Soergel, 1985; Lancaster, 2003). From a library and information science perspective, the creation of a document representation contains two tasks: 1) cataloguing (descriptive bibliographic analysis), and 2) indexing (subject analysis) (Soergel, 1985). Eventually, the processed document representation constitutes the basis for IR system indexes. The processes of automatic indexing and thesaurus construction are treated in the Chapters 3 and 4.

IR systems give access to documents by matching user search queries to indexed document representation (Kowlaski & Maybury, 2000, p. 28). A query represents the user’s information need. A query consists of one or several terms and possible search operators that define the scope of the documents to be retrieved. In IR research, a term is an unbroken string of alphanumeric characters that serve as the basic unit for searching. During query processing, query terms are matched to index terms, which define the searchable terms within the database.

There exist three main categories of mathematical matching models in IR, which are set theoretic, algebraic (geometric), and probabilistic models (Baeza-Yates & Ribeiro-Neto, 1999, p. 21). Two principal matching systems are related to these models. Exact match systems return documents that precisely satisfy a structured query expression, of which the best-known type is the set theoretic Boolean queries (Belew, 2000; Kowlaski & Maybury, 2000). For large and heterogeneous document collections, the result sets of exact match systems usually are either empty, or huge and unwieldy, so recent work has concentrated on best match retrieval systems that rank documents according to their perceived or estimated relevance to the query (Belew, 2000; Kowlaski & Maybury, 2000). The classic best match retrieval models include algebraic models, such as the vector space model (e.g., Salton & McGill, 1983) and latent semantic indexing (Deerwester et., 1990), and probabilistic models such as the *BM25* (e.g., Robertson et al., 1994). In chapter 3 we outline the vector space model due to its widespread use in automatic indexing and automatic thesaurus construction. In addition, an alternative search strategy to document-query matching is that of

browsing (Marchionini, 1995). IR systems can adopt one or several of the models to provide access to documents (Kowlaski & Maybury, 2000).

IR systems may implement natural language (free text) searching, or may rely on a controlled vocabulary (Soergel, 1985; Lancaster, 2003). A natural language search does not limit the searchable terms available to a set of predefined terms. Conversely, a controlled vocabulary, such as thesaurus descriptors, limits the search vocabulary to a set of preferred standardized terms that have been applied to index the individual documents in the system (Soergel, 1985; Lancaster, 2003). Controlled vocabularies rely on developed knowledge organizing systems such as thesauri to ensure standardized usage of vocabularies. The next section introduces manual indexing, and the characteristics and functions of natural language versus controlled vocabularies.

2.2 Natural versus controlled languages

The main purpose of manual indexing is to construct subject representations of documents in a form suitable for inclusion in IR systems (not necessarily in electronic form) to support retrieval (Soergel, 1985; Lancaster, 2003). In order to establish the subject matter of a document a subject analysis is required. The manual approach consists of a conceptual analysis in which the subject matter of the document is established and subsequently 'translated' into controlled index terms from an appropriate controlled vocabulary (Lancaster, 2003).

Paradoxically, the basic problem in IR is language (e.g., Holm & Rasmussen, 1961; Furnas et al., 1987; Blair, 1990; 2003). Language consists of words, and words derive meaning from their surrounding context (Blair, 1990). Many words have several meanings, and many concepts can be named individually by a set of different synonyms. As a result, a vocabulary problem exists in IR, that is, people use different words to describe the same concepts (Furnas et al., 1987). According to Holm and Rasmussen (1961, p. 184), natural language problems can be separated into three groups:

- Individuals consider objects, ideas, facts or images from different *viewpoints*;
- Almost every concept implies the use of broader or narrower concepts, this creates generic problems, as a given topic may be described and searched from several hierarchical levels;

- Semantic problems that concern the relationship between words and their meanings, this includes synonymous and quasi-synonymous problems.

The notion of controlled vocabulary exists in indexing to be able to limit some of the language problems inherent in IR. A controlled vocabulary is a limited set of chosen standardized descriptors, subjected to a terminological and semantic control, which is used to secure redundant indexing and redundant searching.

The opposite to manual subject analysis of documents is automatic indexing. Automatic indexing identifies, compares, and extracts uncontrolled index terms from the natural language text in documents (e.g., Kowlaski & Maybury, 2000). This literally means no terminological or semantic control of the indexed terms, and, consequently, uncontrolled index terms are subjected to the natural language problems outlined above (Lancaster, 2003). Chapter 3 outlines the assumptions and processes of automatic indexing.

Experimental research that compares retrieval results based on controlled index terms and uncontrolled natural language terms, show that the two vocabularies compliment each other (Lancaster, 2003, p. 257). Natural language has a number of advantages over controlled language and *vice versa*. Controlled language can be used to compensate for the deficiencies of natural language and *vice versa* (Aitchison, Gilchrist & Bawden, 2000). For example, Aitchison, Gilchrist and Bawden (2000, p. 6) point out that natural language is more specific and up-to-date than controlled languages. Likewise, natural language reflects the actual terminological usage of authors in a domain. But, as mentioned above, natural language is ambiguous. The aim of controlled index terms is to reduce the ambiguity of natural language. The terminological and semantic treatment of controlled index terms makes them less specific since they have to convey more meaning as 'broader concepts' (Lancaster, 2003). The dichotomy of natural language and controlled languages is further discussed in Fugmann (1982), Lancaster (1986) and Rowley (1994).

The more advanced the design of a controlled vocabulary, the more likely it will enhance IR performance (Aitchison, Gilchrist & Bawden, 2000). Nevertheless, the major predicament with controlled vocabularies is their complexity. They are typically conservative systems that are not up-to-date, and perhaps most important, they are extremely time consuming and expensive to construct and maintain, especially if they are advanced conceptual networks (Soergel, 1985). It is the drawbacks that makes automatic indexing attractive since it yields reasonable results, it is more manageable, and essentially less time consuming (Lancaster, 2003). Nevertheless, the interest in,

and necessity of, controlled vocabularies has been revived during the last decade with the rapid expansion of new information technologies and online (full text) information systems (e.g., Aitchison, Gilchrist Bawden, 2000). This makes research into new construction methods of controlled vocabularies interesting.

It is our opinion, as stated in Chapter 1, that manual construction of controlled vocabularies is mandatory due to the complexities and dynamics inherent in languages. Further, advanced controlled vocabularies serve a number of secondary roles essential for the understanding of conceptual structures within knowledge domains, this also makes controlled vocabularies highly attractive (Soergel, 1999). However, we fully acknowledge the complexities of construction and maintenance of such vocabularies. We therefore find it beneficial and necessary to combine manual approaches with automatic approaches due to their complementarity. Automatic approaches reduce manual workload but are deficient if used in isolation. Thus, a hybrid of less resource demanding, semi-automatic construction methods, are attractive, as noted by Soergel (1974).

Manual indexing usually means assignment indexing that involves the representation of subject matter by means of terms selected from some form of controlled vocabulary (Lancaster, 2003). The selection of index terms from a controlled vocabulary ideally secures uniformity of indexing, but, inter-indexer consistency ratios vary a great deal (Lancaster, 2003, p. 68). However, this variation is not necessarily due to a defective controlled vocabulary. It is more likely that the results of individual subject analyses differ, so that different index terms are chosen to describe similar content (Rowley, 1994).

In the following sub-section, we present Medical Subject Headings® to display a well-designed and efficient controlled indexing language.

2.2.1 An example of a controlled index language

Different types of controlled vocabularies exist amongst others subject headings and thesauri (e.g., Hodge, 2000; Lancaster, 2003). Medical Subject Headings® (MeSH®) (www.nlm.nih.gov/cgi/mesh), the controlled vocabulary of the MEDLINE® database, which despite its name is a thesaurus, is used for empirical investigations in the present dissertation. What follows is a brief presentation of MeSH®. The MeSH® system is a good example of an advanced and well-balanced controlled vocabulary (Lowe & Barnett, 1994). The MeSH® vocabulary covers the biomedical literature and contains over 22,500⁵ controlled index terms. Each MeSH® term represents a single concept,

⁵ As of 01-02-2004.

and the average assignment of MeSH® terms per document is 10 to 12 (Lowe & Barnett, 1994). MeSH® terms are organized into a set of 15 hierarchies called the ‘MeSH® Tree Structure’ (www.nlm.nih.gov/mesh/meshhome.html). In addition, MeSH® contains a group of some 80 terms called ‘MeSH® Sub-headings’. Subheadings are used to qualify the use of MeSH® terms and allow the user to make searches that are more specific. Empirical evidence suggests that MeSH®-based searches may be superior to natural language searches in MEDLINE® (Lowe & Barnett, 1994). Natural language searches in titles and abstracts produced significantly lower recall than MeSH®-based searches (Lowe & Barnett, 1994).

The ‘MeSH® Tree Structure’ supports a number of useful search strategies. By traversing the tree to the most specific term that represents a concept, the searcher can create very specific search queries. The indexing policy underlines the use of the most specific term available to describe a topic in order to enhance precision (Lowe & Barnett, 1994). The ‘MeSH® Tree Structure’ also allows one to broaden the scope of a search, for example by use of the ‘exploding’ search technique (see Chapter 7). This improves the possibility of higher recall. MeSH® contains a very large set of ‘Entry Terms’ that are used to map non-MeSH®-concept descriptors to appropriate MeSH® terms (Lowe & Barnett, 1994). As is the case in most exhaustively indexed databases, the use of MeSH® indexing terms, when searching MEDLINE®, tend to enhance recall by making it possible to retrieve many of the relevant documents (Lowe & Barnett, 1994). In addition, sufficient devices are available in MeSH® in order to make precision searches as well.

The successful application of these searching principles assumes a high quality of MeSH® indexing consistency. Funk, Reid, and McGoogan (1983) concluded that the MeSH® controlled vocabulary is generally considered a state-of-the-art thesaurus, which is applied consistently. Hence, MeSH® is a fine example of a very useful controlled vocabulary, but not necessarily an easy one to master or manage. The quality of the MeSH® vocabulary means that it is applied in the present dissertation as a benchmark against which the empirically selected candidate thesaurus terms are validated.

The rest of this chapter will present some of the special characteristics of thesauri that eventually can lead to an advanced controlled vocabulary if properly designed and applied. The primary and secondary functions of a thesaurus are presented in section 2.3. In section 2.3.1 and onwards, the main steps and principles of manual thesaurus construction is outlined. Special attention is given to principles of relation to the empirical work presented in Chapter 8.

2.3 Thesauri

The word *thesaurus* derives from Greek and Latin and means “[a] treasury or storehouse; hence, a repository, especially of knowledge; often applied to a comprehensive work, like a dictionary or encyclopaedia” (www.hyperdictionary.com/dictionary/thesaurus).

A thesaurus is essentially a tool for vocabulary control used in both indexing and searching (e.g., Aitchison, Gilchrist & Bawden, 2000). The classical thesaurus is defined as a controlled index vocabulary that consists of a set of preferred controlled terms (descriptors) (Soergel, 1974). Index terms function as conceptual labels for sets of natural language entry words (synonyms or quasi-synonyms), with indication of relations between terms in the form of references from terms to hierarchical broader terms, hierarchical narrower terms or conceptual related terms (Soergel, 1974; Aitchison, Gilchrist & Bawden, 2000). Accordingly, a thesaurus is a structured system of concepts with an indication of hierarchical and associative relationships between these concepts. This makes a thesaurus the most complex type of controlled vocabularies (e.g., Soergel, 1974; Hodge, 2000).

Soergel (1999, p. 1119) points out a range of different functions for a thesaurus:

- A thesaurus is a tool that supports information retrieval, for example:
 - It supports searching, particularly knowledge-based support for end-user searching, including hierarchically expanded searching;
 - It also supports indexing, e.g., it provides a rich set of terms that eventually increases the chances of successful retrieval by redundant indexing.
- A thesaurus functions as a ‘semantic road map’ to individual domains and the relationships among domains, for example:
 - It maps a concept space, relate concepts to terms, and provide definitions, thus providing orientation and serving as a reference tool.

These functions correspond to the conceptions of others, that a thesaurus serves two overall functions, a primary function in relation to IR, and a secondary function as a ‘semantic road map’ (Aitchison, Gilchrist & Bawden, 2000). We can establish that a thesaurus is a tool that supports and mediates the understanding and use of concepts and terms for different purposes through its internal structure (Soergel, 1999).

Numerous definitions of thesauri exist across fields such as computer science, artificial intelligence and library and information science (e.g. Sparck Jones, 1992; Aitchison, Gilchrist & Bawden, 2000). They vary from quite modest definitions that

do not specify types of conceptual relations, to more specific definitions that clearly define the conceptual relations. Schütze and Pedersen (1997, p. 307) is an example of a modest definition: “[w]e define a thesaurus as simply a mapping from words to other closely related words”. In contrast, Miller (1997, p. 489) gives a more elaborate definition of a thesaurus as “a lexico-semantic model of a conceptual reality or its constituent, which is expressed in the form of a system of terms and their relations, offers access via multiple aspects and is used as a processing and searching tool of an information retrieval unit. [...] Hence, it appears that a process of thesaurus construction is a process of simulation in a lexical form: 1) of the whole universum of realities and concepts or its part, and 2) of hierarchical and associative connections and relations between these realities and concepts” (Miller, 1997, p. 489).

Essentially, these two definitions reflect very different views upon the role and functions of thesauri. Schütze and Pedersen (1997) reflect the view basically held in computer science that a thesaurus is a recall device used in information retrieval. Chapter 4 will elaborate on this view. Whereas Miller (1997) reflects views in library and information science that tends to accredit the thesaurus with more advanced structural functions as well. We work along the lines of Miller (1997).

2.3.1 *Approaches to traditional manual thesaurus construction*

This section outlines the main steps in the traditional thesaurus construction process. The purpose is to give an overview of the process; though some attention is given to aspects of interest to the empirical work presented in Chapter 8.

Before thesaurus construction commence, some preliminary considerations concerning its purpose and scope must be made. Thesaurus construction is a technical process that follows the general principles and recommendations outlined in the numerous standards and guidelines established for this purpose (e.g., Soergel, 1974; ISO 2788, 1986; Aitchison, Gilchrist & Bawden, 2000; ANSI/NISO Z39.19-2003, 2003). However, the results of the thesaurus construction process vary according to several factors. The result of the construction process is the product of numerous subjective decisions made by the thesaurus constructor(s) in relation to the purpose and scope perceived for the thesaurus.

Construction of a subject specific thesaurus must take into consideration the needs of the information system and the users in a domain (i.e., information environment) (Holm & Rasmussen, 1961; Soergel, 1974; Foskett, 1981; Lykke Nielsen, 2002). For example, whatever the guidelines and standards may suggest, decisions concerning associative relationships must be made according to the way concepts are used and related by users in the domain. The scope and complexity of the domain will provide

some indication of the scope and complexity of the thesaurus. Accordingly, a thorough preliminary analysis of the domain, its users, information sources, etc., is essential before actual thesaurus construction commences. As will be illustrated theoretically in Chapter 5, and empirically in chapter 8, bibliometrics is a useful tool in this respect. Implicitly, most guidelines do recommend that thesaurus construction is based on some form of analysis of the information system as a whole, pointing to considerations concerning the subject area, type of literature, quantity of literature, system users, resources available, request and query types etc. (Aitchison, Gilchrist & Bawden, 2000).

Different views exist on how to obtain the necessary domain knowledge needed for thesaurus construction. According to Lykke Nielsen (2002, p. 32), the so-called user-centered approaches are the most dominant in this respect. The user-centered approaches are originally based on the development of indexing systems from empirical studies of individual users' information needs. In this sense, the thesaurus construction process essentially becomes a system design process (e.g., Mark Pejtersen, 1980; Soergel, 1985). However, user characteristics are not uniform and contextual and situational factors have a profound influence on user's information search behaviour. According to Lykke Nielsen (2002), this diversity in user characteristics makes it difficult to develop user models on which to base information system design. Today the user-centered approach has evolved into a 'holistic view' that integrates all components of the interactive IR process (e.g., Ingwersen, 1996; 2001; Vakkari, 1997). To this view, knowledge of the situational context that surrounds the user is especially important.

Criticism has been levelled at the user-centered approach. It is accused of being defective due to individualism, which means that it is based solely on studies of individual users' information needs and information behaviour (e.g., Hjørland, 1997; Jacob & Shaw, 1998; Jacob, 2001). The main argument of these critics is that an individual user's behaviour and knowledge structures are formed through participation in a socially grounded domain (Lykke Nielsen, 2002). To the critics, knowledge is created through, and embedded with acting, so system analysis should study the domain and its characteristics and not the individuals acting in it (Lykke Nielsen, 2002). Hjørland (1997) provides a framework for analysis of scientific knowledge domains (discourse communities). To him, individual information needs is not a valid starting point for information system analysis. Instead, he recommends an examination of the information and communicative structures of a scientific knowledge domain by use of 'domain analysis' (Hjørland & Albrechtsen, 1995; Hjørland, 1997; Hjørland, 2002). This includes a range of different approaches such as

bibliometrics, historical, epistemological, and terminological studies etc. Jacob (2001), however, does not reject individual information needs as a basis for analysis of an information environment (domain). Nevertheless, Jacob (2001) argues that individual users as objects of study cannot be isolated. She focuses on the socio-cognitive nature of information behaviour and on the environmental dimensions that influence the user's need for and use of information. To Jacob (2001), an understanding of human cognitive activities should incorporate these factors.

According to Lykke Nielsen (2002), this critique is an important one but essentially unnecessary. Researchers within the user-oriented approach, as mentioned above, have pointed out the importance of studying individual users "... with a view to obtaining a general picture of the work domain and the work tasks that trigger the information need and govern the use of information" (Lykke Nielsen, 2002, p. 33).

The approach to thesaurus construction and maintenance presented in this dissertation is based on bibliometrics. According to Hjørland (2002, p. 450) "[b]ibliometric studies organize sociological patterns of explicit recognition between individual documents". Accordingly, we apply bibliometric methods in an attempt to identify and study agreed upon concepts, and relations between such concepts, from the communicative structure within a scientific domain. Our focus is therefore on written scientific knowledge structures, as suggested by Hjørland (e.g., 1997). Even so, depending on domain, work tasks, and scope, scientific knowledge structures may not be a sufficient basis for thesaurus construction. We therefore agree with Lykke Nielsen's (2002) 'holistic approach' to thesaurus construction that emphasizes the usefulness of different methods for different purposes.

2.3.2 *Main steps in thesaurus construction*

The construction of a thesaurus requires a major intellectual effort (e.g., Soergel, 1974). The knowledge gained about the information environment should be used as a guide and govern the decisions taken during the main steps of thesaurus construction. The main steps of manual thesaurus construction are:

- Term selection
- Concept analysis
- Structuring of concepts and terms
- Displaying the thesaurus.

2.3.3 Term selection

Thesaurus construction requires selection of a set of candidate terms. In a controlled vocabulary, an index term can be either a preferred term or a non-preferred term (Aitchison, Gilchrist & Bawden, 2000). The preferred terms serve as focal points where the information about a concept is stored; as a result, they are used in indexing as descriptors. According to ANSI/NISO Z39.19-2003 (2003), literary warrant is the guiding principle for the selection of the preferred form of a descriptor. Analogously, the non-preferred terms denote 'equivalent' terms to the preferred term, i.e., spelling variants, synonyms, and quasi-synonyms. Non-preferred terms are not used in indexing, but provide an entry point from which the user may be directed by the instruction USE to the appropriate preferred term (Aitchison, Gilchrist & Bawden, 2000).

Three overall methods are suggested for selection of candidate terms, the committee method, the empirical method, and a combination of the two (ANSI/NISO Z39.19-2003, 2003). In the committee method, experts in the subject domain of the thesaurus draw up a list of key candidate terms in the field and indicate the relationships among them. The empirical method extracts terms from documents, either manually by humans, or automatically by use of computers.

The empirical thesaurus construction method can be divided into a deductive method and an inductive method (Aitchison, Gilchrist & Bawden, 2000). In the deductive method, no attempt is made to control the vocabulary, or to determine relationships between terms, until a sufficient number of terms has been collected. A group of experts then reviews all terms. They should first identify terms that represent the broadest classes, and then allocate remaining terms to these classes on the basis of their logical relationships, so that the hierarchies tend to be established on a broader-to-narrower basis (Aitchison, Gilchrist & Bawden, 2000, p. 146). In contrast, if the inductive method is used, terms are admitted to the thesaurus when they are encountered in the literature. Vocabulary control is applied from the outset and when each term is selected, it is designated as a member of one or several broader classes, which are constructed on an *ad hoc* basis (ANSI/NISO Z39.19-2003, 2003). The thesaurus is therefore constructed on a narrower-to-broader term basis.

In practice, more than one of these approaches is likely to be employed at one stage or another during the construction of a thesaurus (ANSI/NISO Z39.19-2003, 2003).

Obviously, it is a major issue in thesaurus construction to identify suitable sources from where candidate terms can be collected. Two types of sources exist, prearranged standardized sources and open-ended non-standardized sources (Soergel, 1974).

Prearranged sources, are already existing descriptor lists, classification schemes, thesauri, dictionaries, etc., require less effort in the gathering of material. The terms

are already standardized and some resources even indicate term relationships. The use of prearranged sources also contains some risk. Clearly, terms collected and arranged for another purpose, and from another viewpoint, may not fit the purpose and scope of the thesaurus being constructed (Soergel, 1974).

Open-ended sources can be grouped into written sources and unwritten sources (Aitchison, Gilchrist & Bawden, 2000). The most traditional written sources are documents within a subject literature, where terms are selected from titles, abstracts, paragraphs, or the full-text. This is also called a text corpus or the global approach to automatic thesaurus construction, see Chapter 3 and 4.

Another written source used for term selection is recorded lists of search requests, or interest profiles, from users (Soergel, 1985). Open-ended sources are attractive as they may reflect the current usage of terminology in a domain, as expressed in its written corpus of text (Lykke Nielsen, 2002, p. 36). Unwritten sources relate to the knowledge and experience of users and subject specialists. As mentioned in section 2.3.1, the user-centered approach, where subject specialists are used as a source for thesaurus construction, has generated some controversy (e.g., Jacob, 2001; Hjørland, 1997). The main criticism is that subject expert's view of a domain and its conceptual structure is not necessarily representative for the domain as a whole, i.e., it is an individualistic view. It is certainly true that opinions can differ between subject specialists. However, consensus on terminology and structure can also be observed (e.g., Small 1978, 1982; Rees-Potter, 1989). Further, differences in opinion are not necessarily a problem since it can point to terminological and conceptual problems relating to the domain. Hence, subject specialists are *one* source for construction and validation of thesauri, but their knowledge should be substantiated by other knowledge sources.

As a result, a thesaurus constructor must choose a set of sources that are current, as complete as possible, and considered authoritative, in order to select appropriate terminology. Nevertheless, it can be very difficult to identify an appropriate set of sources that exhaustively cover a knowledge domain and its topics (Lykke Nielsen, 2002, p. 36). See Chapter 7 for a bibliometric approach to identification of an appropriate set of 'written sources'.

The extraction process of candidate terms from written open-ended sources is done in two tempi. First, a 'literature scanning' of the source identify candidate terms, where possible, terms in a thesaurus should be nouns or noun phrases (Aitchison, Gilchrist & Bawden, 2000). Second, terms are extracted from the documents either manually or automatically. Soergel (1974; 1985) outlines several criteria for extraction:

- Usefulness for searching and retrieval;
- Part of logical structure;
- Frequency of use in documents, indexing, search requests etc.;
- Frequency of inclusion in other conceptual sources such as classification schemes etc.

According to ANSI/NISO Z39.19-2003 (2003), words and phrases drawn from the literature of the field should determine the formulation of descriptors. Sparck Jones (1992, p. 1606) states: "... to function effectively, thesaurus descriptors must be derived from, or at least be strongly motivated by, the particular scientific literature for which they were to be used. [...] The set of terms for a thesaurus would thus be obtained by an examination of the literature of a field". She continues, "... it should be possible to identify its terms by considering the way natural language words behave in the text of subject literature" (Sparck Jones, 1992, p. 1607). Consequently, when two or more term variants have literary warrant, the most frequently used term should be selected as the descriptor (ANSI/NISO Z39.19-2003, 2003). Similarly, Aitchison, Gilchrist and Bawden, (2000) state that a frequency of occurrence list will assist not only in suggesting the leading candidate terms, but also the choice of preferred terms among synonyms.

Chapter 5 outlines the feasibility of bibliometric methods for analysis of subject literatures. Likewise, automatic term selection is a major focus of the present dissertation. Chapter 3 and 4 outlines the generic process of automatic thesaurus construction, which is rooted in statistical term frequency analysis. As an alternative, Chapter 6 presents our semi-automatic approach to thesaurus construction based on bibliometric methods and statistical term frequency analysis. This approach suggests that automatically selected index terms, and their statistically derived co-occurrence relationships, should instead provide guidance for thesaurus constructors on term groupings and term relationships. In other words, complement the manual thesaurus construction process.

2.3.4 Concept analysis

The next step of thesaurus construction concerns conceptual analysis (Aitchison, Gilchrist & Bawden, 2000). The first obvious action is to regulate the form of the admissible terms by vocabulary control. This can involve control of the grammatical forms, spelling variants, singular and plural forms, abbreviations and acronyms, punctuation and capitalization, as well as compound forms (e.g., ISO 2788, 1986).

Similarly, decisions must be made whether to admit, and how to treat, certain types of terms, such as loan words, trade names, proper names, loan words, slang words, abbreviations, acronyms, homographs, etc. For a detailed examination of these issues see ISO 2788 (1986), ANSI/NISO Z39.19-2003 (2003), and the practical manual by Aitchison, Gilchrist and Bawden (2000). However, we will consider some essential aspects. Usually, indexing terms consist of nouns and noun phrases. The most common form of noun phrases is the adjectival phrases. The less common, but still admissible, form is the prepositional form. The ANSI/NISO Z39.19-2003 (2003) recommends that when possible, the noun phrase should exclude prepositions and that, for example 'EDUCATIONAL PHILOSOPHY' should be preferred to 'PHILOSOPHY OF EDUCATION'. Indexing terms in the form of prepositional noun phrases should be limited to concepts which cannot be expressed in any other way, or that have become idiomatic.

According to Aitchison, Gilchrist and Bawden (2000), the decision on singular and plural form should depend on the traditions of the thesaurus' potential language communities. In English language communities, indexing terms are usually divided into two groups, concrete entities and abstract concepts. Special prescription then exists of when to use singular or plural form on these groups.

As defined by ANSI/NISO Z39.19-2003 (2003) an index term (single or compound) is 'the representation of a concept'. ANSI/NISO Z39.19-2003 (2003, p. 35) defines a concept as "[a] unit of thought, formed by mentally combining some or all of the characteristics of a concrete or abstract, real or imaginary object. Concepts exist in the mind as abstract entities independent of terms used to express them". The concepts represented by descriptors can be grouped into general types, such as a) things and their physical parts; b) materials; c) activities or processes; d) events or occurrences; e) properties or states of persons, things, materials, or actions; f) disciplines or subject fields; g) units of measurement etc." (ANSI/NISO Z39.19-2003 (2003, p. 35). Alternatively, concepts can be represented by descriptors that belong to fundamental categories applied in facet analysis (Ranganathan, 1987). The ANSI/NISO Z39.19-2003 guidelines (2003) outline three main categories and their sub-divisions, concrete entities, abstract concepts, and proper nouns.

Scope notes (SN) are used in a thesaurus to indicate restrictions on meaning of a preferred term; to specify the range of topics covered by a concept; and to convey instructions to indexers on how indexing terms should be used, especially regarding compound terms (Aitchison, Gilchrist & Bawden, 2000). Definitions are related to the role of scope notes. Definitions are attached to descriptors to clarify meaning and usage due to language differences, and to supplement the meaning conveyed by the

thesaurus structure (Aitchison, Gilchrist & Bawden, 2000). Definitions tend to be most required in social science and humanity thesauri, to clarify imprecise terminology, which occurs more often in these subject areas (Hudon, 1996; Svenonius, 1997; Aitchison, Gilchrist & Bawden, 2000). Nevertheless, from a work domain study in a pharmaceutical company, Lykke Nielsen (2002) concludes that thorough definitions are also needed within pharmacology, a discipline within the life sciences, which normally is considered to have a well-defined and consistent vocabulary. It is not clear whether the cross-disciplinary settings of the work domain dictates the need for definitions, or if indeed the vocabulary of pharmacology is blurred and in need of descriptor definitions.

2.3.4.1 Compound terms

The concept of specificity is essential to IR, natural language, and controlled vocabularies. The specificity of a controlled vocabulary depends on the ability of the descriptors to express the subject in depth and detail (Lancaster, 2003). The degree of specificity affects the ability of a controlled vocabulary to retrieve precise information and avoid unwanted information (Soergel, 1974).

The use of compound terms (level of pre-coordination) strongly affects the degree of specificity in a controlled vocabulary (Aitchison, Gilchrist & Bawden, 2000). The more complex the index terms, the more specific the vocabulary will be, and the more numerous the total number of terms. The ISO 2788 standard (1986) and ANSI/NISO Z39.19-2003 guidelines (2003) recommend that descriptors should represent simple or unitary concepts as far as possible. However, compound terms are more precise and specific than single, unitary terms combined during search to represent the same concept. It is thus a very important aspect of thesaurus construction to decide when to factor compound terms into simpler terms, and when it is appropriate for system performance to retain the compound term.

A thesaurus with a majority of single terms is said to have a low pre-coordination level, and one with many two- or three-word compound terms is said to have a high pre-coordination level (Aitchison, Gilchrist & Bawden, 2000, p.38). Compound terms are of special interest to this dissertation. We apply a sophisticated natural language parser to identify and extract noun phrases from the citation context of documents. Most often noun phrases are compound terms, see Chapter 3. Candidate thesaurus terms are to be selected from the set of extracted noun phrases, however, not until the phrases have been subjected to conceptual analysis. Chapter 4 introduces parsing, Chapter 6 presents the parser we apply, and Chapter 8 outlines parsing results from a case study in periodontology. Similarly, the syntactical structure of compound

terms is of special interest to this dissertation. Recommendations have been set out for manual syntactical factoring of compound terms (Aitchison, Gilchrist & Bawden, 2000). Some of the principles used in syntactical factoring are also exploited in linguistic approaches to automatic thesaurus construction (e.g. Ruge, 1992; Grefenstette, 1994), see Chapter 4 and Chapter 8 for empirical results.

A compound term can be separated in two parts for syntactical analysis, the *focus* and the *difference* (Aitchison, Gilchrist & Bawden, 2000, p. 39). The *focus*⁶ is the *noun* component. The noun component identifies the broader class of things or events to which the term as a whole refers. For example, DISEASE in the adjectival phrase PERIODONTAL DISEASE and PHILOSOPHY in the prepositional phrase PHILOSOPHY OF EDUCATION. In a one-word term such as DISEASE or PHILOSOPHY the word *is* the focus.

The *difference*⁷ is the part of the compound term that refers to a characteristic, or a logical difference, which, when applied to a focus, narrows its connotation and so specifies one of its subclasses (Aitchison, Gilchrist & Bawden, 2000, p. 39). For example, PERIODONTAL, which specifies a subclass of diseases in PERIODONTAL DISEASE and OF EDUCATION which specifies a subclass of philosophy in PHILOSOPHY OF EDUCATION.

According to ANSI/NISO Z39.19-2003 (2003), for a compound term to be acceptable as a descriptor, it should express a single concept or unit of thought, which is capable of being arranged in a hierarchical relationship. Further, a compound term may be employed so frequently within the literature of the domain covered by the thesaurus that splitting the term into its components would be unacceptable to users who consider it a lexeme. ANSI/NISO Z39.19-2003 (2003, p. 12) therefore recommends some rule-of-thumbs of when to include a compound. For example, if splitting the parts would lead to ambiguity or a loss of meaning like in PLANT FOOD. Another example is if one component of the term is not relevant to the scope of the thesaurus, or is too vague to exist as an independent term, like in COMPOSITE DRAWINGS or FIRST AID (non-distinctive element underlined) (Aitchison, Gilchrist & Bawden, 2000, p. 42).

Syntactical factoring is the technique applied to compound terms that are amenable to morphological analysis into separate components, each of which *can* be accepted as an indexing term in its own right. For example, COTTON SPINNING factors into two terms COTTON + SPINNING. If the factored compound term is likely to be sought as

⁶ Focus is also known as the genus term or the *head* noun in a noun phrase (see Chapter 3).

⁷ Difference is also known as the species term or *modifier* in a phrase (see Chapter 3).

an entry point, the term as a whole should be entered in the thesaurus as a non-preferred term, and a reference should be made to the components of the term used in combination (Aitchison, Gilchrist & Bawden, 2000, p.43).

Three recommendations for syntactical factoring are set out in ANSI/NISO Z39.19-2003 (2003, p. 12) and Aitchison, Gilchrist and Bawden, 2000 (pp. 44-46), that is 'focus and differences', 'transitive action' and 'intransitive action'. They are general criteria and not mandatory instructions to be rigidly applied in all circumstances. An additional recommendation, suggested by Lancaster (1986), is that a compound term that represents two different principles of division should always be factored. Finally, the differencing rules mentioned in the *Precis manual* (Austin, 1984) reject as 'lead-terms' those compound terms which include a focus modified by more than one difference at the same level.

As mentioned, the ANSI/NISO Z39.19-2003 (2003) does allow factoring rules to be overridden, for example, when the indexing in a special field requires special treatment. This means that the nature of the terminology in a given field can call for special criteria to regulate the treatment of compound terms. Factoring rules are indeed *only* recommendations; but, as pointed out in Chapter 8, they are valuable recommendations for terminological control of automatically extracted compound terms.

Compound terms are specific, they are able to prevent 'false drops' and ensure precision in a search. The use of compound terms in a thesaurus specifies relational information about descriptors otherwise lost if the compound term is factored (Aitchison, Gilchrist & Bawden, 2000). A thesaurus that contains a high number of compound terms is expensive to construct and maintain due to its greater size and complexity. Factored terms enhance recall because they are less specific, and they allow for a smaller, less complex vocabulary. To be most effective, a thesaurus should include a substantial entry vocabulary that links the compound form of the factored terms to the factored components (Aitchison, Gilchrist & Bawden, 2000).

Factoring of compound terms, and thereby the decision on the specificity of the vocabulary, essentially rests on domain-specific conditions, as well as the coverage of the thesaurus. The generality of a subject area affects the pre-coordination level of a vocabulary. In a limited field, thesaurus terms can be used without qualification to represent concepts that can be ambiguous in a wider context (Aitchison, Gilchrist & Bawden, 2000). It is, however, a recurring and very difficult problem of thesaurus construction to determine the level of specificity (Lykke Nielsen, 2002, p. 41).

2.3.5 Structuring of concepts and terms

The one thing that separates a thesaurus from other controlled vocabularies is its ability to distinguish and display relationships between the concepts and terms it contains (Aitchison, Gilchrist Bawden, 2000). The conceptual network of a thesaurus provides important conceptual knowledge and understanding to the user of the domain (Soergel, 1999). A conceptual network makes it possible for the user to explore the knowledge domain by browsing its conceptual structure. Thus, the meaning and usage of the vocabulary of the knowledge domain can be more easily comprehended. A basic rule in thesaurus construction and maintenance is that a thesaurus structure should be kept relative simple (Aitchison, Gilchrist & awden, 2000).

There are two broad types of relationships within a thesaurus. The first is at the micro level and concerns the semantic links between individual terms. The second is at the macro level and is concerned with how the terms and their inter-relationships relate to the overall structure of the subject field (Aitchison, Gilchrist Bawden, 2000).

It is difficult to discover the first type of relationship, the micro level inter-term relationship, without first establishing the structure of the thesaurus at the macro-level (Aitchison, Gilchrist & Bawden, 2000). To determine the structure of a subject field, and the inter-term relationships, classification techniques are used. These include class arrangement, systematic classification, facet analysis, and automatic classification methods (Aitchison, Gilchrist & Bawden, 2000); the latter is discussed in Chapter 4 and applied in Chapter 8. Three basic inter-term relationships exist; these are the *equivalence* relationship, the *hierarchical* relationship, and the *associative* relationship (ANSI/NISO Z39.19-2003 (2003)).

2.3.5.1 Equivalence relationship

According to ANSI/NISO Z39.19-2003 (2003), the *equivalence relationship* is the reciprocal relation between preferred and non-preferred terms where two or more terms are regarded, for indexing purposes, as referring to the same concept. Thus, they form an *equivalent* set of terms. The preferred term is the one chosen to represent the concept in indexing, see above, while the non-preferred terms are the ones not selected. The non-preferred terms form an entry vocabulary that directs the user from terms not selected to those that are. In a traditional thesaurus used for indexing the user is lead to the 'correct' descriptor by non-preferred synonym expressions. In a thesaurus used for searching the non-preferred alternative search terms, representing the same concept, are listed to assist the user in increasing recall or selecting more specific terms (Lykke Nielsen, 2002). As we outline in Chapter 4, systems that

facilitate automatic *query expansion*, basically broaden a query by automatically incorporating 'synonymous relations' in the search (Efthiminades, 1996).

The equivalence relationships include synonyms, lexical variants, quasi-synonyms, as well as factored and unfactored forms of compound terms. ANSI/NISO Z39.19-2003 (2003) defines synonyms as terms whose meaning can be regarded as the same in a wide range of contexts, so that they are virtually interchangeable. In common linguistics, synonyms are not common, but they do occur more frequently in scientific terminology (Aitchison, Gilchrist & Bawden, 2000). This is due to the proliferation of trade names, popular names, and other variations depending on local usage and opinion or on etymological root (Aitchison, Gilchrist & Bawden, 2000). A logic consequence of controlled vocabularies in all fields, not only scientific, is that there are more synonyms in a controlled language than in natural language, because the meanings of terms are intentionally limited in controlled language (Aitchison, Gilchrist & Bawden, 2000). In ANSI/NISO Z39.19-2003 (2003), a distinction is made between synonyms, which are different terms for the same concept, and lexical variants, which are different word forms for the same expression, such as spelling, grammatical variation etc. Quasi-synonyms (also known as near-synonyms) are defined as terms whose meanings are generally regarded as different in ordinary usage, but they are treated as though they are synonymous for indexing purposes ANSI/NISO Z.39.19-2003 (2003). As a result, quasi-synonyms include terms that have a significant overlap. The following conventions are used to express the reciprocal relation: UF (use for) written as a prefix to the non-preferred term, and USE (or U) written as a prefix to the preferred term.

The selection of preferred terms from possible synonyms should always be influenced by the needs of the thesaurus. These will vary, depending on such factors as whether subjects are treated at a general or detailed level, or from the popular or specialist scientific viewpoint. Whichever alternatives are adopted, they should be applied consistently throughout the thesaurus. Every effort should also be made to find all appropriate non-preferred synonyms for the preferred terms, in order to enrich the entry vocabulary and with it the usefulness of the thesaurus as a recall improvement device.

2.3.5.2 Hierarchical relationship

According to ANSI/NISO Z39.19-2003 (2003), the *hierarchical relationship* illustrates levels of superordination and subordination. A superordinate descriptor represents a *class* or *whole*, whereas the subordinate descriptors refer to its members or parts. In a thesaurus, hierarchical relationships are used to locate broader and narrower concepts

in a logically progressive sequence. The hierarchical relationship is a basic feature of a thesaurus that distinguishes it from unstructured controlled vocabularies (Foskett, 1981; Aitchison, Gilchrist & Bawden, 2000).

A hierarchical structure is an important factor for improvement of recall as well as precision (Aitchison, Gilchrist & Bawden, 2000). In addition, the display of a descriptor, and its hierarchical relations to broader and narrower terms, in effect helps the users to clarify the scope and use of such a term (Aitchison, Gilchrist & Bawden, 2000). As is the case with equivalent relationships, hierarchical relationships are reciprocal and are set out in a thesaurus using the following conventions: BT (i.e. broader term) written as a prefix to the superordinate term, and NT (i.e. narrower term) written as a prefix to the subordinate term). The hierarchical relationship comprises the generic, whole-part, instance, and polyhierarchical relationships:

- The *generic* relationship identifies the link between a class or category and its members or species (ANSI/NISO Z39.19-2003, 2003). It has the mathematical property of inheritance, whereby what is true of a given class is also true of all the classes subsumed under it. The relation is also known as the inclusion relationship and it has long been used in biological and zoological taxonomies (Aitchison, Gilchrist & Bawden, 2000).
- The hierarchical *whole-part* relationship covers situations in which one concept is inherently included in another, regardless of context, so that the descriptors can be organized into logical hierarchies, with the whole treated as a broader term (ANSI/NISO Z39.19-2003, 2003). Some examples are organs of the body, geographical location, discipline or field of study. Part-whole relationships are most common in thesauri of narrow subject fields (Aitchison, Gilchrist & Bawden, 2000). In some cases whole-part relationships can in fact be regarded as associative relationships, with the above mentioned examples as an exception (Aitchison, Gilchrist & Bawden, 2000).
- The *instance* relationship identifies the link between a general category of things or events, expressed by a common noun, and an individual instance of that category, often a proper name (ANSI/NISO Z39.19-2003, 2003). An example would be MOUNTAIN REGIONS, narrow terms: ALPS and HIMALAYAS (Aitchison, Gilchrist & Bawden, 2000, p.59)
- The *polyhierarchical* relationship is used to express the fact that some concepts belong, on logical grounds, to more than one category (ANSI/NISO Z39.19-2003, 2003). This phenomenon can be applied to both the generic and hierarchical whole-part terms. In fact, polyhierarchical relations can be a problem in thesaurus

construction since too many hierarchies tend to give conceptual overload (Aitchison, Gilchrist & Bawden, 2000).

2.3.5.3 Associative relationship

An associative relationship covers associations between descriptors that are semantically or conceptually associated to such an extent, that the link between them should be made explicit in the thesaurus, on the grounds that it may suggest additional descriptors for use in indexing or retrieval (ANSI/NISO Z39.19-2003, 2003). The associative relation is reciprocal, and is distinguished by the abbreviation RT (related term), for example: CELLS related term CYTOLOGY.

The associative relationship is vital to the users of a thesaurus. To display the associative relations, to and from a concept, essentially gives the user more knowledge of the concept and its context. This feature is of great importance for a conceptual network such as a thesaurus. The ANSI/NISO Z39.19-2003 guidelines (2003) clearly state that the associative relationship is the most difficult one to define. For example, in automatic thesaurus construction, the fact that two terms co-occur together in a significant number of texts warrant a potential relationship (Lancaster, 2003). The relationship between two frequently co-occurring terms is most often ruled out as an equivalence or hierarchical relation (Sparck Jones, 1992). As a result, the relation between two terms, grounded on the fact that they co-occur frequently, is defined as an associative relationship (Lancaster, 2003). Obviously, this is not a precise definition of the associative relationship. In connection, Aitchison, Gilchrist and Bawden, (2000) mention the risk that thesaurus constructors can overload the thesaurus with 'valueless' associative relationships. It is important to make explicit the *nature* of the associative relationship between descriptors linked in this way and settle for criteria for inclusion⁸. Otherwise, RT references may be too numerous or established inconsistently (ANSI/NISO Z39.19-2003, 2003). As a general guideline, whenever one term is used, the other should always be implied within the common frames of reference shared by the users of the thesaurus (ANSI/NISO Z39.19-2003, 2003). Moreover, one of the terms is often an explanation or definition of the other; the term CELLS, for example, forms a necessary part of the definition of CYTOLOGY.

Molholt (1996) draws attention to the weakness of 'associatively-related inter-term links' found in many thesauri. There is a lack of rules on defining and constructing related terms and this cause inconsistency and idiosyncratic application (Aitchison, Gilchrist & Bawden, 2000). However, the *Art and Architecture Thesaurus* (1994)

⁸ As an example, RT could be classified as processes, tools, activities etc.

presents a well-understood and highly stable basis for the application of associatively related links.

Recently, there has been an increase of research into associative relationships ultimately to set out some guidelines (e.g., Bean, 1996; Green, 1996; Schmitz-Esser, 1999). For example Bean (1996) reports on her interesting research to determine which non-hierarchical relationships in MeSH® could be characterized, organized, and structured. She investigates whether there is a consistent operating logic in the pattern of relationships, and whether the selection of relationships can be automated by use of, for example, morphological analysis of medical terminology.

The ANSI/NISO Z39.19-2003 (2003) does give some examples of the most frequently mentioned types of related terms. The most common type is the whole-part associative relationship (ANSI/NISO Z39.19-2003, 2003). Apart from the instances of hierarchical relationships listed above, the whole-part relationship is associative, as it links between terms in different fundamental categories (Soergel, 1974). A thesaurus constructor can apply the so-called 'all-and-some' hierarchical test of validity to establish the type of a relationship, that is, if the test shows that the terms are not hierarchically related, then they must be associatively related (Aitchison, Gilchrist & Bawden, 2000). According to Aitchison, Gilchrist and Bawden (2000, p. 66), it is easier to judge the correct relationship between terms in the science and technology fields than in the social sciences and humanities.

The existing broad categorization of related terms, and their somewhat inconsistent application, will possibly remain acceptable so long as thesauri are used mainly to guide human users in term selection. However, the increasing automated use of thesauri is likely to create a demand for thesauri with relationships that are sub-categorized and more consistently applied (Aitchison, Gilchrist & Bawden, 2000; Dextre Clarke, 2000). A rather different method to identify associatively related terms is by automatic statistical means, this described in Chapter 4, and empirical results are presented in Chapter 8.

2.3.6 *Thesaurus displays*

A final aspect of thesaurus construction is that of thesaurus display. Methods of thesaurus display vary from static printed displays to more or less sophisticated interactive electronic displays. Aitchison, Gilchrist and Bawden (2000) categorize traditional printed thesaurus displays: 1) alphabetical display that shows scope notes, equivalence, hierarchical and associative relationships; 2) hierarchical displays generated from the alphabetical display; and 3) systematic and hierarchical display, that shows the overall structure of the thesaurus and all levels of hierarchy. Two

traditional thesaurus displays are exemplified below (Figure 2.1 and 2.2), in order to illustrate the product of the thesaurus construction process. The first example is from the *Thesaurus of ERIC Descriptors* (1995). It is an entry from the alphabetical display, which shows scope note, equivalence, hierarchical, and associative relationships.

<i>Bibliometrics</i>	
Scope Note	The application of mathematical and statistical methods in the study of bodies of writings to reveal the historical development of subject fields and patterns of authorship, publication, and use.
Broader Terms	Documentation
Narrower Terms	Citation Analysis
Related Terms	Books Cataloguing Classification Cluster Grouping Indexing Information Sources Information Utilization Library Collection Development Literature Reviews Periodicals Statistical Analysis
Used For	Statistical Bibliography

FIGURE 2.1. Example of the term 'bibliometrics' in the *Thesaurus of ERIC descriptors*.

This entry defines the descriptor BIBLIOMETRICS in the scope note. We are informed about the descriptor's hierarchical relationship, with DOCUMENTATION as a broader term (BT), and CITATION ANALYSIS as a narrower term (NT). Hence, BIBLIOMETRICS is part of a three-level hierarchy in this thesaurus. In addition, we are informed of one equivalence relationship: BIBLIOMETRICS is used for (UF) the term STATISTICAL BIBLIOGRAPHY (see Chapter 5 for a historical explanation of this equivalence relation). Finally, eleven associative relationships are indicated. At first, this number of related terms (RT) seems relative high compared to the rest of the entry. However, when looking at the number RT attached to the descriptor entries in the ERIC Thesaurus (1995) in general this number is by no means extraordinary. As mentioned above, there certainly is a risk of lowering the information potential of individual thesaurus entries if the list of RT becomes too long. In such cases, some form of structuring of the associative relations is appropriate (Dextre Clarke, 2000).

The next example is the Medical Subject Heading tree structure. Remember that MeSH® is in fact a thesaurus despite its name.

MeSH® Tree Structures	
Dentistry [E06]	
Periodontics [E06.721]	
Dental Prophylaxis [E06.721.189]	
Gingivectomy [E06.721.321]	
Gingivoplasty [E06.721.384]	
Guided Tissue Regeneration [E06.721.485]	
Periodontal Dressings [E06.721.595]	
Periodontal Index [E06.721.658]	
Subgingival Curettage [E06.721.874]	

FIGURE 2.2. Example of the term 'Guided Tissue Regeneration' in the MeSH® tree structure.

The MeSH® tree structure depicts hierarchies of broader and narrower terms of up to several levels within broad classes and subclasses. Terms can appear in more than one hierarchy, and terms are ordered alphabetically within the same hierarchical level. The tree structure do not include scope notes, synonyms or related terms. As indicated, GUIDED TISSUE REGENERATION is found at the third level, and is a subclass to PERIODONTICS, which again is a subclass to DENTISTRY. An expressive notation links the terms in the tree structure with the alphabetical section of MeSH®, this is illustrated in Figure 2.3 for GUIDED TISSUE REGENERATION:

MeSH® Heading	Guided Tissue Regeneration
Tree Number	E06.721.485
Annotation	note category: a periodontal technique
Scope Note	The repopulating of the periodontium, after treatment for periodontal disease. Repopulation is achieved by guiding the periodontal ligament progenitor cells to reproduce in the desired location by blocking contact of epithelial and gingival connective tissues with the root during healing. This blocking is accomplished by using synthetic membranes or collagen membranes.
Entry Term	Regeneration, Guided Tissue
Previous Indexing	Periodontal Diseases (1986-1991); Periodontal Ligament (1986-1991); Periodontitis (1986-1991); Periodontium (1987-1991); Regeneration (1986-1991)
Online Note	use GUIDED TISSUE REGENERATION to search REGENERATION, PERIODONTAL TISSUE & REGENERATION, ROOT SURFACE 1992-96
History Note	1992; REGENERATION, PERIODONTAL TISSUE & REGENERATION, ROOT SURFACE were see GUIDED TISSUE REGENERATION 1992-96

FIGURE 2.3. Example of the term 'Guided Tissue Regeneration' in the alphabetical display of MeSH®.

The annotation establishes the function of the descriptor, and the scope note depicts its actual role. We are informed of 'previous indexing' and the history of this descriptor. On this basis, we can establish how the periodontal technique of GUIDED TISSUE REGENERATION previously was expressed in the MeSH® vocabulary until it became a descriptor in 1992. In addition, equivalent relations are indicated by providing information on the use of GUIDED TISSUE REGENERATION instead of REGENERATION, PERIODONTAL TISSUE & REGENERATION, and ROOT SURFACE. No related terms are indicated for this descriptor even though, MeSH® does allow this by use of 'see related' references. Further, no eminent hierarchical relationships are provided. Albeit, the tree number does indicate that the descriptor is

residing at the third level of a hierarchy, but no contextual information is given in relation to the hierarchy.

The major difference between ERIC and MeSH® is that MeSH® does not illustrate conceptual hierarchies within the alphabetic display. Conversely, as can be seen in Figure 2.3, MeSH® does provide a substantial description of the role and function of the descriptor, as well as its indexing history and context in MeSH®.

In addition an alternative to the traditional methods of thesaurus display is that of graphic display. This is a rather elusive definition, in that the examples illustrated above do indicate structure and relations through visual arrangement.

Perhaps the simplest type of graphic display is the tree structure. Nevertheless, graphic display usually means presentation of indexing terms and their inter-relationships in the form of a two-dimensional plane (Aitchison, Gilchrist & Bawden, 2000).

The basic notion behind the graphic display of concept structures is to allow users to grasp these structures more readily by making use of a 'spatial metaphor' (e.g., Chen, 2003). Graphic displays appeal to the intuitive powers of the mind (e.g., Tufte, 1983). It is therefore likely that a user enhances his or her understanding of the conceptual relations by studying the visual representation of them. In effect, this should improve the accessibility of the thesaurus (Aitchison, Gilchrist & Bawden, 2000). In this way, the function of graphic displays is similar to traditional displays, except that they appeal to other perceptual abilities of the users.

Early graphic displays of thesauri in library and information science has been centred on graph approaches, such as the hand made 'terminographs' (*SPINES thesaurus*, 1976) and 'arrow graphs' (Rolling, 1965). A conceptual structure can generally be considered a type of graph, as defined in graph theory (e.g., Buckley & Lewinter, 2002). A graph consists of a set of nodes together with a set of edges (undirected) or arcs (directed) that link the nodes together (Buckley & Lewinter, 2002). In the case of a conceptual structure, the nodes are concepts, usually represented by terms, and the arcs are conceptual relations. The conceptual relations may be either paradigmatic (semantic) or syntagmatic (syntactic). The 'arrow graph' shows reciprocal relations that works in both direction between two terms in a network (see Figure 2.4).

FIGURE 2.4. 'Arrow graph' by Rolling (1965), reproduced from Foskett (1981, p. 128).

A more complex overlay to the graph display is the clustered 'box-chart' that not only shows 'neighbour' or contiguous relationships but also include hierarchical relationships to the main descriptor in each box (Rolling, 1965). For example, the 'box chart' reveals that the main descriptor (the one that defines the cluster) TERMINOLOGY has a narrower term in WORDS. Conversely, the 'arrow graph' shows that three main descriptors have relationships pointing *to* TERMINOLOGY. In addition, two relationships go out from TERMINOLOGY; the one that points to LANGUAGES is reciprocal.

Graphic thesaurus display can be traced back to Doyle's (1961; 1962) seminal work on term association maps. Essentially, such an 'aerial view' of a larger part of a conceptual network of relations allows for somewhat more flexibility and comprehension, than traditional tree structures provide.

An often-neglected aspect of graphic displays is that they can be useful as tools during the thesaurus construction process as well (Aitchison, Gilchrist & Bawden, 2000). The 'spatial metaphors' inherent in graphic displays can perhaps give thesaurus constructors a more thorough and intuitive understanding of a conceptual network, when it is projected in a larger two- or three-dimensional context? Soergel (1999) emphasizes, that advances in computer technology has made it possible to construct

much more sophisticated and dynamic thesaural displays. He points to interactive graphical representation of conceptual networks, such as graphs, in two- or three-dimensions. Essentially, what Soergel (1999) calls for is to modify or update the early approaches to graphical thesaurus display, such as the 'arrow graph' with the available technology of today. An example of this is the Visual Thesaurus (www.visualthesaurus.com). Obviously, the display of this thesaurus is inspired by Rolling's (1965) early work on the 'arrow graph'. In addition, the vital aspects of scaling and dynamics are incorporated into the thesaurus. This enables the user to focus on different parts of the network, as well as browsing the conceptual relations.

An approach to support and fulfil Soergel's (1999) vision is the research of 'information visualization'. Information visualization is a cross-disciplinary research field that investigates possible computerized graphic displays of information phenomena (e.g., Lin, 1997; Chen, 1999). The goal of information visualization is to reveal patterns and trends, and give new insights to information phenomena, by use of computer supported, interactive, visual representations of abstract data, in order to amplify cognition (Card, Mackinlay, & Schneiderman, 1999). Thus, automatic algorithmic approaches are especially well suited for such graphical displays (Soergel, 1999). Consequently, we will apply visualization techniques in order to investigate their potential value for thesaurus construction. Chapter 8 presents graphical displays of conceptual networks.

2.4 Summary

Chapter 2 has introduced some important characteristics concerning thesauri, their purpose, setting in relation to IR and the manual construction process. A thesaurus is a highly complex controlled vocabulary, characterized by its three types of conceptual relations: the equivalence, hierarchical, and associative relationships (Aitchison, Gilchrist & Bawden, 2000). A thesaurus is a language tool which primary function is to aid and support indexing and retrieval of documents in an information retrieval system (Soergel, 1999; Aitchison, Gilchrist & Bawden, 2000; Lancaster, 2003). The fundamental problem in IR is language, so a tool like the thesaurus is indispensable, though it is not perfect, and never will be (e.g., Blair, 1990). Natural and controlled languages are said to be complementary in relation to indexing and retrieval (Lancaster, 2003). Consequently, both types of indexing languages, natural language and controlled vocabularies are essential, because they express different aspects,

relations, and structure of language (Soergel, 1985; 1999; Aitchison, Gilchrist & Bawden, 2000; Lancaster, 2003).

Thesauri also serve a number of secondary functions essential for understanding of conceptual structures within knowledge domains (Soergel, 1999; Aitchison, Gilchrist & Bawden, 2000). This raises the thesaurus above the mere role of a recall device in IR (e.g., Peat & Willet, 1991). At the same time, this also increases the intellectual effort needed to construct such an advanced language tool.

Manual thesaurus construction is a huge, time-consuming task of term selection, conceptual analysis, and relational structuring of concepts and terms (Aitchison, Gilchrist & Bawden, 2000). In this chapter, we have outlined the main steps, and indicated some principles, guidelines and problems pertaining to construction. Beyond doubt, a well-designed thesaurus will eventually do good to IR, as well as the secondary functions intended for it; the case of MeSH® is an appropriate example (Lancaster, 2003).

The ever present problems of natural language, effectively rules out automatic methods being applied alone if effective language tools are to be created (Soergel, 1974; Aitchison, Gilchrist & Bawden, 2000). Consequently, it is attractive to investigate the possibilities of a hybrid construction approach that subjugate automatic methods to manual control and is less time-consuming; this is the focus of this dissertation. The aim is to develop and explore a semi-automatic approach to construction and maintenance of thesauri.

Finally, the possibilities for display of a thesaurus or a conceptual network have dramatically changed due to advances in computer technology (e.g., Soergel, 1999; Chen, 1999). It is assumed that graphic displays of conceptual networks can enhance the understanding of such networks, not only from the users' perspective (searcher and indexer), but also for the thesaurus constructor during the construction process. We therefore investigate the usefulness of visualization techniques in thesaurus construction in the dissertation. The semi-automatic thesaurus construction approach is based on bibliometrics, but also utilizes automatic thesaurus construction methods. Consequently, the next two chapters outline the general assumptions and methods behind automatic thesaurus construction.

3. Vocabulary construction: Term selection

Automatic thesaurus construction⁹ is in effect an extension to automatic indexing (e.g., Srinivasan, 1992). Automatic indexing concerns the construction of an indexing vocabulary¹⁰. Automatic thesaurus construction ‘enhances’ this vocabulary by indicating significant term associations, and imposing a structural organization on the selected index terms (Srinivasan, 1992). As a result, automatic thesaurus construction comprises three overall steps, vocabulary construction (term selection), term association and vocabulary organization. Characteristic for these steps are the application of a number of related automatic methods and techniques. They comprise NLP techniques, such as stemming and parsing; statistical methods, such as frequency analysis, weighting and co-occurrence analysis; mathematical models, such as vector space representation; and statistical ordination techniques, such as cluster analysis (Manning & Schütze, 1999; Belew, 2000; Jurafsky & Martin, 2000; Börner, Chen & Boyack, 2003).

The basis for application of these methods is a text corpus in the form of a document collection. The object of study is the full text or an excerpt of a document. The units of analysis are natural language words in these documents. Essentially, NLP techniques standardize words, and statistical methods identify and select index terms and significant term associations. Index terms are weighted and represented in the vector space model, and subsequently grouped by ordination techniques.

Vocabulary construction is the basis for thesaurus construction. The crucial aspect of vocabulary construction is *term selection*. Term selection is based on assumptions about the distributional behaviour of words in natural language text, as well as the position of these words within a document structure. Vocabulary construction and its assumptions are the focus of this chapter. Chapter 4 treats term associations and vocabulary organization in a similar manner.

The present dissertation concerns the development of a semi-automatic thesaurus construction approach. Semi-automatic thesaurus construction means that results produced by automatic methods and techniques are subjected to manual intellectual

⁹ We use the term automatic thesaurus construction when automatic indexing methods are applied to thesaurus construction with minimal human involvement. In addition, we use the term semi-automatic thesaurus construction, when automatic indexing methods are used to support humans in thesaurus construction.

¹⁰ We use the term automatic indexing synonymously with vocabulary construction.

analysis before their actual application (Soergel, 1974). Hence, semi-automatic thesaurus construction also utilizes automatic methods and techniques in the process of vocabulary construction, term association and vocabulary organization.

Our approach does however differ from traditional automatic thesaurus construction approaches on one major account. Similar to Rees-Potter (1987; 1989), our starting point is not distributions of words in natural language text, but instead, distributions of citations in subject literatures. We use citations and references as a preliminary tool to identify and group candidate thesaurus terms. By use of bibliometric methods, citation contexts of a candidate thesaurus term are identified in full text documents. Automatic methods and techniques are subsequently used to process these citation contexts to derive term associations. The basis for application of bibliometric methods for semi-automatic thesaurus construction approach is presented in Chapter 5.

The major difference between a bibliometric approach and a term-based approach to thesaurus construction is therefore related to the assumption applied in the latter, that term distributions are the pivotal basis for construction. Besides this major difference in origin, semi-automatic construction, and bibliometrics in general, share a range of common automatic methods and techniques with automatic thesaurus construction approaches. Consequently, we find it important to outline thoroughly the process of automatic thesaurus construction. In this way, we introduce and discuss automatic methods and techniques used in thesaurus construction, many of which also forms part of the semi-automatic thesaurus construction approach presented in Chapter 6.

We designate vocabulary construction as step 1 in the automatic thesaurus construction process. Step 2 is term associations and step 3 is vocabulary organization, they are both treated in Chapter 4. The aim and process of vocabulary organization can be outlined as below:

1. Vocabulary construction:

The aim is to select and represent candidate thesaurus terms based on the distribution of words in natural language text and perhaps their position within the structure of documents. The process comprises of the following main automatic methods and techniques:

- a. Lexical analysis: identification of individual tokens of words in a text; removal of function words by use of a stop list; conflation of morphologically related words through stemming;

- b. Vector space representation of the remaining word stems; term selection based on word distribution models, and/or structural position in documents;
- c. Addition of weights to the selected index terms in order to reflect relative term importance.

The purpose of the present chapter is to explain the basis, assumptions and processes behind the construction of a vocabulary. Sections 3.1 to 3.3 concern some basic properties of natural language text that are important for automatic index term selection. Section 3.1 focuses on linguistic and lexical aspects of natural language. Section 3.2 point out that document structures can be an important indicator of where to identify appropriate index terms. Traditional automatic thesaurus construction approaches tend to neglect this feature. They treat the whole document text as an unstructured ‘bag of words’ and identify candidate index terms on the basis of word distributions instead. Section 3.3 outlines the most important of these word distributions used for vocabulary construction.

Sections 3.4 to 3.6 concern the automatic process of vocabulary construction. Section 3.4 briefly introduces the notion of stemming and stop words and how these features are utilized in vocabulary construction. Section 3.5 introduces the vector space model, which is the most widely used mathematical model in automatic indexing. The model is also widely used within bibliometrics. We pay special attention to this model, as it is a central aspect of this dissertation. Section 3.6 discusses the application of term weights to indicate the relative importance of index terms within the vector space representation.

Section 3.7 selectively outlines some interesting alternative approaches applicable for identification of index terms. The focus is on document structure and exemplary documents. We give special attention to these alternative approaches since our own approach to identification of candidate thesaurus terms deviates from the ‘bag of words’, and instead concentrates on structural markers (references) in document text. Finally, section 3.8 summaries the main points of this chapter.

3.1 Linguistic and lexical properties of natural language texts

The aim of this section is to introduce some basic properties of natural language, which is the object of study in automatic indexing and this dissertation. Natural language is the most elaborate symbolic system that human beings control and an essential tool for communication (Sperber & Wilson, 1995, p. 173). The representation power of

natural language is unrivalled and provides an effective and expressive tool for communication of content (Sparck Jones, 1991). Automatic indexing of text documents relies on linguistic and statistical processing of natural language in order to represent the contents of texts. The individual words and phrases, and their ordering, in a text, manifest the content of that text (Moens, 2000). In the context of this dissertation, the object of study is the natural language text of documents within a document collection (van Rijsbergen, 1979; Salton & McGill, 1983). In practice, a document refers generically to the unit of text indexed in the system and available for retrieval. It can be a complete logical unit, like a research article, a book or a manual. It can also be part of a larger text, such as a paragraph, a passage of text etc. (Moens, 2000). Texts can be distinguished in many ways. An important distinction is between expository text and narrative text. In expository texts, more emphasis is given to the topics and subtopics of the text (Rau, Jacobs & Zernik, 1989). We focus on scientific texts, which is an important part of expository texts.

Below, we outline some essential characteristics of the basic items of natural language, their mutual ordering, and the influence of text structure, for indexing. We should note that our focus is on the English language. The English language uses compound terms, separated with blanks, and this affects index term identification.

Different conceptions exist to the definitions of word, term, and concept within the fields of terminology, computational linguistics, automatic and manual indexing (Aitchison, Gilchrist & Bawden, 2000; Jacquemin & Bourigault, 2003). In this dissertation, word denotes the unprocessed lexical item as it is found in the text of a document. Term denotes the selected and normalized index term, either a single term or a phrase. Finally, concept denotes entities and abstract concepts found in an indexing language and expressed by a range of index terms. In addition, in automatic indexing the term token is typically used to denote a morphologically processed word. We use token to denote the normalization process of a word until it becomes an index term.

From the perspective of automatic indexing, the basic items of natural language are words and phrases. A word can consist of three components, called morphemes, a root form (stem), and additional affixes (prefix and suffix) (Jurafsky & Martin, 2000). The root form comprises the essential meaning of a word. Words are further divided into lexical classes called parts of speech (e.g., nouns, verbs, etc.). Parts of speech can also be partitioned into content words and function words. The former class contain words that are suitable indicators of the content of a text. The latter class contain words that have important functional or grammatical roles in a text (Halliday, 1989). Function

words belong to syntactic classes, such as articles, pronouns, particles, and prepositions. Function words tend to define how content words are to be used in the sentence, and how they relate to each other (Halliday, 1989). Further, function words are typically very short words and they tend to occur frequently in a text (Jurafsky & Martin, 2000). In the English language, four important lexical classes of content words are considered. Nouns describe objects, events, or substances. Adjectives describe properties of objects. Verbs describe relationships between objects, activities, and occurrences. Finally, adverbs describe properties of relationships or other properties (Jurafsky & Martin, 2000). Lexical meaning refers to the meaning or semantics of words, that is, what words symbolize including their denotations and connotations. The origins and usage of words in certain contexts of texts define lexical meaning.

3.1.1 *The properties of noun phrases in indexing*

Constituency is the notion that groups of words may behave as a single unit. The single unit is a phrase (a constituent). A phrase consists of a headword and its modifiers. The head gives the central concept of the phrase and the modifiers serves to make it more precise (Jurafsky & Martin, 2000). Basically, phrases can be divided into four broad classes of phrases: noun phrases, adjective phrases, verb phrases, and adverbial phrases. A very important aspect in relation to indexing is the fact that phrases are less ambiguous in meaning than the individual words of which they are composed (Jurafsky & Martin, 2000). We focus on noun phrases, as they are the most important semantic units for indexing (e.g., Moens, 2000). We use bibliometric methods in combination with a noun phrase parser to identify candidate thesaurus terms, see Chapter 6 and Chapter 8. Noun phrases are used in indexing to capture a ‘richer linguistic representation’ of document content. Below, we present some of the motivations given by Anick and Vaithyanathan (1997, p. 317) behind the use of noun phrases to describe concepts in natural language text:

- Noun phrases are widely used across sublanguage¹¹ domains to describe concepts succinctly.
- Noun phrases are contiguous and relative easy to detect and extract from natural language text corpora.

¹¹ A sublanguage (SL) is the written or spoken language that is used in a particular field or discipline by people working in the field (Grishman & Kittredge, 1986).

- Unlike many phrasal constructions, which reflect transient relationships among objects noun phrases are generally applied to express tighter, more 'long lived' relationships between concepts. Noun phrases therefore contribute less 'noise'.
- Nouns phrases (single nouns inclusive) account for the bulk of the phrases that show up in actual queries.
- In most cases, the relationship between a noun phrase and the head noun of the compound is a strict conceptual specialization.
- The degree to which a noun participates in phrases appears to be a good measure of the importance of that concept within the sublanguage domain.

The simplest noun phrase consists of a single pronoun, noun, or proper name (Jurafsky & Martin, 2000). The remaining forms of noun phrases consist of a head and one or several modifier words that qualify or specify the head. A core noun phrase (*NP*) has the general form:

$$NP = det^* pre^* \textbf{head} post^*$$

Where *det* is a determiner (article, quantifier, number, etc.); *pre* is a pre-modifier (adjective, noun, etc.); **head** is the central noun in the phrase; *post* is the post-modifier (prepositional phrase, relative clause, etc.); and * denotes a list of zero or more elements.

Normalization of phrases is common when they are applied as index terms (Strzalkowski, 1995). For example, determiners in noun phrases can be eliminated because they are of little interest in indexing (Jurafsky & Martin, 2000). In addition, pre- and post-modifiers are considered to modify the head noun independently from each other. Consequently, $[h, m_1; m_2]$, may be expanded as $[h, m_1]; [h, m_2]$, where *h* is the head and *m* the modifier. This resembles the process of compound factoring described in Chapter 2¹². Indexing of single terms is the most commonly used approach in IR (Belew, 2000). It has turned out to be the most cost-effective approach hitherto (e.g., Kowalski & Maybury, 2000). The primary object of single term indexing is the head nouns in noun phrases. Here, the head is used as an abstraction of the phrase, though this causes the noun phrase to loose specificity.

¹² Although the head-modifier relation implies semantic dependence, it is a pure syntactic relation. The intention is to identify meaningful indexing terms without deep semantic analysis. Therefore, the precise semantic interpretation of any head-modifier relation is a forborne ordered relation.

3.1.2 Lexical cohesion

Linguistic knowledge about the structural cohesion of text is important when natural language is processed and modelled for automatic thesaurus construction (van Dijk, 1997; Moens, 2000). This aspect gives insight to the thematic structures of a text, i.e. the major and minor topics, their central concepts, and most important where to find them.

The first aspect concerns the basic items of natural language text, word and phrases, and how they make a text cohesive through chaining (Halliday & Hasan, 1976). One of the main properties of a connected text is grammatical and lexical cohesion (Halliday & Hasan, 1976; De Beaugrande & Dressler, 1981; Moens, 2000). Cohesion refers to the relations between words and phrases that essentially connect different parts of the text into a meaningful whole. The referring item and the item it refers to, bring about cohesion. There are a number of forms of cohesion, such as reference, substitution, ellipsis, conjunction and lexical cohesion. From the point of view of automatic text analysis, lexical cohesion is of interest.

Lexical cohesion deals with the actual usage of words and phrases in a text and how this creates lexical chains readily available for automatic text analysis (Moens, 2000). A lexical chain is a sequence of semantically related lexical items in the text that spans either short (adjacent words or sentences) or long distances (entire text) (Halliday & Hasan, 1976; Morris & Hirst, 1991). A lexical chain can for example provide a context for the resolution of ambiguous words, enable identification of a concept that words may represent, point to collocations, or help identifying the thematic and discourse structures of a text (van Dijk & Kintsch, 1983).

Two features of lexical cohesion are important in this respect: reiteration and collocation (Boguraev & Neff, 2000). Reiteration is a form of lexical cohesion that involves the repetition of a lexical item. This may involve the simple repetition of the word or phrase, but also includes the use of synonyms, near-synonyms, hyponyms, meronyms, anaphora or acronyms (Boguraev & Neff, 2000). Obviously, these phenomena are the focus of indexing; especially important is the phenomenon of repetition of words. The knowledge of word repetition, and its important cohesive function in text, supports the use of word frequency distributions as models for identification of topics within texts. Section 3.3 presents the most important word distributions used in automatic indexing.

Collocation refers to any pair of lexical items that stand to each other in some recognisable lexical-grammatical or lexical-semantic relation, for example, ‘information’ and ‘retrieval’, or ‘retrieval’ and ‘indexing’ respectively (Vechtomova, Robertson & Jones, 2003). In automatic indexing, lexical-grammatical associations are typically used to identify phrases, and lexical-semantic associations are used to

identify semantically related words that describe the same topic (Vechtomova, Robertson & Jones, 2003). The notion of collocation comes from computational linguistics and is strongly related to the co-occurrence methods used within automatic indexing. Chapter 4 discusses the notion of collocation and its various conceptions.

This section has presented some basic linguistic and lexical characteristics of importance for automatic vocabulary construction. We now turn to the aspect of how and where to identify potential index terms. The following section draws attention to document structures as an important indicator of where to identify appropriate index terms. Section 3.3 focuses on word distributions in individual documents and over a collection of documents as well.

3.2 The use of document structure to identify index terms

Another aspect that can give insight to the thematic structures of a text is that of document structures (Mann & Thompson, 1988; van Dijk, 1997; Moens, 2000).

Experiments by Dillon (1991) clearly demonstrate that readers who are experienced in reading certain document types possess a schematic structure or model of the text that enables them to predict with high levels of accuracy where specific information is located. For example, Hearst and Plaunt (1993), assume that the main theme of scientific articles is discussed throughout the entire text, while the discussion of subtopics is restricted to narrow passages of the text.

The themes, topics and sub-topics of a document characterize its thematic structures. Thematic structures in documents are influenced and shaped by the formal (physical) structure of a document and the discourse structures that somehow govern the written communication in a particular text (e.g., Moens, 2000). Documents have a formal structure. For example, a scientific journal article comprises of a title, an abstract, sections that include section titles, paragraphs within each section etc. (e.g., Bazerman, 1988). Earlier research have shown that formal structural positions within some types of documents, such as titles, abstracts, the first paragraph etc., contain significant content words most appropriate for indexing (Buxton & Meadows, 1977; Diodato, 1982; Bernstein & Williamson, 1984; Bonzi, 1984; Jonak, 1984; Liddy & Myaeng, 1993; Paice & Jones, 1993; Wilkinson, 1994; Losee, 1996; Ries et al., 2001; Yitzhaki, 2002).

Communication by means of natural language text is governed by discourse patterns. Consequently, a specific document text has a specific discourse structure that indicate how language and rhetoric is used to communicate the content (van Dijk,

1997; Moens, 2000). It is acknowledged that knowledge of discourse patterns is indispensable in text understanding, even if this understanding is only partial (Moens, 2000). Discourse structure is therefore a key to the identification of the thematic structure of a document. Discourses usually belong to a specific genre, and genres typically define the conditions for the discourse structure and its rhetoric (Fairclough, 1995). For example, analysis of the discourse structure of scientific documents has confirmed the differences between a normalized and highly structured style of writing in the case of the natural sciences and the more idiosyncratic style of authors in social sciences and humanities (Milas-Bracovic, 1987). Analysis of scientific articles shows that they usually follow a basic structure that contains the purpose of the research, methodology, results, discussion of the results, and conclusions in that order (Pinto Molina, 1995). In a related study, Liddy (1991) claims that a prototypical empirical text has a discourse level structure of seven major segments: subjects, results, purpose, conclusions, methodology, references, and hypothesis.

Discourse and genre analysis has identified some common rhetorical means used by authors in specific text genres (e.g., Swales, 1990; Allen, 1995; Hyland, 2000). Discourse is characterized by special rhetoric, and rhetoric is seen as a standard way of organizing text to achieve certain communicative effects. Rhetoric is often hinted by typical lexical language expressions observable in text and it is assumed that these language expressions can help identify the thematic structures in a text (Allen, 1995). In addition, special rhetorical markers in a text, such as italics, bolds, underlining, and orthographic markers, provide cues for relevant text segments. Of special interest to this dissertation, is the use of reference markers in the rhetoric of scientific discourse (e.g., Small, 1982). Reference markers indicate a specific area in the scientific text where, most likely, a subject matter with relations back to earlier cited documents, are discussed. Chapter 5 introduces the special function of references in scientific text. Further, the potential use of the textual context, surrounding the reference markers, is presented as well.

Several studies have focused on the rhetorical composition of scientific argumentation (e.g., Swales, 1990; Teufel, 1999; Hyland, 2000). For example, Swales (1990) shows how patterns of 'argumentative moves' (prototypical rhetorical expressions) by authors can be used to describe the rhetorical structure of introduction segments of physics articles. Importantly, the 'argumentative moves' describe the rhetorical status of the text segment with respect to the overall content of the document, and not with respect to adjacent text segments, this makes introductions very interesting for indexing purposes, at least in this case. Losee (1996) shows empirically that there are significant differences between the vocabulary, grammar,

and style, used in full text documents in different disciplines and sub-languages. These differences are indicated by where significant content words and phrases appear in the documents. For example, in physics the most significant content words and phrases appear in the beginning of the text (Losee, 1996). Thus, these empirical results support the findings by Swales (1990).

Consequently, from the point of view of indexing, knowledge about formal and discourse structures, as well as the function of specific rhetorical expressions within genres, is valuable, as it may reduce ambiguity and lead to the extraction of semantically more important index terms. Section 3.7 selectively outlines some interesting alternative approaches used for identification of index terms, which are based on knowledge of document structures.

Structural features, such as lexical cohesion, lexical chaining and document structures, show that content words and phrases in a document are not isolated units that appear at random (e.g., Bookstein et al., 2003). Nevertheless, traditional automatic thesaurus construction approaches tend to neglect this feature. They treat the whole document text as an unstructured 'bag of words' and identify potential index terms on the basis of word distributions instead (e.g., Belew, 2000, p. 213). The next section discusses how the distributional properties of natural language are modelled for single term vocabulary construction

3.3 Distribution of words in natural language

To be able to select suitable index terms in automatic indexing, distributional knowledge of different words within individual documents and across all documents in a collection is important (Manning & Schütze, 1999). According to Bookstein et al. (2003, p. 620) "[t]he great mystery of Information Retrieval [...] is that one of the most intellectual of human tasks, making decisions on the basis of a document's content, can be accomplished surprisingly well by completely mechanical means". The explanation for this success is that the process of producing intellectual artefacts, made up of words related to each other semantically in extremely complex ways (lexical chains), incidentally leaves traces that have simple statistical regularities (Bookstein et al., 2003, p. 620).

3.3.1 *The empirical findings of Zipf and Luhn's notion of resolving power*

The number of occurrences of natural language words in text are not uniformly distributed (Egghe & Rousseau, 1990; Wolfram, 2003). Instead, the distribution of words in a text is highly skewed. If plotted as a graph, the distribution will be a reversed *J* shaped curve with a long tail, where the tail expresses the large number of rare events in the distribution (Baayen, 2001). The basic evidence for this phenomenon comes from Zipf (1949). He studied the frequency of different words in a text by use of the type-token ratio. This is the ratio of the number of different words in a text to the total number of words in that text. Zipf (1949) found that the relationship between the frequency of a word and its rank can be represented as:

$$frequency \times rank \cong constant \quad (1)$$

The result of Zipf's (1949) study shows an inverse relationship, where a large majority of unique words were found to occur a few times, while a small minority of a few hundred words had high frequencies of occurrence that accounted for more than 50% of the text (Francis & Kucera, 1982). Zipf's (1949) results are universal because it appears that the highly frequent words in any natural language text are primarily function words.

Luhn (1957; 1958) offers the first theory of how to exploit frequency patterns of words in texts for automatic indexing. Luhn (1957) proposes that the frequency of a word in an article furnishes a useful measurement of the significance of the word in the text. That is, if a word occurs frequently within a text, more frequently than expected, then it reflects emphasis on the part of the author about a topic. Based on Zipf's (1949) findings, Luhn (1957) excludes high frequency function words because they are non-content bearing. Likewise, Luhn (1957) argues that low frequency words in an article are not significant either when it comes to defining the subject matter of the text. Therefore, Luhn (1957) suggests that the most significant words, the ones with most *resolving power*, are the medium-frequency words within a text. Resolving power is defined as a word's ability to represent subject matter. Luhn (1957) assumes that resolving power is maximal at the middle range of a bell curve as illustrated in figure 3.1.

FIGURE 3.1. *Illustration of Zipf's findings and Luhn's notion of resolving power.*

The reversed *J* shaped curve in Figure 3.1 illustrates Zipf's (1949) findings of the inverse relationship between the frequency of words and their rank. The specification of an upper and a lower cut-off level in order to define the range of significant mid-frequency index terms is problematic. While the upper cut-off level may be suggested by Zipf's rank-frequency distribution, the lower cut-off level is more problematic to establish (van Rijsbergen, 1979).

The distributional regularities of natural language put forward by Zipf (1949), and Luhn's (1957; 1958) central finding that repetition is an indication of emphasis, are fundamental to all aspects of automatic indexing. This is the starting point for identification of suitable index terms in a document collection, and the foundation for a number of the classical term weighting schemes introduced later in section 3.5 (van Rijsbergen, 1979; Salton & McGill, 1983; Frakes & Baerza-Yates, 1992; Belew, 2000; Kowalski & Maybury, 2000).

The important distributional regularities of words in a text presented above are, however, of a very general nature. They are based on empirical observations and not on a general probabilistic model of language use (Dunning, 1993, p. 61). To some researchers, the latter is considered an advantage when the object is the processing of natural language (Manning & Schütze, 1999). The assumption is that a probability distribution can capture *a priori* information about the likelihood of word occurrences in a text (Manning & Schütze, 1999). Consequently, the likelihoods or probabilities of word occurrences indicate their distributional behaviour within and between documents. Word distribution models try to characterize how 'informative' a word is. One could cast the problem as one of predicting content from non-content (or function)

words, or ‘good’ or ‘bad’ index terms, though most models have a graded notion of how informative a word is (van Rijsbergen, 1979).

3.3.2 The Poisson distribution

The process of text generation is often viewed as a stochastic process where the text is created by the random selection of words (Manning & Schütze, 1999). The assumption is that words, used by authors to write documents about a selected topic, occur with a certain average frequency in these documents. A stochastic process is modelled by a probability distribution. In IR, the most commonly used probability distribution is the Poisson model (Bookstein & Swanson, 1974; 1975; Harter, 1975a; 1975b; van Rijsbergen, 1979). Two assumptions are needed in order to model the occurrences of a word by a Poisson distribution. The probability of an occurrence of a word in a text is proportional to the length of that text, and the word occurs independently in relation to earlier or later occurrences (Manning & Schütze, 1999). Yet, both these assumptions are violated in this context, see below (Losee, 1998; Manning & Schütze, 1999).

Early experimental research on word distributions established that the statistical behaviour of ‘speciality’ words was different from that of ‘function’ words (Damerau, 1965; Dennis, 1965; Jones & Curtice, 1967; Stone & Rubinoff, 1968; Carroll & Roeloffs, 1969). This research established that ‘function’ words are randomly distributed among documents of a collection, while the ‘specialty’ words are distributed in a non-random fashion.

Later, Bookstein and Swanson (1974; 1975) elaborated on this, and proposed that ‘specialty’ words are content bearing whereas ‘function’ words are non-content bearing. In addition, non-content bearing words are Poisson distributed from document to document of approximately the same size. This means that a word, which is randomly distributed according to a Poisson distribution, is not informative about the document in which it occurs (Bookstein & Swanson, 1974; 1975). At the same time, the fact that a word does not follow a Poisson distribution is assumed to indicate that it conveys information as to what a document is about (Bookstein & Swanson, 1974). For that reason, Bookstein and Swanson (1974; 1975) suggested that content bearing words could be identified by significant deviations from the random distribution. This means that the difference in the distributional behaviour of single words can be used as a guide to whether a word should be assigned as an index term or not. However, in order to model both content and non-content words, a mixture of at least two Poisson distributions is needed (Losee, 1988; Srinivasan, 1990; Margulis, 1992; Church & Gale, 1995).

Harter (1975a; 1975b) proposes the two-Poisson model for IR. In addition to the word discriminating function of the model, the two-Poisson model is also used to measure the degree to which a content word is effective as an index term for a document (Harter, 1975a; 1975b). This is the basis for probabilistic term weighting schemes (e.g., Robertson & Walker, 1994).

As mentioned above, the Poisson model typically violates its two basic assumptions. When modelling events according to the Poisson distribution, uniformity of size is one of the two assumptions. Yet, documents in many collections differ widely in size, and are thus not uniform units of measurement, unless their length is normalized. Further, the tendency of content words to cluster is a problem when the Poisson distribution is used to model the importance of words as index terms for a document. The word independence assumption holds approximately for non-content words, but once a content word occurs in a document it is more likely to occur again. Words appear in a text, not in a 'memoryless' fashion, but following a pattern governed by the thematic progression; words tend to clump and word distributions are usually 'bursty' (Church & Gale, 1995; Katz, 1996; Bookstein, Klein & Raita, 1998; Bookstein & Raita, 2001; Bookstein et al., 2003).

3.3.3 The 'burstiness' of word distributions

Katz (1996) considers the Poisson distribution as an inappropriate model for the distribution of content bearing words due to its incapability to take into account 'topicality' within documents. He states that when a concept, named by content bearing words, is 'topical' for the document, then that content bearing word tends to be characterised by multiple and bursty occurrences. Contrary to most word distribution models, Katz (1996) focuses specifically on the dependency between words within a document. He argues that the observed word occurrences are not independent of each other, and multiple instances of content bearing words are observed close to each other more often than it would be the case if they were an outcome of a Poisson process (Katz, 1996). Thus, Katz (1996) claims that it is desirable not to base probabilistic language modelling on a priori defined specific stochastic mechanisms, such as Poisson processes. Instead, word dependencies in documents should be represented on some observable language characteristics like basic discourse properties pertinent to text formation (Katz, 1996, p. 16).

Katz (1996) also claims that unlike function words, the number of instances of a specific content word is not directly associated with the document length, but is rather a function of how much the document is about the concept expressed by that word. According to Katz (1996), the treatment of concepts in documents introduces new

content bearing words into the discourse, and these new words introduce new related concepts. As a consequence, the number of occurrences per document of individual content bearing words (tokens) grows relatively slowly with the length of a document. But the number of different content bearing words (types) per document grows fast with the length of a document. This weak dependence between document length and the number of occurrences of content bearing words is ignored by many statistical methods.

However, a word or a phrase, which is related to one of the main concepts of a document, may be used throughout the entire document. This means that its occurrence must depend on the length of the document. Katz (1996) argues that in such cases the document length can be viewed, not as a variable that affects the number of occurrences, but as a side effect of continuous and intensive treatment of the concept related to this word or phrase. Likewise, a tendency of longer documents to contain, on average, more instances of a given content bearing word than short documents, is because longer documents tend to discuss a given topic more exhaustively than short documents (Katz, 1996).

Katz (1996, p. 18) introduces the notion of *burstiness* in order to model topicality. Burstiness is used to characterize two closely related but distinct phenomena:

- *document-level burstiness*, that is, multiple occurrence of a content word or phrase in a single document, which is contrasted with the fact that most other documents contain no other instances of this word or phrase at all; and
- *within-document burstiness*, that is, the close proximity of all or some individual instances of a content word or phrase within a document exhibiting multiple occurrence.

Katz (1996) classifies the documents of a collection into three groups according to their topicality with regard to a particular content bearing word or phrase. *Unrelated* documents contain no instances of the particular word or phrase; *non-topical* documents contain a single instance of the word or phrase; and *topical* documents contain more than one instance of the particular content bearing word or phrase. Single occurrence is simply equated with non-topical occurrence, and multiple occurrences with topical occurrence. According to Katz (1996), a ‘within-document burst’ of a given word always indicates that this multiple occurrence of the word is an instance of document-level burstiness as well, but not necessarily vice versa. A given content word or phrase may occur frequently in a document but at long intervals, without any within-document burst. A more bursty word or phrase will be found in

fewer number of documents than a less bursty word, and its multiple occurrence is usually localized narrowly in a document, often being confined to a single burst occupying just a paragraph or two (Katz, 1996).

The G-model derived by Katz (1996) models the distribution of content bearing words and phrases along an extent of documents within a collection as to how likely the word or phrase occurs in a document, how likely it is used topically when it occurs, and how intensively, on average, the word is used when it is used topically. Therefore, the characteristic distribution pattern of topical content words, which contrasts markedly with that of non-topical and non-content words, could provide a useful aid in identifying the semantically relevant words within a text.

Katz's (1996) model is hitherto the most elaborate word distribution model within statistical NLP. It relates to the problem of word dependency, and the model focuses on word behaviour both within and between documents in a collection.

The distributional behaviour of words is the basis for automatic index term selection. More specific, characteristics of the different word distributions, mentioned in this section, are incorporated into some of the weighting schemes discussed in section 3.6 to reflect the relative importance of an index terms.

In the preceding sections 3.1 to 3.3, we have presented the basis and assumptions behind automatic vocabulary construction. We now move on to the actual process of vocabulary construction. The next section describes the preliminary process of lexical analysis necessary before actual indexing.

3.4 Step 1a: Lexical analysis

Preliminary to actual index term selection in automatic indexing, is the process of lexical analysis. Lexical analysis is the process of identification of individual words (tokens) in a text, the removal of function words by use of a stop list, and finally, conflation of morphologically related words through stemming (Fox, 1992; Frakes, 1992). Stop list and stemming procedures are both implemented to improve efficiency and performance of an IR system (Fox, 1992; Frakes, 1992).

The use of a 'stop word' list in automatic indexing address the issue of what words should be allowed into the index. A stop list is simply a list of high frequency words that are eliminated from the representation of both documents and queries. Two motivations are normally given for this strategy. The first motivation is based on the findings of Zipf (1949) and Luhn (1957), that is, high frequency, mostly function words, are seen as carrying little semantic weight and are thus unlikely to help with

retrieval. However, highly frequent content words, in a specific document collection, can also be very poor index terms and thereby become candidates for a stop list (Fox, 1992). Secondly, the elimination of high frequency words can save considerable space in the index files as especially function words make up a large fraction of text (Francis & Kucera, 1982; Frakes, 1992). There is though a major problem with the application of a stop list. Function words have important syntactical and semantic functions in a sentence and its phrases are therefore essential for NLP tasks (Riloff, 1995).

The basic question addressed by stemming is whether the morphological variants of a lexical item should be listed (and counted) separately, or whether they should be collapsed into a single root form (stem) (Frakes, 1992). The premise is that the stem carries the meaning of the concept associated with the token and the affixes (endings) introduce subtle modifications to the concept or are used for syntactical purposes (Frakes, 1992). It is therefore assumed, that tokens with the same stem are semantically related. For example, without stemming, the terms *process*, *processing* and *processed* will be treated as distinct items with separate term frequencies; with stemming they will be conflated to the single term *process* with a single summed frequency count.

The most common stemming algorithms remove suffixes and/or prefixes (e.g., Lovins, 1968; Salton, 1968; Porter, 1980). Stemming can, however, result in over-stemming and under-stemming (Frakes, 1992). The former refers to the case when stemming is too strong, which causes unrelated tokens to be conflated to the same stem. The latter refers to the case when stemming is too weak, which prevents related tokens from being conflated. The application of stemming algorithms reduces the index vocabulary and in effect improves recall rates. However, stemming also cause problems for NLP due to the loss of information needed for aggregate levels of natural language parsing.

Lexical analysis changes the initially derived word distribution. Some words are removed from the vocabulary and some word frequencies are raised or lowered due to stemming. The remaining word stems are represented in a vector space model together with their parent documents. The vector space model is crucial in automatic thesaurus construction. The representation of objects and their attributes in this model enables a range of statistical operations such as recalculation of new word distributions to be used for index term weighting, calculation of term association, vocabulary organization, as well as query and document matching in IR. In addition, the vector space model is also widely used in bibliometric methods such as co-citation analysis, where word stems are replaced by references (Small, 1973). The next section introduces the vector space model.

3.5 Step 1b: The vector space model

In IR, mathematical models are used to represent documents and their content. The most widely used mathematical model is the vector space model (Salton 1968; 1971; Salton, Wong & Yang, 1975; Salton & McGill, 1983). In the vector space model, entities, such as documents and queries, are represented as vectors in a multi-dimensional Euclidean space $\mathbf{R}^n$ (Salton, 1968). The vector space model is based on the notion, that in some rough sense, the meaning of a document is conveyed by the words used (Salton, 1968). If one can represent the words in a document as a vector, then it is possible, by use of linear algebra to compare documents with documents, and documents with queries, in order to determine their 'closeness'. However, by a slight change in representation and interpretation, it is also possible to compare terms with terms for thesaurus construction purposes (see Chapter 4). The following introduction of the vector space model is based on Manning and Schütze (1999).

Each component in a vector represents an index term. There is one value in a vector for every distinct term that occurs in the indexing vocabulary. The indexing vocabulary is equal to the number of different terms selected to index the documents in a collection. In addition, each index term corresponds to a dimension in the vector space. If the component value is binary, it indicates the presence or absence of a term within the document or query. A problem with the use of binary values for terms is that it fails to capture the fact that some terms are more important to the meaning of a document than others are. If the component value is a numerical weight, it indicates the relative importance of a term within the document or query; if a term is absent, the value is zero. Term weighting is treated in section 3.5. An n -dimensional document vector that contains weights can be represented as follows:

$$\vec{d}_j = (w_{1,j}, w_{2,j}, w_{3,j} \dots, w_{n,j})$$

In this notation, $\vec{d}_j$ denotes a particular document j represented as a vector, while the various w components represent the weights for the n -dimensional vocabulary of terms that occur in the collection. This characterization of documents as vectors of term weights allows us to view the document collection as a whole as a matrix of weights, where $w_{i,j}$ represents the weight of term i in document j . Such a matrix is typically called a term-by-document matrix. Under this view, the columns of the matrix represent the documents in the collection, and the rows represent the terms.

Correspondingly, the weights that serve as values for those components serve to locate documents in that space. We denote an $n \times m$ matrix $\mathbf{M}$ with w components as:

$$\mathbf{M} = \begin{pmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \cdots & w_{nn} \end{pmatrix}$$

Where the rows represent index terms, the columns represent documents and w denotes the term weights. In IR, $n \times m$ matrices are sparse as they contain a significant number of zero values. This is the result of tabulating the complete index vocabulary of n -dimensions against individual document vectors, as the latter usually contain only a small set of the index vocabulary¹³.

Consider as an example of the vector space model, the space shown in Figure 3.2, where D denotes documents.

FIGURE 3.2. A simple vector space representation of three documents.

This figure shows a simplified space $\mathbf{R}^3$ that consist of three dimensions corresponding to the terms 1, 2 and 3. In a vector space, some vectors can be characterized as basic vectors. Basic vectors correspond to the dimensions (axes) in the vector space, that is, the terms of the indexing vocabulary. The main assumption is that basic vectors must be orthogonal, which means that they are conceived of as independent of each other.

¹³ It is desirable to attempt to reduce the number of dimension in a vector space and at the same time preserve the salient representations in the matrix. This is called ordination, and it is the focus of several multivariate statistical techniques, see Chapter 4.

If we use the raw term frequency in a document as a weight, then the documents are represented as points in the positive quadrant of a Cartesian rectangular coordinate system. D_1 is represented by the vector (1, 2, 1), D_2 by (6, 0, 1) and D_3 by (0, 5, 1). Each document in the collection is represented as a vector that contains the linear combination (sum) of the basic index term vectors in the space. In a Euclidean space, the value of the vector components becomes coordinates, and the totality of vector coordinates decide the vector placement in the coordinate system. From Figure 3.2 it appears that this space captures certain intuitions about how these documents are related. Document 1, being general, is fairly similar to both document 2 and 3. Conversely, documents 2 and 3, are distant from one another since they cover a different set of topics.

Unfortunately, this instantiation of a vector space places too much emphasis on the absolute values of the various coordinates of each document. For example, what is important about the *term 2* dimension of document 2 is not the value 6 but more so that it is the dominant contributor to the meaning of that document. Similarly, the specific values of 1, 2, and 1 for document 1 are not important, what is important is that the three dimensions have roughly similar weights. It would be sensible, for example, to assume that a new document 4 with weights 3, 6, and 3 would be quite similar to document 1 despite the magnitude differences in the term weights. We can accomplish this effect by normalization of the document vectors. Normalization simply means the conversion of all the vectors to a standard length. Conversion to a unit length can be accomplished by dividing each of their dimensions by the overall length of the vector (the Euclidean norm), which is defined as:

$$\|\vec{d}_j\| = \sqrt{\sum_{i=1}^n w_i^2} \quad (2)$$

This, in effect, eliminates the importance of the exact length of a document's vector in the space, and emphasizes instead the direction of the document vector with respect to origin.

The final step in the vector space model is to determine the closeness between vectors by use of a proximity measure. We analyze the characteristics of proximity measures in Chapter 4. The basic operation that indicates inter-vector closeness is the inner product¹⁴. The inner product of two vectors that have the same dimensionality is denoted as:

¹⁴ The inner product is also known as the scalar product or dot product.

$$\text{sim}(\vec{d}_j, \vec{d}_k) = \vec{d}_j \cdot \vec{d}_k = \sum_{i=1}^n w_{i,j} \cdot w_{i,k} \quad (3)$$

In general, the inner product between two vectors is of no use as a proximity measure, since it is too sensitive to the absolute magnitudes of the various dimensions. However, the inner product between vectors that have been normalized has a useful and intuitive interpretation as it computes the cosine of the angle between two vectors. Note, that if for some reason the vectors are not stored in a normalized form, then the normalization can be incorporated directly into the proximity measure as follows:

$$\text{sim}(\vec{d}_j, \vec{d}_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\|\vec{d}_j\| \|\vec{d}_k\|} = \frac{\sum_{i=1}^n (w_{i,j} \cdot w_{i,k})}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \cdot \sqrt{\sum_{i=1}^n w_{i,k}^2}} \quad (4)$$

In this context, the cosine ranges from 1.0 (cosine 0°) for vectors that point in the same direction, two identical documents, to 0.0 for orthogonal vectors (cosine 90°). We compute how well $\vec{d}_j$ and $\vec{d}_k$ correlate, and then divide by the Euclidian length of the two vectors to scale for the magnitude of the individual $\vec{d}_j$ and $\vec{d}_k$.

The cosine measure of the three document vectors mentioned above in Figure 3.2, results in a symmetric document-by-document matrix $\mathbf{M}$, where columns and row represent document 1, 2 and 3:

$$\mathbf{M} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

Interestingly, if we extend the matrix with the new document 4 (3, 6, 3), which in term profile resembled document 1, then document 1 and 4 end up as identical vectors (1, 0.47, 0.88) due to vector length normalization.

It is convenient to assume that the terms in a vector space are independent, in which case the dimensions are orthogonal. This simplifies the task of computing the closeness between two vectors to that of measuring the angle between them, based on the cosine measure. Obviously, most content bearing index terms that occur in a collection will be highly correlated with other terms (Raghavan & Wong, 1986), but

the assumption of linear independence¹⁵ among variables is commonplace in many real-world applications of statistics, and especially in IR. To address the 'faulty' term independence assumption in IR, co-occurrence information is required (e.g., van Rijsbergen, 1977; Yu et al., 1983). Chapter 4 deals with the principles of term co-occurrence analysis within the vector space model and its use in different automatic thesaurus construction approaches.

The next section presents the generic composition of term weighting schemes. The value derived from term weighting indicates the relative importance of an index term in relation to a specific document. The value is based on the term's distributional behaviour within a specific document and across all other documents in the collection. This is the individual vector-component value represented in the document-term matrix.

3.6 Step 1c: Term weighting

The two most important factors that govern the effectiveness of an index language are the exhaustivity of indexing and the specificity of the index language (Lancaster, 2003). For any document, exhaustivity is defined as the number of different topics indexed, whereas specificity of the index language is its ability to describe topics precisely (Lancaster, 2003, pp. 7-9). Manual and automatic indexing is defined by a tension between exhaustivity and specificity (Lancaster, 2003, p. 30). Unfortunately, these two intuitively reasonable desiderata are in some sense at odds with one another (Belew, 2000). The best explanation of this trade-off is in terms of precision and recall. High recall is best achieved when indexing is exhaustive but not very specific; high precision is best accomplished when indexing is not very exhaustive but is highly specific.

In close connection to this, van Rijsbergen (1979, p. 14) points out that two conflicting ways exist of how to look at the problem of document characterization for retrieval. One way is to characterise a document through a *representation* of its contents, regardless of the way in which other documents in the collection may treat similar content (van Rijsbergen, 1979). This is known as 'representation without discrimination', and the focus is on exhaustive indexing of the potential topics in a document (a recall enhancing function) (van Rijsbergen; 1979). The opposite way, is to insist that in characterising a document, one is *discriminating* it from potentially all

¹⁵ In linear algebra, a set of elements of a vector space is linearly independent if none of the vectors in the set can be written as a linear combination of finitely many other vectors in the set.

other documents in the collection (van Rijsbergen, 1979). This can be labelled ‘discrimination without representation’. Here attention is on the specificity of index terms. They must be specific enough to distinguish between the documents in the collection (a precision enhancing function) (van Rijsbergen, 1979).

Ultimately, the fundamental purpose of an indexing language is to reconcile the many document descriptions in a collection with the many anticipated user queries (Belew, 2000). In practice, one seeks some sort of optimal trade-off between representation and discrimination (Belew, 2000). Traditionally, this has been attempted through balancing the exhaustivity of document indexing against the specificity of the indexing language by use of term weighting schemes (Sparck Jones, 1972; 1973; Salton & Yang, 1973; Salton & Buckley, 1988). Term weighting schemes evaluate the ‘worthiness’ of individual index terms in relation to their dual function of representation and discrimination in the document collection to be indexed.

Index terms belong to the general class of content bearing words in a collection, but index terms are not equally important in relation to automatic indexing. Instead of just a binary indication of the presence or absence of an index term in a document, *term weights* also indicate the relative importance of the term in expressing the contents of a document. In practice, the weighting scheme used to assign term weights in the vectors (document and query) has an enormous impact on the effectiveness of a retrieval system.

An effective term weight is composed of two factors, 1) term frequency within a single document, and 2) the distribution of terms across a collection. The former factor is referred to as a *local* representation factor, and the latter factor is referred to as a *global* discrimination factor (van Rijsbergen, 1979, p. 14). These two factors, the term frequency within a single document, and the distribution of terms across a document collection are dealt with in the following two subsections.

3.6.1 *Term frequency within a single document*

The first factor is called term frequency (*tf*) and is based on the observation of Luhn (1957) that word frequency in a document is a useful measure of word importance, see section 3.3 above. Terms that occur frequently within a document may reflect and represent its content more strongly than terms that occur less frequently, and thus should be given higher weights. This is the principle of lexical cohesion, specifically those of repetition and chaining. In its simplest form, term frequency is simply the raw frequency of a term within a document (Luhn, 1957; Salton & Buckley, 1988). However, there is a fundamental problem with the use of raw term frequencies. Documents vary in length. Longer documents are more verbose than short documents,

as longer documents usually use the same words repeatedly; this is called the verbosity problem (Robertson & Walker, 1994). In addition, longer documents also contain numerous different terms compared to a short document; this is called the scope problem (Robertson & Walker, 1994). The former increases the term frequency of longer documents, and the latter increases the probability of a match between longer documents and a query (Singhal, Buckley & Mitra, 1996). As a result of both cases, longer documents have an advantage over short documents in retrieval (Belew, 2000). What is preferable is a relative term frequency weight that enables comparison between documents of varying length (Salton & Buckley, 1988). Several length normalization functions have been applied. The most popular are the cosine function (Salton & Buckley, 1988), the augmented normalized term frequency function (Salton & Buckley, 1988), and the pivoted length normalization (Singhal, Buckley & Mitra, 1996; Singhal et al., 1996). The cosine and the augmented normalized functions tend to over-penalize longer documents (Singhal, Buckley & Mitra, 1996; Singhal et al., 1996). Conversely, in accordance with Katz (1996), Singhal, Buckley and Mitra (1996) show that longer documents in general are more likely to be relevant to topics than short documents. Singhal, Buckley and Mitra (1996) approach the problem of length normalization by doing a post hoc analysis of the distribution of retrieved versus relevant documents as a function of their length. The fact that these two distributions cross suggests a corpus-specific length normalization pivot-value, below which match scores are reduced, and above which they are increased. Thus, the generic pivoted length normalization function makes the normalization factor weaker or stronger by reducing the deviation in the retrieval probabilities from the likelihood of relevance (Singhal, Buckley & Mitra, 1996; Singhal et al., 1996).

3.6.2 The distribution of terms across a document collection

The second factor to consider is the distribution of terms across the collection as a whole. Terms that occur in a few documents are useful for discriminating those few documents from the rest of the collection. Conversely, terms that occur frequently across the entire collection are less useful in discriminating among documents. Therefore, what is needed is a measure that favours terms that occur in fewer documents. The fraction N/n_i , where N is the total number of documents in the collection, and n_i is the number of documents in which term i occurs, provides exactly this measure. The fewer documents a term occurs in, the higher this weight. The lowest weight of 1 is assigned to terms that occur in all documents. Due to the large number of documents in many collections, this measure is usually squashed with a log

function leaving us with the following *inverse document frequency* (*idf*) term weight (Sparck Jones, 1972):

$$idf_i = \log \left(\frac{N}{n_i} \right) \quad (5)$$

When the term frequency factor is combined with the *idf* factor, it results in a generic scheme known as $tf_{i,j} \times idf_i$ (Salton & Yang, 1973, p. 354). That is, the weight of term i in the vector for document j is the product of its overall frequency in j with the log of its inverse document frequency in the collection. With some minor variations, this weighting scheme is used to assign term weights to documents in nearly all vector space retrieval models (Salton & Buckley, 1988).

While term frequencies in individual documents are static, the inverse document frequency weight is dynamic. The *idf* is collection dependent and changes when the document collection changes over time. Because the primary focus of *idf* is on global discrimination among documents, local document frequencies of the same index term become insignificant. This makes *idf* very sensitive to the definition of how document boundaries are defined (Belew, 2000, p. 90).

In this section, we have introduced the three basic components of a term weight, the term frequency (*tf*), the term importance (*idf*), and document length normalization. There are several ways to compute the three components. For a survey and evaluation of different applications of the $tf \times idf$ scheme, we refer to Salton and Buckley (1988). Although there is no single optimal weighting scheme, some heuristics have been established (Salton & Buckley, 1988). Weighting schemes that use collection frequency generally perform better, and if documents vary considerably in length, the use of document normalization is imperative (Salton & Buckley, 1988).

3.6.3 Term discrimination value

The term discrimination value is comparable with the inverse document frequency weight and may replace the latter in a $tf \times idf$ weighting scheme, such as $tf \times DISCVALUE_i$ (Salton, Yang & Yu, 1975). The term discrimination model assumes that the most useful terms for content identification of natural language texts are those capable of discriminating the document vectors in the vector space from each other (Salton & Yang, 1973; Salton, Yang & Yu, 1975; Salton & McGill, 1983). A measure for the average similarity between vectors is compared to the individual terms. Good discriminators are those terms that decrease the average similarity by their presence.

Poor discriminators increase the average similarity, while neutral discriminators have no effect on average similarity. This means that terms that are positive discriminators can be included in the vocabulary and the rest rejected. Salton and Yang (1973) show that terms with high frequency have a negative discrimination value, terms with low frequency a near zero discrimination value, and only intermediate frequency terms a significantly positive value. This supports the findings of Luhn (1957). One major problem with the term discrimination value is that for a document collection of any size the calculation may be expensive (Belew, 2000). Salton, Yang and Yu (1975) have provided a reasonable alternative; they suggest the use of document frequency df_i as an approximation to the discrimination value. Interestingly, there is a strong correlation between Poisson distributions, document frequencies, and discrimination values (Srinivasan, 1990). This means that in many cases the document frequency can be used, as it is the most easy to compute (Srinivasan, 1990).

The weighting schemes presented above are *ad hoc* with respect to their mathematical validity (Kageura & Umino, 1996). The weighting schemes take an empirical or pragmatic standpoint regarding the 'meaning' of a weight. This can be inferred from the existence of variations based on the same idea, for example, taking or not taking the log or square or applying division or subtraction to the same components. In that sense, they are essentially the application of statistical or quantitative methods but not of statistical models (Kageura & Umino, 1996). This problem can also be observed at the stage of evaluation of the selected index terms. The index terms are evaluated either by subjective comparison with human made lists of index terms or by retrieval performance in experimental IR contexts. In either case, what is examined and evaluated is the result obtained by the application of the weighting scheme, and not the theoretical validity of the scheme or underlying model itself. These criticisms are, of course, not fair since these schemes are effective in practice (e.g., Baeza-Yates, 1999; Belew, 2000; Kowalski & Maybury, 2000). As long as the schemes are situated in the context of IR applications, the ultimate purpose of automatic indexing is the practical improvement of IR performance. Further, the schemes have to pay attention to the applicability of their methods within practical IR environments, which necessarily imposes a certain restriction on their approach. Nevertheless, for further improvement of indexing, totally effect-oriented evaluation has its own limitations. At that point, the problem of the theoretical validity of the statistical model, with respect to language in general and index terms in particular, has to be addressed (Dunning, 1993).

3.7 Alternative approaches for identification of index terms

In section 3.2, we pointed out that document structures influence the location of the content in a document. In IR, it has become accepted that somehow document text processing must be more narrowly focused in order to overcome the problems of the ‘bag of words’ approach (e.g., Al-Hawamdeh & Willet, 1989; Salton & Buckley, 1991; Callan, 1994; Hearst & Plaunt, 1993; Lalmas & Ruthven, 1998). This can, for example, entail the segmentation of documents into smaller semantic sub-units. These subunits can be retrieved individually, or they can bring about retrieval of their parent documents (Salton & Buckley, 1991; Belew, 2000). The purpose is to treat the document as an entity where some parts are more important than others for indexing and retrieval purposes, and not as a ‘bag of word’, where the constituent word order is ignored (Belew, 2000; Moens, 2000). The use of document structures for indexing and retrieval is not new. Online bibliographic retrieval systems have for decades allowed field specific searches for document retrieval. However, the dissemination of markup languages has accentuated the interest in automatically derived methods for more structured indexing and retrieval (e.g., Bishop, 1999).

From the point of view of indexing and retrieval, there are several good reasons why a full text document should be decomposed into minor subunits based on formal, discourse or thematic structures. For example, minor text units are relatively short and local, thereby embodying focused context for the treated concepts (Belew, 2000; Moens, 2000). Likewise, content words tend to cluster in local areas of a document, thus locality is one way to identify and extract semantically coherent words from texts (Bookstein, Klein & Raita, 1998). Another important feature of minor text units is that decomposition can be used as a tool for document length normalization (Belew, 2000).

Knowledge about formal and discourse structures in documents, and their signalling lexical and linguistic phenomena, can help to identify the thematic structures and thereby to help selecting words from the text that are more reflective of its content than a simple ‘bag of words’ (e.g., Hahn, 1990; Lewis & Sparck Jones, 1996). For example, index term selection and weighting can be determined by the formal structural position of the term within the text, for example title, abstract, first paragraph etc. (e.g., Bernstein & Williamson, 1984; Jonák, 1984; Liddy & Myaeng, 1993; Wilkinson, 1994; Al-Halimi & Tompa, 2003). This idea can be traced back to work on automatic abstracting, where it was recognized that the first or the last sentence of a paragraph, and sentences at the beginning or the end of a document text, are usually the most central to the theme of the text (Luhn, 1957; 1958; Baxendale, 1958). From an indexing perspective, Dennis (1967) determined the importance of a

word based upon its frequency of occurrence within a text paragraph and across preceding and succeeding paragraphs. Of special interest to this dissertation, and discussed in Chapter 5, is the use of citation contexts in documents as the target for index term selection (Small, 1978; Rees-Potter, 1989). Index terms selected from citation contexts refer back to previously cited documents. Reference markers indicate citation contexts within the documents, though the boundaries are fluent. O'Connor (1982; 1983) has developed several algorithms that can detect citation contexts of varying length within scientific documents, and subsequently extract single index terms from these contexts. Similarly, Web page indexing often uses the anchor text (citation context) from links in source pages to represent the referenced page (e.g., Lawrence, Giles & Bollacker, 1999). There is also a lot of research into decomposition of texts according to different thematic structures, which can be useful for identifying important topic terms in text (Salton & Buckley, 1991; Hearst & Plaunt, 1993; Hearst, 1997; Singhal, Mitra & Buckley, 1997). This is known as passage retrieval, where the document is divided into uniform-sized passages that are indexed individually (Knaus et al., 1995; Zobel et al., 1995; Kaszkiel & Zobel, 1997), and locality based retrieval where the passage boundaries are more dynamic (Kretser & Moffat, 1999). The types of passages explored by researchers can be grouped into passages based upon formal structural units, passages based upon the subject or content of the text, or passages based upon a fixed number of words in so-called windows. For example, the *TextTiling* algorithm searches for parts of a text where the vocabulary shifts from one subtopic to another in order to detect the subtopics of a text (Hearst & Plaunt, 1993; Hearst, 1997). The algorithm is based upon the assumption that the main topics of an expository text occur throughout the text, and the subtopics only have a limited extent in the text.

Bookstein, Klein and Raita (1998) revisit the random non-random distributional aspect of words in the context of full text documents in their notion of term clumping. Term clumping is based on the Poisson model (Bookstein & Swanson, 1974; 1975) and is related to the ideas of 'burstiness' and 'topicality' suggested by Katz (1996). Bookstein, Klein and Raita (1998) find that the occurrences of an important word in a full text document tend to clump in restricted regions. The assumption behind term clumping is that if a word conveys meaning in a semantic sense, then its occurrences will be associated with textual units related to that meaning. For such a word, occurrences will not be random, but rather will tend to occur in clumps, just as content terms does, and this tendency produces sequential patterns that can be used to distinguish units about that word from units not about the word (Bookstein, Klein & Raita, 1998). Thus, important content bearing words tend to appear largely in text

segments about the subject matter of that word. So, while it is not possible to detect meaning directly, it is possible to exploit statistical correlates of meaning (Bookstein, Klein, & Raita, 1998; Bookstein & Raita, 2001; Bookstein et al., 2003). Moving through a document sequentially, one notices how subject matter often shifts. The longer the discussion of a topic in a document, the larger the segment that contains the topical words used, and the more important the topic with respect to the document. Minor topics may be discussed in smaller segments of text (Bookstein, Klein & Raita, 1998; Bookstein & Raita, 2001). Strongly clumped words have indexing and retrieval value, and if text is segmented to minimize clumping strength, such stretches of text are likely to be content homogeneous. Experiments by Bookstein, Klein and Raita (1998) show that the ‘quality’ of terms selected for indexing, by use of term clumping measures, improves when compared to, for example, the inverse document frequency. Several different term clumping measures have been developed and tested to detect the unevenness of distributions as a way of predicting the significant words within a text (Bookstein, Klein & Raita, 1998; Bookstein & Raita, 2001). In addition, Bookstein et al. (2003) extends the notion of term clumping to construct measures of semantic term association. This is further discussed in Chapter 4.

Kim and Wilbur (2001) investigate the ability of three different statistical methods, including the Poisson model and term clumping measures, for identification of content bearing words in titles and abstracts of the MEDLINE® database. The two structural elements chosen, titles and abstracts are traditionally conceived of as good indicators of document content, at least in the journal articles within the natural and life sciences (e.g., Resnick, 1961). However, the prevalent automatic indexing methods usually avoid the individual use of abstracts, or at least titles, as they do not contain enough text to exhibit a term’s distributional behaviour needed for such indexing methods. Nevertheless, Kim and Wilbur (2001) demonstrate that it is possible to identify important content bearing words from titles and abstracts in a domain specific database like MEDLINE® by use of statistical methods, based on the same distributional parameters as used in full text indexing. In addition, Losee (1996) shows that ‘theoretical’ disciplines have lower type-token ratios, with words being re-used more often than in ‘experimental’ disciplines with higher type-token ratios. The ‘theoretical’ disciplines, with more word re-use, produce less ‘rich’ document segments such as abstracts. This results in less overlap between the abstracts and the full text. What Losee (1996) calls ‘experimental’ disciplines have fewer word tokens per type of words, essentially this makes the abstracts semantically ‘richer’. Thus, words from abstracts have a larger intersection with the terms in full text. Ries and

colleagues (2001) confirm this finding in a study, where the distribution of words in abstracts and full texts of medical journals are compared.

The final aspect we consider in this section does not refer directly to document structure, it focuses on the use of exemplary documents as 'containers' of candidate index terms (Blair & Kimbrough, 2002). As described in Chapter 2, the use of exemplary documents for identification of index terms is very common in manual thesaurus construction. Conversely, the use of exemplary documents in automatic indexing is not widespread. Blair and Kimbrough (2002) suggest that exemplary documents are well worth considering for automatic indexing as well. They find it inexpedient to base the vocabulary in a full text collection on all the index terms found in the documents. Instead, they argue that it is well worth to consider the application of fewer, but perhaps better, index terms, selected from some canonical literature (Blair & Kimbrough, 2002). According to Blair and Kimbrough (2002), exemplary documents are those documents that describe or exhibit the intellectual structure of a particular field of interest. In so doing, they provide both an indexing vocabulary for that area and, more importantly, a narrative context in which the indexing terms have a clearer meaning. Many document collections then can be divided into two kinds of documents: the documents to which access is desired, and a much smaller set of exemplary documents that may be used to build an intellectual structure on which some or all of that access can be based. According to Blair and Kimbrough (2002), it may be possible to exploit these exemplary documents as an organizing principle for a collection of documents.

3.8 Summary

This chapter focuses on the assumptions behind, and procedures of, the construction of a vocabulary. This process is also known as automatic indexing. Vocabulary construction is the first of three steps in automatic thesaurus construction. The other two succeeding steps are term association and vocabulary organization. They are treated in the next chapter.

The chapter draws attention to the important problems in automatic indexing of how to identify and represent words (phrases are treated in Chapter 4) that reflect the content of a natural language text. In general, automatic indexing has dealt with this problem by removing instances of a few hundred common words and treating the remaining words as though they were more or less content bearing. This approach

relies upon simple assumptions about the distribution patterns of individual words in full text documents. Such existing methods are originally developed to index heterogeneous document collections, which explain the rather shallow approach where almost all words are considered index terms. Yet, the simple ‘bag of words’ approach is still the most cost-effective approach to automatic indexing of full text documents. Hitherto, alternative indexing features have not significantly improved IR performance. Nevertheless, indexing is more than cost-effective IR full text performance. From a LIS perspective, it is also the assignment of more ‘costly’ quality controlled index terms. As Kim and Wilbur (2001) point out, when index terms are to be examined by a human, as a means of accessing the literature, it greatly improves efficiency if most of the unimportant words and phrases can be eliminated from the indexing. Consequently, the aim of semi-automatic thesaurus construction is not necessarily the same as automatic thesaurus construction despite the methods and techniques are practically the same. Semi-automatic thesaurus construction primarily focuses on identification of fewer, but more meaningful, index terms. It is our belief, that the use of statistical natural language processing can help identify important index terms beyond those identified in a ‘bag of words’. But, it is necessary to localize statistical NLP to specific domains and to combine it with knowledge of document structures.

This chapter has presented some of the characteristics of natural language. We have emphasized the inappropriateness of single term indexing, and pointed to the noun phrase as the basic unit for indexing. We have also indicated that document structures can be used as a tool to identify good index terms. This has motivated us to investigate the possibility of using some of these features for semi-automatic thesaurus construction. The basic indexing unit in our semi-automatic approach is the noun phrase. The target for identification of noun phrases in the text structure is citation contexts. Identified noun phrases somehow refer back to the cited document pointed to in the citation context. As outlined in this chapter, it is a fair assumption to believe that noun phrases identified in a localized segment of a domain specific document is likely to be semantically related. This domain specific and localized focus for index term extraction creates a homogenous corpus of approximately same length and content. It is also assumed that these characteristics will lead to identification of significant domain specific terms with less ambiguity. In addition, our approach resembles that of Blair and Kimbrough (2002) where ‘exemplary documents’ are used as the basis for index vocabulary construction. However, our ‘exemplary documents’ are highly cited documents if they act as an agreed upon concept symbol to a majority of users in a scientific domain.

The next chapter presents the two remaining steps of automatic thesaurus construction, term association and vocabulary organization.

4. Term association and vocabulary organization

As outlined in Chapter 3, the process of automatic thesaurus construction can be divided into three generic steps. Whereas Chapter 3 treats step 1 that concerns vocabulary construction, Chapter 4 concerns the remaining two steps of term association and vocabulary organization:

2. **Term associations:**

The aim is two-fold:

- Identification of term associations by use of statistical co-occurrence analysis and/or linguistically based syntactical-statistical co-occurrence analysis;
- Statistical or syntactical phrase identification, phrases are weighted and used as index terms (optionally), and can be applied without thesaurus construction);

3. **Vocabulary organization:**

The aim is to structure the thesaurus terms by means of cluster analysis or other related statistical ordination techniques.

The two steps of term association and vocabulary organization extend automatic indexing to automatic thesaurus construction.

The purpose of the chapter is to present the assumptions behind and procedures of term association and vocabulary organization. They are generic to other applications, where textual objects are compared and clustered according to their co-occurring attributes. In addition to presenting the steps of automatic thesaurus construction, the chapter also functions as the basis for related bibliometric methods to be presented in Chapter 5, and is furthermore the foundation of the proposed semi-automatic approach put forward in Chapter 6.

Phrase indexing is a special case. Phrases can play a major role in automatic thesaurus construction, although their use is not restricted to thesaurus construction. Phrase indexing could be treated as a separate step in automatic indexing. But automatic phrase identification can also be seen as a form of term association. We have decided to treat phrase indexing in connection with the more general aspects of term associations. An important feature of term associations and vocabulary

organization is that of proximity measures. Proximity measures are the statistical interpretation of association between two co-occurring objects. Proximity measures are therefore also extensively used in some bibliometric methods, such as co-citation analysis (Small, 1973). The basic characteristics of proximity measures apply regardless of its application. Nevertheless, the application and use of these measures is not always apparent. Consequently, we present and discuss the composition, characteristics and meaning of proximity measures at some length in section 4.3, in between step 2 and step 3. The steps presented in Chapter 3 and the present chapter are generic steps of automatic indexing and automatic thesaurus construction. The actual research on automatic thesaurus construction can be divided into two main approaches: the statistical approach and the syntactic-statistical approach. The chapter closes with a presentation of the most significant results produced by these approaches.

Section 4.1 presents the notion of term association and its fulfilment through co-occurrence analysis. Section 4.2 introduces statistical and syntactical phrase identification in natural language text. Section 4.3 discusses the composition, characteristics and meaning of proximity measures. Section 4.4 outlines the basic procedures of cluster analysis used for vocabulary organization. Section 4.5 reviews approaches to automatic statistical thesaurus construction, and section 4.6 does the same to syntactic-statistical approaches. Finally, section 4.7 summaries the main points of the chapter.

4.1 Step 2: Term associations

The common basis for automatic construction of word associations is a specific text corpus (Dagan, 2000). The assumption is that the 'meaning' of words is related to their patterns of co-occurrence with other words in the text. This assumption was proposed in early linguistic work. Harris (1968), for example, expressed it in his 'distributional hypothesis': "... the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities" (Harris, 1968, p. 12). Another expression of this assumption is the famous statement by Firth (1957, p. 11): "... you shall know a word by the company it keeps".

In automatic thesaurus construction, index term associations are identified through co-occurrence analysis. Researchers understand the phenomenon of co-occurrence differently (Manning & Schütze, 1999). Some use the terms collocation and co-occurrence to be able to differentiate between different types of co-occurrences

(Manning & Schütze, 1999). Manning and Schütze (1999) define collocation as grammatically bound items occurring in a certain order that are characterised by limited compositionality. They admit the existence of word associations across larger expanses of text, but they suggest calling such associations ‘co-occurrences’ and to reserve the term ‘collocation’ only for grammatically bound combinations (Manning & Schütze, 1999). In addition, Halliday and Hasan (1976) point out that collocation is a realisation of lexical cohesion in text. They argue that words collocate because they are in some kind of lexical-semantic relation. Accordingly, two linguistic factors cause word co-occurrence, lexical-grammatical restrictions and lexical-semantic relations (Vechtomova, Robertson & Jones, 2003). Further, these linguistic factors reflect the distance between words in text. We simplify matters and use co-occurrence to cover both short and long-span collocation types between words.

As noted above, two lexical categories of word co-occurrences exist. The first category comprises grammatical relations, which refer to the co-occurrence of words within restricted lexical-grammatical relations. The second category comprises non-grammatical relations, which refer to the co-occurrence of words within a certain distance within a text. This is often referred to as a window of text (Dagan, 2000). These two types of co-occurrence are outlined below.

4.1.1 Co-occurrence based on grammatical relations

Restricted lexical-grammatical relations operate within very short distances in sentences of natural language text (Jurafsky & Martin, 2000). Lexical-grammatical restrictions limit the choice of words that can be used within the same grammatical structure with a particular word in question (Dagan, 2000). Thus, the meaning of a word usually restricts the identity of other words that can co-occur with it within specific syntactic relationships (Dagan, 2000). Two examples of short distance co-occurrences, motivated by lexical-grammatical constructions, are syntactic phrase expressions, such as noun phrases and predicate-argument relations such as subject-verb and object-verb (e.g., Hindle, 1990; Ruge, 1992; Jurafsky & Martin, 2000). In order to extract grammatically based lexical relations, it is necessary to use a syntactic parser, see section 4.2. Parsed short distance grammatical relations serve two primary functions in automatic indexing, syntactic phrase identification and syntactical-statistical thesaurus construction. The set of co-occurrences for a word within its syntactic relations can provide a reflection of its semantic properties; hence, grammatical relations are of interest to thesaurus construction (Ruge, 1992; Grefenstette, 1994).

4.1.2 Co-occurrence of words within a text window

Lexical-semantic relations exist between semantically related words that are used to describe the same topic within a text (Halliday & Hasan, 1976; Boguraev & Neff, 2000). These relations are not confined to the same grammatical structure, but usually span over longer distances in text. Lexical-semantic relations are usually domain specific (Moens, 2000). Several sub-types of non-grammatical co-occurrences exist, such as *n*-grams (Jelinek, Mercer & Roukos, 1992), symmetric and non-symmetric co-occurrence within small or larger windows, extending to whole documents. The latter equals the maximum window size. For example, lexical-semantic co-occurrence between two words within a relatively large window in a text suggests that both words are related to the general topic discussed in the text (Dagan, 2000). Conversely, lexical-semantic relations within smaller defined text windows are more likely to capture context words that identify different topics of discourse in a text (Gale, Church & Yarowsky, 1993). This hypothesis will usually hold for frequent co-occurrences, that is, for pairs of words that often co-occur in the same text window.

Distance based co-occurrences are easier to compute than grammatical ones as no parsing is involved. Large window sized co-occurrence analysis is used to identify term dependencies, term associations and statistical thesaurus construction (Sparck Jones, 1971; van Rijsbergen, 1977). Small window sized co-occurrence analysis are used to generate multiple-word index terms. In contrast to syntactic phrases, a statistical phrase is a statistical surrogate of a genuine linguistic phrase (e.g., Fagan 1989, Croft, Turtle & Lewis, 1991; Smadja, 1993).

4.1.3 First and second order co-occurrences

Above we have characterized word co-occurrences in a corpus as either dependent on lexical-grammatical restriction in a very narrow context, or because of lexical-semantic relations within a broader context defined by a text window. However, word co-occurrences or word associations can also be categorized as first or second order co-occurrences, or direct or indirect co-occurrences respectively (Soergel, 1974; Matsumoto, 2000).

First order term co-occurrence analysis measures how often term *X* and term *Y* tend to occur together in a text window (Matsumoto, 2000). Second order co-occurrence analysis associate terms with similar context (Matsumoto, 2000). Suppose that term *D* almost never occurs without term *W*, and that term *T* also tends not to occur without term *W*, yet *D* and *T* never co-occur in documents. One concludes that some relationship exist between *D* and *T*, that is, they are related by the fact that each one co-occurs strongly with *W*. In all probability, term *D* and term *T* are synonymous or

near synonymous in this context. Synonyms tend not to occur with each other, yet the terms they co-occur with will be very similar (Lancaster, 2003). The assumption behind first order co-occurrence analysis is that semantically *related* terms tend to appear in some predefined context. Whereas the notion behind second order co-occurrence analysis is, that semantically *similar* words have a tendency to share similar contexts (Matsumoto, 2000). This entails, that a first order co-occurrence analysis can generate an association profile for a specific term (Soergel, 1974). This profile contains other terms that frequently co-occur with the specifically chosen term. Subsequently, two different term association profiles can be compared through second order co-occurrence analysis. If they have similar association profiles, we can expect the two terms to be in a definitional relationship (Soergel, 1974).

Given the distributional hypothesis and the notions of lexical-grammatical and lexical-semantic relations, we can expect that words that resemble each other in their ‘meaning’ will have similar co-occurrence patterns with other words. To capture this ‘similarity’, each word can be represented as a co-occurrence vector, which represents the statistics of its co-occurrence with all other words in the corpus. The association of two words is then computed by applying some proximity measure to the two corresponding co-occurrence vectors. Section 4.3 will follow up on the characteristics of the proximity measures used to identify significant term associations. The following section, section 4.2, introduces the special case of term association that concerns the identification and extraction of phrases for automatic indexing. We present two approaches, one based on statistically defined small text windows, and one based on lexical-grammatical collocation.

4.2 Step 2: Identification of phrases

Whereas Chapter 3 focused on single words as index terms, this section introduces multiple-words as index terms and how they are identified in natural language text. In automatic indexing, multiple-words refer to statistically or syntactically identified phrases. The rationale for the use of phrases as index terms is that they represent more meaningful concepts than individual words (e.g., Moens, 2000). Especially noun phrases are believed to be content bearing units and thus good indicators of a text’s content (Earl, 1970; Salton, Buckley & Smith, 1990; Smeaton, 1992). It is conceived that phrases, in a combination with single terms, may constitute a more meaningful representation of contents in documents (Zhai et al., 1997). This rationale is based on

some common facts about phrases. Phrases are less ambiguous in meaning compared to the single words they are composed of. A phrase can be considered as a specification of a concept (Anick & Vaithyanathan, 1997; Moens, 2000). It may signify an important concept in certain subject domains (e.g., Justeson & Katz, 1995; Anick & Vaithyanathan, 1997). For example, the term 'joint venture' is an important term in financial texts, while neither 'joint' nor 'venture' are important by themselves (Strzalkowski et al., 1999). Phrases, even ones made up of high frequency words, may only occur in a few documents, thus becoming good discriminators (Salton & McGill, 1983). Phrases improve the specificity of the indexing language, and hence theoretically increase precision in retrieval operations (Fagan, 1989).

Despite extra computational requirements for their recognition (Callan & Lewis, 1994), phrases are prime index term candidates and should be included in a document representation (Earl, 1970; Salton, Buckley, & Smith, 1990; Smeaton, 1992). When phrases are employed as index terms, three aspects are essential: 1) automatic phrase identification, 2) normalization to a standard phrase form, and 3) statistical weighting of the phrase (Moens, 2000). Two approaches exist to phrase identification in automatic indexing. One is the statistical co-occurrence approach that identifies phrases within a small text window, and the other is the lexical-grammatical approach that identifies phrases from the syntactical structure of sentences, likewise in small text windows. Yet, in practice hybrid methods, which combine syntactic and statistical approaches, are common in both automatic indexing and NLP (Smadja, 1993; Grefenstette, 1994).

4.2.1 Statistical phrase identification

Simple statistical phrase identification assumes that when a set of words, usually two, co-occur in small-defined text windows, at a rate greater than would be expected by random chance, then they may denote a phrase (e.g., Salton, Buckley & Smith, 1990; Damerau, 1993). The idea of using statistical associations between words as an indicator of phrases goes back to the early 1960s (e.g., Edmundson & Wyllys, 1961; Giuliano, 1965; Spiegel & Bennett, 1965). Typically, pairs of adjacent, morphologically normalized, non-stop words are considered as candidate phrases. A statistical phrase is then defined by constraints upon the frequency of occurrence of the phrase, upon the occurrence and co-occurrence frequency of its components, and/or upon the proximity of its components in a text window (Salton, Buckley & Smith, 1990; Croft, Turtle & Lewis, 1991).

Indexing phrases needs to be persistent expressions (Church & Hanks, 1990). A persistent expression, for example, could be a phrase like *solar system* (Godby, 2002,

p. 5). In statistical terms, this entails that when *solar* and *system* occur in a collection of specialized text, they usually occur together. Only rarely is one seen without the other. However, a noun phrase such as *additional activities* is not persistent because *additional* combines with hundreds of other words, forming phrases such as *additional assessment*, *additional candidates* etc. (Godby, 2002, p. 5). The narrower the text window, the more likely it is that the syntagmatic combination (co-occurrence) is in fact a linguistic phrase (Dagan, 2000).

Statistical phrase identification methods are based on proximity measures such as association measures or probabilistic measures; see section 4.3 for a thorough definition. The association measures use corpus frequency as input (e.g., Salton & McGill, 1983; Choueka, 1988; Smadja, 1993; Justeson & Katz, 1995), and the probability measures use a training corpus to derive estimates for ‘correct’ phrase constructions (Church & Hanks, 1990; Damerau, 1993; Dunning, 1993). Further, the probabilistic measures are asymmetric; this means that they are able to depict the ‘correct’ composition of a phrase.

Occurrence frequency and proximity parameters do not always yield correct and meaningful phrases. Statistical phrases can be ambiguous and they often do not reveal the correct composition of a phrase. Two or more terms possibly co-occur for reasons other than being part of the same phrasal concept. Finally, it is very difficult to identify phrases of more than two words by use of statistical co-occurrence analysis. It is therefore not surprising, that Fagan (1989) found that the use of statistical phrases did not significantly increase retrieval performance.

4.2.2 Syntactical phrase identification

A syntactic phrase is a grammatical part of a sentence that is identified by linguistic criteria embedded in a natural language parser. Parsing refers to grammatically rule-based algorithms used for syntactical analyses of a text to determine its structure (Schwarz, 1990; Smeaton & Sheridan, 1991). For automatic indexing purposes, it is sufficient to identify noun phrases, as well as, lexical-grammatical relations for co-occurrence analysis. Syntactic phrase identification can be accomplished through a restricted syntactical analysis called *shallow parsing*. Shallow parsing traditionally consists of part-of-speech tagging and syntactic analysis based upon the output of the part-of-speech tagging (Sheridan & Smeaton, 1992; Strzalkowski, 1995; Zhai et al., 1997). Words in sentences are tagged by their part-of-speech (word class). Subsequently, shallow parsers use the part-of-speech tagging to identify the major constituents of the sentences, i.e., noun phrases and verb phrases. In addition, shallow parsing may also detect the dependency relations within phrases. Dependency

relations are of interest to automatic thesaurus construction. In this respect, dependency relations refer to the predicate-argument structure in a sentence, such as verbs and their arguments, as well as, head-modifier relations within a phrasal expression (Sheridan & Smeaton, 1992; Strzalkowski, 1995).

Figure 4.1 illustrates the general principles of shallow parsing with focus on application in automatic indexing.

FIGURE 4.1. The process of shallow parsing.

The result of the parsed sentence is three noun phrases: *the periodontal ligament*, *the tooth*, and *the alveolar bone*. From an indexing perspective, shallow parsing identifies phrases with a low degree of ambiguity. In an indexing situation, the preposition *the* will be excluded, thus *periodontal ligament*, *tooth*, and *alveolar bone* remains as candidate index terms.

Shallow parsing algorithms operate with low error rates and parsed syntactical phrases are very precise (Moens, 2000). In addition, syntactic analysis can identify phrases even when the words of which they are composed of do not co-occur with greater than chance frequency. However, extraction of phrases by purely syntactic means alone is seldom effective as it is likely to extract many phrases of little value for characterizing the topics of a given document (Croft, Turtle & Lewis, 1991). Most phrases, often 85% or more, occur only once in a corpus and can be safely eliminated. It is therefore necessary to identify persistent expressions. A syntactic phrase is typically selected on the basis of its occurrence frequency, the co-occurrence of its components, and document frequency (Salton & McGill, 1983; Croft, Turtle & Lewis, 1991; Strzalkowski, 1995; Strzalkowski et al., 1999).

4.2.3 *Normalization and weighting of phrases*

Selected phrases, whether statistical or syntactical, still need to be subjected to phrase normalization and subsequent phrase weighting before they can be incorporated into a vocabulary. Part of the discouraging effect of the use of phrases in IR arises when phrases are not normalized to a standard form (Moens, 2000). Normalization refers to the fact that phrasal concepts can be expressed by use of different syntactic structures (e.g., *garden party* and *party in the garden*), possibly combined with lexical variations in word use and morphological variants (Moens, 2000). It is necessary to map these different phrase expressions to a single form. Of possible methods used are a machine-readable dictionary (Evans et al., 1991); to omit function words and neglect the order of the remaining content words (Dillon & Gray, 1983; Fagan, 1989); or to recognize syntactic variants (Sparck Jones & Tait, 1984; Tzoukerman, Klavans & Jacquemin, 1997; Jacquemin, 1999).

Phrase weighting serves the same function as in the case of single term weighting outlined in Chapter 3. Phrase weighting identify good index phrases based on the notion of representation and discrimination (van Rijsbergen, 1979). Because of their lower frequency and different distribution characteristics it is, nevertheless, important not to weight phrases by use of traditional weighting schemes applied for single terms (Strzalkowski et al., 1997). Several different approaches have been suggested for phrase weighting (Lewis et al., 1996); either they treat the phrase as a separate concept (Dillon & Gray, 1983; Croft, Turtle & Lewis, 1991; Strzalkowski, 1994), or as a set of words (Salton, Yang & Yu, 1975; Fagan, 1989; Croft, Turtle & Lewis, 1991; Evans et al., 1991). In addition, Jones, Gassie and Radhakrishnan (1990) employ a combined approach and weigh phrases proportional to the frequency of occurrence of the complete phrase and to the frequency of occurrence of its composing words.

Automatic phrase indexing has until now shown only a slight improvement in retrieval performance (e.g., Fagan, 1989; Mitra et al., 1997, Zhai et al. 1997). Some of these results indicate that other features than the phrases themselves cause this rather meagre improvement. For example, Fagan (1989) showed that results varied between text collections. Likewise, Fagan (1989) and Strzalkowski (1995) have shown that syntactically derived phrases, if applied on a sufficient scale, out-perform simple statistically derived co-occurrences. In addition, Strzalkowski et al. (1997) point out that the retrieval performance of phrases is improved, if the weighting scheme is adapted to the special characteristics of phrases.

Phrases are a vital part of indexing and thesauri. Identification of two-word phrases by use of statistical co-occurrence analysis is computationally inexpensive, yet the results are typically ambiguous. Conversely, syntactically derived phrases can consist of more than two words and such phrases are often very precise. In this dissertation, we apply shallow syntactic parsing to identify domain specific multi-word index terms from citation contexts; this application is described in Chapter 8. The present section has illustrated how short-span word associations in the form of phrases are identified. The following section treats the characteristics behind proximity measured use to identify significant long-span term associations.

4.3 Proximity measures

Once the appropriate thesaurus vocabulary has been identified, perhaps including phrases, the next step is to determine the association and relations between pairs of index terms in the vocabulary. In automatic thesaurus construction, this is done by use of statistical proximity measures¹⁶. To measure the degree of association between two objects by use of proximity measures is a central issue, not only in IR, but also in bibliometrics. Whether we measure term associations in a thesaurus vocabulary or co-citation networks between documents, the proximity measures are the same. We discuss the composition, characteristics, and meaning of proximity measures on a general level that refer to a wide range of applications in IR and bibliometrics.

¹⁶ Some terminological confusion exists in relation to the measurement of association. For practical reasons, we use the term *proximity* to cover *distance* (dissimilarity) measures, as well as *similarity* measures.

The convention in IR and bibliometrics is to represent objects¹⁷ as vectors in an $n \times m$ matrix, where the components of the vector correspond to the attribute-values¹⁸ (Tanimoto, 1958; Salton, 1968). The $n \times m$ vector-component matrix is the basis for computation of inter-vector proximities. Vectors (objects) are compared through co-occurrence analysis of their components (attributes). The vector components contain either binary values or non-binary values, in the form of weights or frequencies. This process generates an $n \times n$ proximity matrix, where n is the objects compared.

To measure the degree of proximity between any two objects involves three steps. Firstly, select the attributes of the objects whose values are to be used for comparison. Secondly (and optionally), a weighting scheme may be implemented that emphasizes certain attributes according to any difference in their relative significance. We have outlined the use of weighting schemes in automatic indexing in Chapter 3. Finally, a value for a measure that represents the degree of proximity between the objects can be derived from an analysis of their various attribute-values (Ellis, Furner-Hines & Willett, 1994). In this context, we may classify different ‘associations’ or co-occurrences according to the objects whose vector representations are compared, for example: document-query association, document-document association, term-term association, bibliographic coupling and co-citation analysis (Ellis, Furner-Hines & Willett, 1994).

No clear-cut criterion for choice among proximity measures has been derived hitherto in IR or bibliometrics (Jones & Curtice, 1967; McGill, Koll & Noreault, 1979; Jones & Furnas, 1987; Leydesdorff, 1987; Oberski, 1988; Wang, Wong & Yao, 1992). Some measures are preferred for different purposes, such as correlation measures in author co-citation analysis (e.g. McCain, 1990; White, 2003a), but the choice of measure is often bound in heuristics rather than validity (Ahlgren, Jarneving, & Rousseau, 2003).

Several proximity measures are referred to in the literature by various names, or are variously represented by different forms of what may be revealed on close inspection to be the same formula (Anderberg, 1973). Likewise, many measures are monotonic with each other. Two measures may be said to be monotonic if the ranking of all measurements of proximity between pairs of objects in a specific set is the same using one measure as it is using the other (Anderberg, 1973). What proximity measures should then be selected? It seems that the history of research into the use of proximity measures in IR and bibliometrics appears to betray a lack of progress,

¹⁷ Object refers to the *thing* investigated and perhaps clustered; other names that denote an object are, for example, case and entity.

¹⁸ Component, variable, attribute, character and feature denote those aspects of the *things* used to assess their proximity.

excluding the renewed debate in bibliometrics on author co-citation analysis (Ellis, Furner-Hines & Willett, 1994; Ahlgren, Jarneving & Rousseau, 2003; White, 2003a; Leydesdorff, 2004). There are, nonetheless, a number of significant aspects that concern the composition, class and meaning of proximity measures, which should be considered before selecting a measure. As these aspects are often not considered, we present some of them in this section (Ellis, Furner-Hines & Willett, 1994).

4.3.1 The composition of proximity measures

All common proximity measures can be defined in a binary form, and most of these can be transformed into a non-binary variant (Anderberg, 1973). We focus on the non-binary variants as this is case in automatic thesaurus construction. A proximity measure consists of one or two components. In the case of two components, the numerator contains a basic vector comparison function and the denominator contains vector functions applied as normalization factors. From the vectors $\vec{k}$ and $\vec{j}$, we can derive two basic vector functions: the *difference sum* (6) and the *inner product* (3):

$$\sum_{i=1}^n |x_{ik} - x_{ij}| \quad (6)$$

The difference sum is the sum of the absolute differences between the corresponding components in the two vectors.

$$\sum_{i=1}^n x_{ik} \cdot x_{ij} \quad (3)^{19}$$

The inner product is the sum of the products of corresponding components. Linearity is the key feature to these functions (Sneath & Sokal, 1973). Therefore, the inner product is a measure of the colinearity of two vectors.

It is possible for either of these two vector functions to be used on their own as a crude measure of the proximity between two vectors. Because of algebraic linearity, their values, however, vary in direct proportion with the value of n , the number of pairs of components that are compared (Anderberg, 1973). Moreover, if the component values under consideration are in non-binary form, these functions' values also vary in accordance with the scale of the component-values.

¹⁹ The inner product (3) is briefly introduced in Chapter 3.

Vector functions have no upper limit if they are not normalized. Normalization ensures that the values of the measures remain within a specific range, such as that bounded by 0 and 1, or by -1 and +1. The characteristic by which most proximity measures may be distinguished is the composition of the factor by which they require the difference sum or the inner product to be multiplied.

A normalization factor within a proximity measure can be composed of different vector functions. Functions for non-binary data that are commonly used in the composition of normalization factors are:

$$\sum_{i=1}^n x_{ik} \text{ and } \sum_{i=1}^n x_{ij} \quad (7)$$

This is the sum of all components in vector $\vec{k}$ and vector $\vec{j}$.

$$\sum_{i=1}^n x_{ik}^2 \text{ and } \sum_{i=1}^n x_{ij}^2 \quad (8)$$

The second normalization factor is the sum of the squares of all components in vector $\vec{k}$ and vector $\vec{j}$.

$$\sqrt{\sum_{i=1}^n x_{ik}^2} \text{ and } \sqrt{\sum_{i=1}^n x_{ij}^2} \quad (2)^{20}$$

The final normalization factor is the Euclidean lengths of vector $\vec{k}$ and vector $\vec{j}$ in the n -dimensional space.

Common for these three factors is that they normalize individual vectors to a standard length by reducing the number of zero-valued components (Jardine & Sibson, 1971). These vector functions are the building blocks in most proximity measures used in IR and bibliometrics.

4.3.2 *Classes of proximity measures*

Sneath and Sokal (1973) describe four classes of proximity measures: 1) distance measures, 2) association measures, 3) correlation measures, and finally 4) the radically different probabilistic similarity measures. The latter are measures based on

²⁰ The Euclidean length (2) is briefly introduced in Chapter 3.

conditional probabilities that take into account the frequency distribution of components over the set of objects to derive probability distributions (Sneath & Sokal, 1973). The notion is that agreement among rare components is a less probable event than agreement for frequent components and should therefore be given higher weights. Asymmetric proximity values are common, for example, object *A* may stand in certain relation to object *B*, but *B* may not have that degree of relation to *A*. Yet, we focus on distance, association and correlation measures, they are the most common and they follow the abovementioned composition.

Table 4.1 illustrates the most commonly used proximity measures in IR and bibliometrics.

TABLE 4.1. *Proximity measures.*

ID	Common name	Formula	Range
<i>D</i>	Minkowski	$\left(\sum x_{ik} - x_{ij} ^r\right)^{1/r}$	∞ to 0 (9)
<i>D</i>	Manhattan	$\sum x_{ik} - x_{ij} $	∞ to 0 (10)
<i>D</i>	Euclidean	$\sqrt{\sum x_{ik} - x_{ij} ^2}$	∞ to 0 (11)
<i>A</i>	Jaccard	$\frac{\sum (x_{ik} \cdot x_{ij})}{\sum (x_{ik})^2 + \sum (x_{ij})^2 - \sum (x_{ik} \cdot x_{ij})}$	0 to 1 (12)
<i>A</i>	Cosine	$\frac{\sum (x_{ik} \cdot x_{ij})}{\sqrt{\sum (x_{ik})^2 \cdot \sum (x_{ij})^2}}$	0 to 1 (4) ²¹
<i>C</i>	Pearson	$\frac{\sum (x_{ik} - \bar{x}_k) \cdot (x_{ij} - \bar{x}_j)}{\sqrt{\sum (x_{ik} - \bar{x}_k)^2 \cdot \sum (x_{ij} - \bar{x}_j)^2}}$	- 1 to + 1 (13)

ID indicates if it is a distance (*D*), association (*A*) or correlation (*C*) measure. *Range* indicates the range of proximity values. The number to the right in the *Range* column indicates 'highest' proximity value, in the case of distance measures that is zero. From Table 4.1, it is apparent that the proximity measures consist of different combinations of the two basic vector functions and different normalizations factors.

Distance measures are based on the difference sum²² and high degree of 'similarity' is indicated by lower values (Anderberg, 1973). Distance measures

²¹ The cosine measure (4) is briefly introduced in Chapter 3.

²² Distance measures are also known as *dissimilarity measures*.

normally have no upper bounds and are scale-dependent, as they do not include normalization factors (Anderberg, 1973). Hence, two objects are identical if each one is described by components with the same magnitudes. The values of distance measures can be equated with measurements of the geometric distances between pairs of vectors in an n -dimensional Euclidean space (Anderberg, 1973). Consequently, geometric theorems can be applied to calculate geometric distances between vectors in Euclidean space. For this validity to be preserved, distance measure should have the properties of metrics (Anderberg, 1973)²³. The most popular distance measures are specific forms of the Minkowski metric (9) (Sneath & Sokal, 1973). Both the Manhattan (10) and Euclidean (11) distance measures are a derivation of the Minkowski metric. Raghavan & Wong (1986) have questioned the validity of the geometric interpretation of the meaning of distance, which is widely accepted in IR. The major problem with this interpretation is the assumption of independent dimensions, an assumption that does not hold true in IR, as outlined in Chapter 3.

Association measures are based on the inner product. Association measures indicate similarity in the range between 0 and 1, 1 indicating complete similarity. An example is the *Jaccard* measure (12) (Jaccard, 1901; Tanimoto, 1958; Doyle, 1962). In its non-binary variant, Jaccard is often referred to as the extended Jaccard or Tanimoto. The extended Jaccard measure uses component sets from two vectors to calculate similarity. The measure considers the cardinality ratio of the number of shared components of vector $\vec{k}$ and vector $\vec{j}$ (intersection) to the number possessed by vector $\vec{k}$ or vector $\vec{j}$ (union). The Jaccard measure is invariant to scale and does not take into account the sizes of vectors in that only non-zero values are treated (Sneath & Sokal, 1973). Many other association measures have been suggested that use different weighting schemes in the composition of the normalizing factor (Ellis, Furner-Hines & Willett, 1994).

A special class of association measures is that of angular measures (Anderberg, 1973). Angular measures also have a geometric interpretation, not as a function of distances between vectors, but as a function of angles between vectors. The best known example is the cosine measure (4) (Driver & Kroeber, 1932; Ochiai, 1957; Salton, 1963). We introduced the cosine measure in Chapter 3. The validity of its basis in Euclidean geometry has been questioned in a similar way to that in which the

²³ A **metric** is a nonnegative function $d(x,y)$ that describe the ‘distance’ between neighbouring objects for a given set. A metric satisfies the *triangle inequality*: $d(x,y) + d(y,z) \geq d(x,z)$ and is *symmetric*, so $d(x,y) = d(y,x)$. A metric also satisfies $d(x,x) = 0$ as well as the condition that $d(x,y) = 0$, which implies that $x = y$. If the latter condition is dropped, then $d(x,y)$ is called a pseudometric instead of a metric. Functions that do not satisfy the axiom of triangle inequality are non-metric.

appropriateness to applications in IR of measures based on geometric distance has been questioned (Raghavan & Wong, 1986).

A special class of angular measures is that of correlation coefficients. Correlation coefficients are based on a third, more complex vector function: the sum of the product of the differences between each component-value and the mean of all component-values for each of the two vectors (Sneath & Sokal, 1973). The difference between each component value and the mean of all component-values in a vector is called moment²⁴. The best known correlation coefficient is the Pearson product-moment correlation coefficient (13) (Sneath & Sokal, 1973). As this measure is based on moments around the mean, it is sensitive to zero component-values. The problem with zero-values can, however, be reduced by the application of a logarithmic transformation to the data (Ahlgren, Jarneving & Rousseau, 2003; Leydesdorff, 2004). Correlation coefficients were originally developed in order to measure the degree of correlation between the attributes of a sample of independent objects taken from a population, so-called R-type studies (Sneath & Sokal, 1973). Their values could then be subjected to test that would determine the statistical significance of such a correlation. Discussion of the use of correlation coefficients in IR often points to this amenability to statistical analysis (Maron & Kuhns, 1960). However, the validity of correlation coefficients for the measurement of similarity between vectors that represent objects, so-called Q-type studies, can be questioned (Jardine & Sibson, 1971). When coefficients are used in R-type studies, it is assumed that the objects are distributed randomly and independently of each other. It is usually not the case in Q-type studies in IR and bibliometrics that attributes are similarly distributed randomly and independently. Likewise, the significance of the mean of a set of values for different attributes is unclear, especially if the attributes of an object are each measured on an individual rather than a standardized scale (Jardine & Sibson, 1971). Consequently, as in the case with zero values, one should apply a logarithmic transformation to the data in order to make tests of significance (Ellis, Furner-Hines & Willett, 1994; Ahlgren, Jarneving & Rousseau, 2003; Leydesdorff, 2004). Nevertheless, it should be noted that the product-moment correlation coefficient has been used with rather less uncertainty as to its validity in a wide range of applications in IR and bibliometrics (Ellis, Furner-Hines & Willett, 1994; White, 2003a).

²⁴ Moments represent specific types of summary statistics of a distribution such as mean, variance, or skewness (Wolfram, 2003, p. 86).

4.3.3 *Linearity of proximity measures*

A measure may be said to be linear if there exist a normalizing function by which each vector may be multiplied such that the value of the inner product of the normalized vectors is equal to the value of the measure under consideration (Ellis, Furner-Hines & Willett, 1994). The cosine measure, for instance, is formed by multiplying the inner product of a pair of vectors by the inverse of the product of their Euclidean lengths. It is hence equal to the inner product of the corresponding pair of vectors that may be derived by normalizing the original pair so that their Euclidean lengths each equal 1 (Ellis, Furner-Hines & Willett, 1994). In this context, we may note the similarity in form of the Pearson coefficient to the cosine measure (Anderberg, 1973). The numerator of the Pearson coefficient takes the form of the inner product of a pair of vectors, *viz.* those whose components are given by subtracting from each of the original vector's components the mean of those components. Without further normalization, this function is known as the covariance of a pair of vectors (Anderberg, 1973). Just as the cosine measure may be formed by dividing the inner product of a pair of vectors by the product of their Euclidean lengths, the Pearson coefficient is formed by dividing the inner product of another pair of vectors, equal to the original vectors normalized by moment, by the product of those vectors' Euclidean lengths (Anderberg, 1973; Ellis, Furner-Hines & Willett, 1994). The essential difference between the two measures is that the cosine is based on the original scores (deviations from the origin) while the correlation coefficient is based on centered scores (deviations about the mean). Thus, correlation coefficients can be considered as a cosine between normalized vectors (Anderberg, 1973). Conversely, the Jaccard measure is not a linear measure (Hubálek, 1982). This is, in fact the probability of a vector having both components *A* and *B* conditional on it having at least one of the components (Kuhns, 1965). This measure does not take into account zero vector component-values. This makes Jaccard a non-linear measure based on set theory instead (van Rijsbergen, 1979).

4.3.4 *The geometric meaning of proximity measures*

Several studies have attempted to explain the difference in output of different proximity measures by analyzing more precisely their geometric meaning (Raghavan & Wong, 1986; Jones & Furnas, 1987; Wang, Wong & Yao, 1992; Rorvig, 1999).

Jones and Furnas (1987) assert that the behaviour of a proximity measure depends on its *semantic sensitivity* to the information of various kinds that is contained in a set of (document) vectors. The two most important of these kinds of information are *topic* and *intensity* (Jones & Furnas, 1987). Topic refers to the content of each document,

indicated by the set of relationships between each individual term weight and others in the same document vector (Jones & Furnas, 1987). Intensity refers to the 'intensity'²⁵ of the content' of each document, indicated by the set of relationships between each individual term weight and other weights for the same term in different document vectors (Jones & Furnas, 1987). The direction of a vector in vector space may be said to represent the topic of a document or query, while its Euclidean length may be said to represent the object's intensity. There is no reason why this interpretation cannot be extended to other types of vectors, such as document vectors that contain references.

Jones and Furnas (1987) identify several important properties of proximity measures in their analysis of the differences amongst measures in their sensitivity to topic and intensity, for example:

- Given two vectors of equal length, the one with the smaller angle from a third particular vector is rated by a measure as the more similar, then we may say that the measure has the property of *angular monotonicity*. All measures that are linear have this property, albeit to different degrees. In the case of the cosine measure, for example, angle monotonicity does not even depend on the two vectors being of equal length, because vectors with the same direction but different lengths are transformed by the cosine formula to the same vector of unit length (Jones & Furnas, 1987).
- Given two vectors of the same angle from a third vector, the longer one is rated by a measure as the more similar, then we may say that the measure has the property of *radial monotonicity*. The inner product is radial-monotone, but the cosine and Pearson measures are not. The latter two measures are thus completely dominated by a *sensitivity* to topic, and *intensity* is ignored (Jones & Furnas, 1987).
- Given two vectors, if an increase in the value of any component in one vector, to a value greater than that of the corresponding component in the other vector, has the effect only of increasing the degree of that vector's similarity with the other vector, or has no effect at all, then we may say that the measure has the property of *component-wise monotonicity*. If a measure is not component-monotone, just as the cosine and Pearson coefficients are not, then it might be argued that it can penalize documents for their 'richness'. An example of this is discussed in Chapter 3, where it is shown that the cosine measure favours shorter documents and penalizes longer ones (Jones & Furnas, 1987).

²⁵ Jones and Furnas (1987) suggest that 'intensity' might be interpreted variously as 'quantity', 'quality', or 'accessibility'.

- If it is possible that an increase in the value of any single component in a document vector can have the effect of increasing the degree of that vector's similarity with a another document vector to an arbitrarily high level, even if the document has very little to do with the other document's topic (and hence its vector's angle of separation from the query vector is very large), then we may say that the measure is subject to unbounded single-component influence. The inner product is subject to this influence, whereas the cosine and Pearson coefficients are not (Jones & Furnas, 1987).

We have outlined some important characteristics of proximity measures in this section. Yet the choice of a proximity measure is largely subjective and often based on tradition or on *a posteriori* criteria such as 'interpretability' of the results (Ellis, Furner-Hines & Willett, 1994). We agree with Kruskal (1964) that each scientific area that has use for different measures of proximity should after appropriate argument and trial, settle down on those measures most useful for its needs. Nevertheless, the conditions mentioned in this section should be considered before choosing the most appropriate measures. The condition that concerns semantic sensitivity should be considered before choosing an appropriate proximity measure. For example, in IR and bibliometrics data matrices are often sparse. Consequently, one has to consider how the larger number of zero component-values will influence the semantic sensitivity of a given measure. Obviously, the final choice depends on what properties of association one wishes to emphasize. Hubálek (1982) suggests that empirical studies would do well to select a linear and a non-linear measure and compare their results. Such results may not necessarily be different; for example, it has been shown that the cosine and Jaccard measures can be monotonic to each other in their rankings of objects (Braam, Moed & van Raan, 1988; Hamers et al., 1989). We follow the suggestion of Hubálek (1982), and apply linear and non-linear measures to be able to compare co-citation results of symmetric proximity measures. The results are presented in Chapter 8.

When the proximity measure is chosen and term associations have been calculated, terms in the vocabulary must be organized or structured. Section 4.4 outlines the technique of cluster analysis, traditionally used for vocabulary organization in automatic thesaurus construction approaches.

4.4 Step 3: Vocabulary organization

In Chapter 1, we stated that classification and indexing is based on classing or clustering of objects (terms or documents) based on similarities of characteristics (Anderson & Pérez-Carballo, 2001a; 2001b). In addition, we pointed out that the term 'clustering' usually implies an automatic process that generates *a posteriori* clusters of objects. This section presents the basic properties of cluster analysis. Above we outlined how proximity measures can be used to construct a proximity matrix that describe the strength of all pairwise relationships among a set of objects (documents, terms, references, citations). This transformation is done by use of the vector space model. In automatic thesaurus construction, it is often preferable to organize the pairwise relationships between index terms. This is the vital step of vocabulary organization (Srinivasan, 1992), and it is traditionally done by use of cluster analysis in combination with graph theoretical interpretations (Soergel, 1974). In this respect, graph theory is a part of cluster analysis, or *vice versa* (Soergel, 1974; Buckley & Lewinter, 2003). Graph theory is a more fine grained structural interpretation of the results of a cluster analysis (Buckley & Lewinter, 2003). Cluster analysis is also an essential part of bibliometrics, and thus, a very important aspect to the present dissertation.

We depict the basic properties of cluster analysis in this section. However, this section is not a thorough introduction to cluster analysis, for this we refer to the general works of Jardine and Sibson (1971), Anderberg (1973), Sneath and Sokal (1973), Aldenderfer and Blashfield (1984), and Everitt (1998). For specific works on cluster analysis in IR, we refer to Sparck Jones (1971), van Rijsbergen (1979), Willett (1988), and Rasmussen (1992), and finally for bibliometrics, Egghe and Rousseau (1990).

4.4.1 *The characteristics of cluster analysis*

Cluster analysis encompasses a range of statistical techniques used to generate a category structure that fits a set of observations (Anderberg, 1973; Rasmussen, 1992). Cluster analysis groups a set of objects into clusters based on a pairwise comparison of the objects using a proximity measure (Anderberg, 1973). Consequently, the $n \times n$ proximity matrix is an appropriate basis for cluster analysis. The clusters formed should have a high degree of association between members of the same cluster and low degree between members of different clusters (Anderberg, 1973). Cluster analysis is an unsupervised process, which means that the generated clusters are not predefined

prior to processing, but are defined by the objects assigned to them (Everitt, 1998). Because there is no need for the clusters to be identified prior to processing, cluster analysis is a useful tool to provide structure in large multivariate data sets, either as a stand-alone tool to get insight into data, or as a pre-processing step for other techniques (Everitt, 1998). Consequently, cluster analysis has been described as a tool of ‘discovery’ because it has the potential to reveal previously undetected relationships based on complex data (Anderberg, 1973; Rasmussen, 1992). Likewise, clustering can be seen as dimensionality reduction techniques, related to Latent Semantic Indexing, Multidimensional Scaling (MDS), and Factor Analysis.

The ability of cluster analysis to categorize objects automatically into groups of association, gives it a natural affinity with the aims of IR and bibliometrics (Willett, 1988; Rasmussen, 1992). Cluster analysis can be performed on documents in several ways:

- Documents can be clustered based on the terms they contain. The aim of this approach has usually been to provide more efficient or more effective retrieval, though it has also been used after retrieval to provide structure to large sets of retrieved documents (e.g., Salton, 1968; Jardine & van Rijsbergen, 1971; van Rijsbergen, 1979; Willett, 1988).
- Documents can be clustered based on co-occurring references or citations, in order to provide insight into the nature of the literature of one or more domains, see Chapter 5 (Kessler, 1963; Small, 1973; Egghe & Rousseau, 1990; Gmür, 2003).
- Terms can be clustered based on the documents in which they co-occur, in order to aid in the construction of a thesaurus or in the enhancement of queries (Sparck Jones, 1971; Salton, 1972).

Sparck Jones (1970) has classified different types of clustering methods in term of some general relational characteristics. Overall, there is the relation between attributes and clusters. This relation is either monothetic, where cluster membership is based upon possession of a specific attribute, or the relation is polythetic, where cluster membership is based upon possession of a sufficient fraction of the attributes that define the cluster. Likewise, there is a relation between objects and clusters. This relation defines whether cluster membership is exclusive or overlapping. Finally, there is the relation between clusters and clusters. Relations between clusters are either non-hierarchic (simple partition) or hierarchic. The major focus in IR has been on methods that produce polythetic clustering of documents, where each cluster is defined by a set

of words and phrases and cluster membership is determined through a threshold value (van Rijsbergen, 1979; Kowalski & Maybury, 2000).

4.4.2 Clustering methods and algorithms

The non-hierarchical methods divide the data set of n objects into k clusters these are known as partitioning methods (Rasmussen, 1992). Each object has a membership in the cluster with which it is most associated, and the cluster is represented by a 'centroid' or 'cluster representative' that is indicative of the characteristics of the objects it contains (Kowalski & Maybury, 2000). This approach has been applied in both document clustering and automatic thesaurus construction, most notably in the SMART project (Salton & McGill, 1983).

The more complex hierarchical methods cluster a set of objects either by a process of division or agglomeration (Anderberg, 1973). Agglomerative hierarchical methods have been the focus in IR and bibliometrics (Voorhees, 1986; Willett, 1988). Agglomerative hierarchical methods produce a nested data set in which pairs of objects are successively joined until every object in the data set is connected (Anderberg, 1973). All objects begin alone as singletons in clusters of size one. Clusters are subsequently joined into new clusters a total of $n - 1$ times, at which point all objects in the set will be in one cluster.

For a given cluster method, there is a choice of different clustering algorithms to implement the method (Sneath & Sokal, 1973). Different agglomerative algorithms vary in the procedure that defines the most 'similar' pair of clusters. While there may appear to be an overwhelming abundance of hierarchical clustering algorithms treated in the literature, all the algorithms seem to be alternative formulations or minor variants of three major clustering concepts: linkage algorithms related to graph theory, centroid based algorithms, and algebraic constructs, such as error of sum of squares or variance algorithms (Aldenderfer & Blashfield, 1984). We focus on two classical linkage algorithms, that of simple and complete linkage. They are they most commonly used in automatic thesaurus construction and bibliometrics.

In single linkage clustering²⁶, an object that is a candidate for an extant cluster has similarity to that cluster equal to its similarity to the closest member within the cluster. Thus, connections between objects and clusters and between two clusters are established by single links between pairs of objects (Sneath & Sokal, 1973). Single link algorithms can be implemented relatively efficiently and have been widely used in IR and bibliometrics (van Rijsbergen, 1979; Willett, 1988; Egghe & Rousseau, 1990;

²⁶ This algorithm is also known as nearest neighbour or minimum method.

Belew, 2000). However, as only a single link is required to merge two clusters, single link algorithms have a tendency toward formation of long straggly clusters. Clusters tend to be prematurely coalesced through a chain of 'intermediate' objects. This tendency to form loosely bound clusters with little internal cohesion is called 'chaining'. Hence, single link algorithms are suitable for delineating 'ellipsoidal' clusters but unsuitable for isolating 'spherical' or poorly separated cluster (Anderberg, 1973; Sneath & Sokal, 1973). In graph theory, single link clusters are known as 'connected component sub-graphs' of a set of objects, where the objects are nodes, and the nodes are connected by links that represent proximities greater than some threshold (van Rijsbergen, 1979). Single link algorithms have been extensively used in index term and document clustering in IR (e.g., Sparck Jones & Jackson, 1970; Auguston & Minker, 1970; Sparck Jones, 1971; Salton, 1975; Salton & McGill, 1983).

Complete linkage clustering is the direct antithesis of the single linkage algorithm. An object that is a candidate for admission to an extant cluster has similarity to that cluster equal to its similarity to the farthest member within the cluster. When two clusters join, their similarity is that existing between the farthest pair of members, one in each cluster (Sneath & Sokal, 1973). This method will generally lead to tight, hyperspherical, discrete clusters that join others only with difficulty. It is called complete link because *all* objects in a cluster are linked to one another within some maximum distance. The strong clustering criterion that pertains to the complete link algorithms can result in a number of singletons that do not cluster at some threshold. Consequently, complete linkage may sometimes impose a structure on the data rather than uncover the actual structure present (Anderberg, 1973). In graph theory, complete link clusters are known as 'maximal complete sub-graphs', where each node is connected to every other node in the sub-graph and the set is maximal with respect to this property, i.e. if one node were included anywhere, the 'completeness' condition would be violated. (van Rijsbergen, 1979).

Based in graph theory, Sparck Jones (1971, p. 56) has identified four prototypical cluster types, see Figure 4.2. Individual clusters are depicted as sub-graphs and their internal proximity connections define their generic type. They are characterized by an increasing sophistication, so that type 1 classes are simpler, and type 4 more complex (Sparck Jones, 1971).

FIGURE 4.2. *Prototypical cluster types (Sparck Jones, 1971, p .56).*

The rationale behind the use of graph theoretical methods is that they depict relational structure. Whereas a cluster in it self only indicate that objects are associated, sub-graphs may indicate how these objects are related. This is of interest of automatic thesaurus construction (Sparck Jones, 1971; Soergel, 1974; Salton & McGill, 1983). According to Sparck Jones (1971, p. 56), the internal structure of single link clusters can be characterized as sub-graph types 1 to 3, ‘strings’, ‘stars’ or ‘clumps’, depending on clustering criterion, proximity data and threshold values. In addition, single link clustering also closely corresponds to a weighted graph’s minimum spanning tree²⁷ (van Rijsbergen, 1979). The internal structure of complete link clusters are typically of type 4 classes, ‘cliques’, likewise dependent on clustering criterion, proximity data and threshold values (Sparck Jones, 1971, p. 56).

Many different interpretations of clustering methods and algorithms have been used for document and term clustering in IR (Willett, 1988; Rasmussen, 1992; Kowalski & Maybury, 2000). Traditionally, the basis for automatic construction of a thesaurus is single link clustering of terms based on the documents in which they co-occur (e.g. Salton & McGill, 1983). Of interest to this dissertation is that the complete link algorithm produces ‘clique-like’ clusters that have the strongest relationships between all of the words in the cluster (Kowlaski & Maybury, 2000). This suggests that the cluster is more likely to describe a particular concept (Kowlaski & Maybury, 2000, p. 150). Further, complete link algorithms provide the highest precision when the statistical thesaurus is used for query expansion. The single link algorithm maximizes recall but can cause selection of many non-relevant objects.

Chapter 3 and section 4.1 to 4.4 in the present Chapter has introduced the generic steps of automatic thesaurus construction: vocabulary construction, term association

²⁷ The minimum spanning tree (MST) is the tree of minimum length that connect objects, where ‘length’ denotes the sum of the weights of the connecting links in the tree. Given the MST, then the single link clusters are obtained by deleting links from the MST in order of decreasing length; the connected sets after each deletion are the single link clusters. The order of deletion and the structure of the MST ensure that the clusters will be nested into a hierarchy (van Rijsbergen, 1979).

and vocabulary organization. The next two sections review recent research done within the two main approaches to automatic thesaurus construction, that is, the statistical approach, section 4.5, and the syntactical-statistical approach in section 4.6.

4.5 Automatic statistical thesaurus construction

The basic purpose of automatically constructed thesauri is to improve retrieval performance by the substitution of an appropriate cluster of terms for one of its members. This process is typically named as *automatic query expansion* (Efthimiadis, 1996). Many different approaches to query expansion methods have been studied in IR (e.g., Efthimiadis, 1996). A key element in query expansion is the source that will provide the terms for the expansion (Efthimiadis, 1996). In this respect, automatic thesaurus construction is of two types, the common *global* corpus type, and the more rare *local* type (Baeza-Yates & Ribeiro-Neto, 1999). These types differ in the choice of sources for query expansion. Global types use *all* documents in a corpus, independent of any search results, to determine a global thesaurus-like structure that defines term relationships (Baeza-Yates & Ribeiro-Neto, 1999). To expand a query, local types identify term relationships from an initially retrieved local set of documents (Attar & Fraenkel, 1977; Doszkocs, 1978). Our focus is on automatic thesaurus construction based on a global corpus analysis.

Automatic statistical thesaurus construction derives from early research on term associations, clustering and the vector space model (Salton, 1971).

Early research established that term associations had significance for retrieval. Term-term associations can create a network of associations among terms and can be used to expand an initial list of search terms (Maron & Kuhns, 1960; Stiles, 1961; Doyle, 1961; 1962; Giuliano & Jones, 1963; Dennis, 1965; Giuliano, 1965; Jones & Curtice, 1967; Lesk, 1969; Sailsbury & Stiles, 1969). In addition, term-term associations can also be used to form discrete term clusters based on the documents in which they co-occur (Sparck Jones & Needham, 1968; Sparck Jones & Jackson, 1968; Sparck Jones, 1971; Sparck Jones & Barber, 1971; Minker, Wilson & Zimmerman, 1972; Salton, 1972; van Rijsbergen, 1977).

Until the beginning of 1990s, the global corpus type of thesauri, where a query was augmented by use of term clusters, failed to yield consistent improvements in retrieval performance with heterogeneous collections (e.g., Sparck Jones, 1970; Smeaton & van Rijsbergen, 1983; Willett, 1988; Peat & Willett, 1991).

Elkalifa (1991) suggests that the failure of these early thesaurus construction approaches is mainly due to the heterogeneity of the collections used rather than to the inefficiency of the methods themselves. Peat and Willett (1991) analyze the repeated failure of automatically constructed thesauri built from term-term proximity matrices. They note that a key problem with the use of term co-occurrence for thesaurus construction is that relatively frequent terms co-occur with other frequent terms (Peat & Willett, 1991). The result is a thesaurus in which one relatively general term is found to be related to another general term. Although these terms are related, they do not improve precision and recall because, due to their relatively high frequency, they are not good discriminators (Peat & Willett, 1991). In addition, Smeaton and van Rijsbergen (1983) had previously shown that randomly selected terms for expansion was sometimes more effective than the use of those generated by the term-term proximity matrix. However, given a Zipfian distribution, most terms appear infrequently (over half of the terms occur only once), so there is a good chance that the randomly selected terms were low frequency, and hence, did not do as much damage as a high frequency non-discriminating term.

Peat and Willett's (1991) conclusions are based on first order co-occurrence analysis. First order co-occurrences between terms are likely to be inappropriate for determination of synonym-like relations, see section 4.1.3 above. Second order co-occurrence analysis is more likely to be appropriate for determination of such relations. Consequently, Peat and Willett's (1991) criticism does not apply to much of the second order co-occurrence research for thesaurus construction, which have gained much attention in the 1990s.

In the last decade, the perception towards automatic statistical thesaurus construction has changed with the appearance of 'modern procedures' for global corpus analysis (Baeza-Yates & Ribeiro-Neto, 1999). Test results from these new construction methods have shown an improvement in retrieval performance of up to 20% (Crouch, 1990; Crouch & Yang, 1992; Qiu & Frei, 1993; Chen et., 1995). In the following, we present some of these new global corpus-based automatic thesaurus construction approaches.

The traditional approach to automatic thesaurus construction is based on a 'bag of words' approach to indexing and the vector space model (Salton, 1968; 1971). The term-document matrix is transposed into a term-term proximity matrix that depicts first order term associations (Salton & McGill, 1983). Terms are represented as vectors with a component for each document in the corpus. The value of each component is based on a derivation of the $tf \times idf$ weighting scheme, where focus is on mid-

frequency terms, as discussed in Chapter 3. Vocabulary organization is generated most often through single link clustering, where membership of related term clusters is given by a co-occurrence threshold value (Sparck Jones, 1971, Salton & McGill, 1983). This is an $O(n^2)$ process so it is common to limit the number of terms for which a related term list is built (Srinivasan, 1992). This is the the generic procedure we have outline in Chapter 3 and in the previous sections of the present chapter. This is also the basic global thesaurus type much criticized for not improving retrieval performance (Peat & Willett, 1991).

In their *similarity thesaurus* approach, Qiu and Frei (1993; 1995) extend the use of the term-term proximity matrix for statistical thesaurus construction. Term-term relationships are not derived directly from co-occurrence of terms inside documents. Instead, the term-term relationships are obtained by ‘indexing’ each term by the documents in which it appears. Another characteristic of the *similarity thesaurus* is that the vocabulary is reduced to a relatively small number of good discriminating index terms (Qiu & Frei, 1993; 1995). Document frequency is used to reduce the vocabulary, where a variant of the $tf \times idf$ weighting scheme is used to find the ‘inverse term frequencies’ (Qiu & Frei, 1993; 1995). The results from this research show improvements of retrieval performance of up to 20% (Qiu & Frei, 1995).

The use of the term-term proximity matrix has also been successfully elaborated in the domain-specific *concept space* approach to thesaurus construction (Chen & Lynch, 1992; Chen & Ng, 1995; Chen et al., 1995; Chen et al., 1997). Chen and Lynch (1992) developed an asymmetrical similarity measure, called the ‘cluster function’, and found that this function represents term associations better than the cosine measure. Symmetric measures may skew some frequent term-infrequent term relationships, whereas asymmetric measures can depict the roles between terms in such relationships (Chen & Lynch, 1992). Several other interesting features pertain to the *concept space* approach. For example, the domain-specific approach means that a more homogenous sub-language are available for thesaurus construction (e.g., Losee, 1996). Term selection is extended beyond the traditional ‘bag of words’ approach. A more sophisticated strategy is devised for the collection of domain-specific terms from different document sources prior to automatic indexing (Chen et al., 1997). Statistical phrases are also included in the thesaurus and given higher weights than single index terms. In addition, index terms that appear in titles of documents are likewise given higher weights. Finally, the ‘cluster function’ is the basis for a specially developed asymmetric similarity measure. No test has yet evaluated the query expansion capabilities of the *concept space* approach. However, the research behind the *concept*

space project conclude that it is feasible and useful to construct a robust domain-specific thesaurus from this automatic approach (Chen et al., 1997).

Recently, Bookstein et al. (2003) have investigate how the 'clumping strength' of one term is influenced by the presence of another term, thereby using term clumping properties to devise a new term association measure. Proximity measures are mostly symmetric. However, term relationships are often of a non-symmetric nature. According to Bookstein et al. (2003, p. 613), terms *A* and *B* are related symmetrically when either term influences the clumping behaviour of the other term in the same manner. If the terms are related asymmetrically, then one term influences the clumping behaviour of the other; however, in the other direction, the influence may disappear, or even be reversed (Bookstein et al., 2003, p. 614). It appears that most often the asymmetric relations appear to be hierarchical in nature (Bookstein et al., 2003). Bookstein et al. (2003) acknowledge that to detect the precise type of relationship between two terms by purely statistical means seems not to be feasible. However, it is appropriate to focus on the term clumping properties to select a pair of semantically related terms and depict the intrinsic relation.

Crouch (1990) presents an approach to automatic statistical thesaurus construction that focuses on low-frequency terms. Low frequency terms have always been a major interest in automatic statistical thesaurus construction (e.g., Salton & McGill, 1983). However, as pointed out by Peat and Willett (1991), most approaches have not been able to model low frequency terms efficiently. Crouch (1990) argue that the efficiency problems experienced with thesauri constructed from term co-occurrence analysis, is related to the use of medium-frequency terms. To be effective, the related terms selected for query expansion must have high term discrimination values (Salton, Yang & Yu, 1975; Crouch, 1990). This implies that terms must have a lower frequency, see Chapter 3 (Salton, Yang & Yu, 1975). However, as low frequency terms occur in relatively few documents, they are difficult to cluster from the traditional term-term matrix. To circumvent this problem, Crouch (1990) uses a document-document matrix to generate a cluster hierarchy of documents instead. This is done by use of a complete link cluster algorithm (Crouch, 1990). Given the document cluster hierarchy the low frequency terms, that compose each cluster, are selected according to the minimum inverse document frequency value. Thereby the document cluster hierarchy become a thesaurus cluster hierarchy of low frequency terms (Crouch, 1990). This approach has shown an improvement in average precision from 10 to 15% for two small collections (Crouch, 1990).

Schütze and Pedersen (1997) use Singular Value Decomposition (SVD) and clustering to construct a thesaurus. SVD is the dimensionality reduction technique used in Latent Semantic Indexing (LSI) (Deerwester et al., 1990). SVD is able to reduce the dimensionality of sparse matrices, such as a term-document matrix, with a minimum of information loss (Landauer & Dumais, 1997). LSI is an automatic statistical indexing approach that identifies ‘latent concept’ relations among terms by means of their context of use in a text corpus (Deerwester et al., 1990). LSI reduces the dimensions in the vector space along some fewer ‘better’ dimensions, where both terms and documents are represented as vectors. Each new dimension is interpreted as a latent concept and vectors are projected against these new dimensions. A latent concept corresponds to a salient concept that is described by several synonymous terms in a corpus, but not necessarily exactly expressed by either one of them. LSI was originally developed to resolve the vocabulary mismatch problem in IR (Furnas et al., 1987; Deerwester et al., 1990). LSI handles synonymy and polysemy by considering the context of terms. LSI is able to associate terms that do *not* co-occur directly in the same text. Instead, terms are associated by their indirect second order co-occurrence; for example, term *A* may be somehow related to term *B* if both occur frequently with term *C*. If a significant number of terms have an indirect association with term *C*, then the latter could subsequently become a latent concept, and its significant second order co-occurrences would be projected on its dimension.

LSI, and thus SVD, is strongly related to other dimensionality reduction techniques, such as factor analysis and principal components analysis. LSI is one of the most sophisticated attempts of high quality automatic indexing. Overall, this approach also seems very promising for thesaurus construction (Schütze & Pedersen, 1997).

In a similar approach, Gauch and Wang (1996) use the context (i.e. surrounding terms) of each term to construct its vector representation. Again, the key to association is not that two terms happen to occur in the same document; it is that two terms appear, more than by chance, in the same context of a third term. The statistical methods that focus on the context of terms are related to the syntactical thesaurus construction methods introduced below. Albeit, the contextual focus in the latter case is on grammatical relations.

4.6 Automatic syntactic-statistical thesaurus construction

Common for the syntactical approaches to thesaurus construction, is the focus on the syntactic relations index terms take part in; hence, the focus is on the immediate context that surrounds index terms. Syntactic dependency and head-modifier relations are the basis for determination of semantic associations between index terms. Shallow parsing is necessary in order to identify the syntactical relations and extract candidate noun phrases. Once identified, these relations are subjected to statistical analysis in a vector space model in order to find semantically similar index terms.

Syntactic term relations can be viewed as first order associations (Ruge, 1992). Term comparison on the basis of these syntactic relations then leads to linguistically based second order associations. According to Ruge (1999, p. 79), the linguistically based first order associations are semantically more compatible than the traditional statistically first order co-occurrences. Therefore, one would suppose the linguistically based second order associations to be more semantically similar than the statistical ones (Ruge, 1999, p. 79).

Hindle (1990) reports on work that uses the syntactical structure of subject-verb-object dependency relations in text to create a predicate-argument-based thesaurus. According to Hindle (1990), nouns, either as subjects or objects, are similar to the extent to which they share the same verb contexts. Firstly, the text corpus is parsed to identify syntactical relations. Secondly, Hindle (1990) forms a co-occurrence vector for each noun. The vector components contain the Mutual Information value (Church & Hanks, 1990) between a noun and the verbs that take it as the head noun of their subjects or objects (Hindle, 1990). The simple matching association measure is used to define the association between a pair of noun vectors (Hindle, 1990). Hindle (1990) concludes that classes of semantically similar nouns can be acquired automatically through this method. As an example, the nouns most similar to *boat* based on common verbs are *ship*, *plane*, *bus*, *jet*, *vessel*, *truck*, *car*, *helicopter*, *ferry*, *man*, in that order (Hindle, 1990, p. 272). Yet, the application of the Mutual Information value as a source for similarity detection has been criticised, since the value apparently favours rarely occurring words (Damerau, 1993).

Lin (1998) takes a similar approach. Instead of using the smaller value of the corresponding components, as in the case of the Mutual Information value used by Hindle (1990), Lin (1998) uses a normalized version of this measure. Lin (1998) creates a thesaurus that shows the top *n* most similar noun, verb and adjective entries for each index term. Pereira, Tishby, and Lee (1993) extend the syntactical

dependency analysis to clustering of similar nouns by their verb-object relation. Similarly, Tokunaga, Iwayama, and Tanaka (1995) construct a hierarchical thesaurus; however, they base the classification of nouns upon their being the subject of verbs.

The other widely used lexical-grammatical relation for thesaurus construction is that of head-modifier relations within noun phrases. In a simple approach, Evans et al. (1991) identified salient noun phrases in documents. The head in these noun phrases was considered the more general term, which subsumes the more specific concepts expressed by the phrase; for example, 'intelligence' subsumes 'artificial intelligence' (Evans et al., 1991). This is the notion of concept hierarchies created by subsumption of head-modifier relations (Woods, 1997).

Woods (1997) also used phrase analysis in addition to a large knowledge base to organize terms into a concept hierarchy. By locating the head and modifier of noun and verb phrases, Woods (1997) was able to make choices about how to classify phrases. For example, in the phrase *car washing*, Woods' (1997) system would identify *car* as the modifier and *washing* as the head of the phrase. Thus, the phrase *car washing* was classified under *washing* and not *car* (Woods, 1997). The success of the technique relied on a large morphological knowledge base of information to help identify phrase components. Woods (1997) used the concept hierarchy to expand non-matching terms of a query. In a set of retrieval experiments, Woods (1997) reported that use of the expansion method significantly improved the effectiveness of his retrieval system.

Sanderson and Croft (1999) propose a more elaborate method for the construction of a concept hierarchy. Firstly, a document set is grouped into monolithic clusters. Secondly, salient words and phrases in these clusters are organized hierarchically using an asymmetric probabilistic measure of subsumption. According to a small validation by Sanderson and Croft (1999), the generated hierarchy possesses properties expected of such a structure in that general terms are placed at the top levels leading to related and more specific terms below.

Different semantic theories suggest that terms that have many heads and modifiers in common are semantically similar (e.g., Katz & Fordor, 1964). The head and modifier comparison implements the smallest context comparison.

Ruge (1992; 1999), Grefenstette (1992; 1994) and Strzalkowski (1995) use head-modifier relations in a different way. Similar to traditional statistical based co-occurrence methods, these authors construct an $n \times m$ matrix of head and modifier relations. This matrix enables the comparison of head nouns based on their common

modifier contexts. Conversely, modifier words can also be compared based on their common share of head nouns.

Ruge (1992) extracts noun phrases and calculates the association of noun heads by comparing the terms that modify them. A special proximity measure was developed as terms both appeared as head and modifiers. The proximity measure used is a variant of the cosine measure with a logarithmic link frequency weight. The measure takes into account the number of shared heads, when the terms were used as modifiers, and the number of shared modifiers, when the terms were used as heads. Table 4.2 gives some examples of head terms together with their most similar terms with respect to head and modifier overlap.

TABLE 4.2. *Head-modifier relations (Ruge, 1999, p. 81).*

Government	Quantity	President
Leader	Amount	Director
Party	Volume	Chairman
Regime	Rate	Office
Year	Concentration	Manage
Weak	Ratio	Executive
Man	Value	Official
Minister	Content	Head
President	Level	Lead
...	...	...

The example in Table 4.2 illustrates that head-modifier analysis can produce several different types of thesaurus relations; for example: synonyms (quantity – amount), hypernyms (president – head), hyponyms (government – regime) and part-of-relations (government – minister).

Grefenstette (1992; 1994) used both predicate-argument and head-modifier relations of nouns in his attempt to automatically locate synonyms. Grefenstette (1992; 1994) parsed the contexts of nouns for syntactical relations. The similarity between nouns was determined through their common syntactical relations. The measure applied was a special weighted variant of the extended Jaccard measure. Using a number of evaluation schemes, Grefenstette (1992; 1994) found that the success of his method varied depending on the corpus and the frequency of occurrence of the words he was analyzing. The best results came from domain-specific corpora. Grefenstette (1992; 1994) attempted to use his derived thesaurus to aid automatic query expansion with mixed success.

Finally, by use of a direct pattern-matching method, Hearst (1992) found that certain key phrases can be an indicator of a hyponym (is-a) and hypernym (is-a-type-of) relations. Hearst (1992) discovered around ten such phrases that were accurate identifiers of the 'is-a' relation. Three of such key phrases are: *such as*, *and other*, and *especially* (Hearst, 1992). Sentences that contain these phrases were parsed to identify the related noun phrases. Hearst (1992) suggested that key phrases can help thesaurus constructors search for new relations.

Caraballo (1999) proposes an extended use of the pattern-matching method by Hearst (1992), for the construction of a noun hierarchy. Caraballo (1999) first applies a hierarchical agglomerative clustering of parsed nouns, where co-occurrence is based on conjunctions and appositives. Then, by using the pattern-matching method, all the possible hypernym relations between nouns are marked. For each internal node in the cluster tree a word (or more than one word) is selected as the hypernym for the cluster based on the amount of evidence collected by the pattern-matching method

4.7 Summary

This chapter has treated step 2, term association, and step 3, vocabulary organization, of the generic automatic thesaurus construction process, whereas Chapter 3 outlined step 1, vocabulary construction. In addition, the present chapter has focused on the specific feature of proximity measures, most important in determining term association and vocabulary organization. Finally, we have presented some of the results obtained within the statistical and the syntactic-statistical approaches to automatic thesaurus construction.

As described in section 4.6 and 4.7 the last decade has seen some promising results in relation to automatic thesaurus construction. Traditionally, statistical automatic thesaurus construction approaches have mostly been concerned with minimizing the costs of construction. This has led to the rather poor semantic term representations and meagre performance results produced by such thesauri until the beginning of 1990s. However, recent research that employ more sophisticated methods, and at the same time are also more resource demanding, have shown that such methods can produce more refined thesaurus entries and improve performance. These methods take into consideration such elements as specific characteristics of term distributions, location of terms in documents, domain terminology, noun phrases, the lexical-grammatical and lexical-semantic context of terms, second order co-occurrence analysis, asymmetric proximity measures or other more refined applications of

proximity measures, alternative clustering and dimensionality reduction techniques etc. In other words, elements we have presented in Chapter 3 and the present chapter that go beyond the less sophisticated 'bag of words' approach to automatic indexing. However, none of these new methods are perfect and in fact, none of them favourably competes with handmade thesauri in terms of semantic liability. As Bookstein et al. (2003) admit to detect the precise type of relationship between two terms by purely statistical means seems not to be feasible. However, such methods, and the results they produce, are very interesting if they are used in combination with manual intellectual analysis (Soergel, 1974; Aitchison, Gilchrist & Bawden, 2000; Kulyukin & Settle, 2001; Bookstein et al., 2003). The merit of these methods is that they can produce a basis of significant corpus-based semantic relations between terms on a scale otherwise insurmountable for humans to process. When used as a tool for thesaurus constructors and not as a mean in itself, then we speak of semi-automatic thesaurus construction (Soergel, 1974). As Bookstein et al. (2003) suggest statistical detection can be used as a front end for an automatic thesaurus construction system. Once statistically interesting term associations are found in a text collection, another procedure, involving a human indexer or a knowledge intensive natural language processing system, can classify each association as to the exact semantic relationship between the terms (Bookstein et al., 2003).

The purpose of this dissertation is to develop a semi-automatic approach to thesaurus construction. As stated in Chapter 3, semi-automatic thesaurus construction utilizes automatic methods and techniques in the process of vocabulary construction, term association and vocabulary organization. However, semi-automatic thesaurus construction often only considers identification of fewer, but more meaningful, index terms, contrary to automatic thesaurus construction, where all terms are considered. We further stated that our approach differ from traditional automatic thesaurus construction on one major account. Our starting point is not distributions of words in natural language text, but instead, distributions of citations in subject literatures. The difference in origin means that we do not strictly obey the ordering of the steps laid out in this and the preceding chapter. However, we do utilize most of the automatic methods and techniques introduced; not necessarily in the same way, some are applied to bibliometric data instead, but for the same purpose, that is, thesaurus construction. For example, our fundamental premise for vocabulary construction is highly cited documents and their citation contexts. The citation contexts of these documents are investigated to determine if these 'exemplary documents' acts as 'concept symbols' to a majority of citing authors. Citation contexts are treated as individual text windows and subjected to shallow noun phrase parsing. We normalize these noun phrases and

represent them in a vector space representation in order to derive first and second order term associations between them. Likewise, the principal technique we use for vocabulary organization is co-citation clustering of highly cited documents. As a result, we cluster semantically related ‘concept symbols’. This opens up for the possibility to cluster semantically related documents that may differ in terminology, a feature that resembles second order term associations or Latent Semantic Indexing. Consequently, the automatic methods and techniques presented in Chapter 3 and the present chapter, directly or indirectly influence, or form part of, the semi-automatic thesaurus construction approach presented in Chapter 6 and Chapter 8.

The next chapter will outline the idea behind, and the assumptions of the bibliometric methods used in the semi-automatic thesaurus construction approach.

5. Bibliometric methods

The aim of the present chapter is to present the bibliometric methods we apply and investigate for semi-automatic thesaurus construction. The chapter gives emphasis to the co-occurrence aspects of these methods in order to underline the resemblance between automatic thesaurus construction and bibliometric methods. The chapter also pays special attention to the underlying assumptions behind the use of references and citations as a bibliometric indicator of content relationships.

Chapters 3 and 4 describe the automatic thesaurus construction process, with its weight on term distribution models, term co-occurrence analysis, vector space representation, and clustering. Bibliometric methods likewise depend on distributions of entities, co-occurrence analysis, vector space representation, and ordination techniques such as clustering. But, the units of analysis are different. Instead of terms identified in natural language text, bibliometric methods rely on document entities extracted from specific fields of bibliographic records. The most notable entities are references from the bibliographies of scientific documents and their corresponding citations accumulated in citation indexes. The semi-automatic thesaurus construction approach to be presented in Chapter 6, share many common basic features with automatic thesaurus construction approaches described in the two previous chapters, but they differ as to the basic units of analysis. The purpose of this chapter, then, is to introduce the bibliometric methods applied in the semi-automatic thesaurus construction approach, to present their potentials, and discuss their assumptions.

The motivation for the use of bibliometric methods for thesaurus construction comes from research into literature mapping and citation indexing (e.g., White & McCain, 1989; 1997; Wilson, 1999), most notably the research of Rees-Potter (1987; 1989) on semi-automatic thesaurus maintenance. The basic notion is that terminology used in the citation contexts of citing papers reflects concepts of a given specialty area. Of special interest is the ability of some of these methods to cluster topically related documents through bibliographic coupling and co-citation analysis (Kessler, 1963; Marshakova, 1973; Small, 1973). It is important to stress that clustering is carried out, not through term co-occurrence in documents, as in traditional automatic thesaurus construction, but through co-occurrence analysis of references and citations. The latter, may have the potential to cluster documents on the same topic(s) that differ in choice of terminology. The assumption is, contrary to most term co-occurrence methods, that bibliometric methods in this way can help identify synonymous and

near-synonymous relations between candidate thesaurus terms. In the present dissertation, we investigate the bibliometric method of co-citation analysis for its ability to cluster topically related documents in connection with thesaurus construction.

Co-citation analysis is applied in combination with citation context analysis to identify candidate thesaurus terms in a specialty area (Small, 1978; Rees-Potter, 1989). A cluster of topically related documents usually refers to a topic and one or several related concepts (Small, 1978). A vital part of indexing is to identify and select preferred and non-preferred index terms that refer to such concepts. Consequently, we investigate whether term and concept identification is possible by use of the bibliometric method of citation context analysis. Former research indicates that highly cited and co-cited documents attain the status of 'concept symbols' (Small, 1978). A concept symbol communicates a specific topic and can resemble a descriptor (Garfield, 1974). Most importantly, the expression of the concept symbol is more or less agreed upon by authors when they refer to the specific cited document. Citation context analysis investigates the citing contexts in documents that cite into a cluster of topically related co-cited documents. This is done in order to identify agreed upon concept symbols, and to extract significant noun phrases related to the concept symbol from these contexts.

Extracted concept symbols and noun phrases are further subjected to the bibliometric method of co-word analysis, in order to investigate whether this method can be used to construct a network of candidate thesaurus terms. As conceptual and terminological changes are important aspects of maintenance of indexing vocabularies, the bibliometric methods are further investigated for their ability to be used in ageing studies. The idea is that because concepts are linked to references and citations it enables investigations into the citation history of these conceptual 'containers', which might give evidence of terminological changes in a specialty area over a period of time.

Accordingly, the bibliometric methods used for ageing studies are also investigated in the present dissertation.

Section 5.1 introduces the characteristics of bibliometric methods and their corresponding bibliometric analyses. The section pays special attention to the significant relationship between references and citations. Section 5.2 returns to the discussion concerning co-occurrence analysis instigated in Chapter 4. This section uses an entity-relationship diagram to present and discuss co-occurrence possibilities within bibliometrics. The following four sections, sections 5.3 to 5.6, introduce the

different bibliometric methods applied and investigated for thesaurus construction and maintenance in the present dissertation work. The bibliometric methods are respectively, document co-citation analysis, bibliometric-ageing methods, co-word analysis, and citation context analysis. Section 5.7 discusses the premise for citation analysis, that is, the different conceptions of citer motivations. Finally, section 5.8 summarises the main findings of the present chapter.

5.1 The characteristics of bibliometric methods

Bibliometrics refer to mathematical and statistical analyses of patterns that arise in the publication and use of scientific documents (Diodato, 1994, p. ix). Bibliometrics can be described as the utilization of quantified, or quantifiable, attributes of scientific literature for the measurement of the content and evolution of scientific endeavour (Tijssen, 1992, pp. 7-8). The basic source of bibliometric analysis is a collection of publications (bibliographic records) most often in the form of scientific journal papers (Wilson, 1999). Each record can be seen as a repository of bibliographic fields with attribute values. Bibliometrics emerges *not* at the level of individual bibliographic records, but at the level of collections of bibliographic records (White & McCain, 1989). The attribute values from different fields of bibliographic records accumulate as records are assembled into a collection and recur when records having a specified value in a particular field are counted (White & McCain, 1989).

Borgman and Furner (2002, p. 4) indicate the scope of bibliometrics when they state: “[b]ibliometrics offers a powerful set of methods and measures for studying the structure and process of [written] scholarly communication”. One way by which written communicative activity can be explained, interpreted, or otherwise understood, is to consider the objects, agents, events, products, and contexts of such activity as *entities* to be counted, measured, or quantified (Borgman & Furner, 2002). In a former study, Borgman (1990) identified three classes of such entities. They include: 1) the producers of communication, e.g. authors, institutions, fields, countries; 2) the formal products of communication, e.g., journal papers, monographs or aggregate levels such as journals; and 3) communication concepts, which cover author’s use of words in titles, abstracts or other textual parts of documents, indexer assigned descriptors and classification codes, and motivations for citing.

One goal for bibliometric analysis is to identify these abstract entities and make them operational. The assumption is that entities of interest for bibliometric analysis can be located in the specific fields of bibliographic records. Accordingly,

bibliometric methods make the entities operational by counting attribute values in the parent fields of bibliographic records. The result is bibliometric indicators of usage, content or impact of the 'communicative activity' in science (e.g., Cronin, 1984; King, 1987; White & McCain, 1989; Egghe & Rousseau, 1990; Van Raan, 1998; Wilson, 1999; Borgman & Furner, 2002). The focus in this dissertation is on bibliometric indicators of content relationships between documents.

Bibliometric methods can be divided according to the specific type of bibliometric analysis they support. Bibliometric analyses can be separated into two categories according to which type of document entity they apply as unit of analysis. Figure 5.1 below illustrates the two categories of bibliometric analysis and their corresponding bibliometric methods. Publication analyses apply document entities *other* than references and citations, whereas citation analyses apply references and citations (e.g., Wilson, 1999; Wolfram, 2003).

FIGURE 5.1. *Bibliometric methods divided according to the specific types of bibliometric analyses they support.*

Bibliometric methods applied for publication analyses focus on productivity counts and co-occurrence analysis of various document entities other than references and citations. Their primary application have been for scientometric purposes (Wilson, 1999). The bibliometric distributions, which are not exactly bibliometric methods, also belong to this category of analyses (Wilson, 1999). We apply co-word analysis in this dissertation; this publication based bibliometric method is introduced in section 5.5.

Traditionally, bibliometrics have been associated with the “... quantitative study of literatures as they are reflected in bibliographies” (White & McCain, 1989, p. 119). In this context, *literatures* are groups of related documents. The relation is based on shared entities such as references, authors, journals, specialty areas²⁸, or even the overarching fields and disciplines themselves (Wolfram, 2003). Bibliometric methods that study entities from document bibliographies, that is references or their corresponding citations, belong to the group of citation analyses (Wilson, 1999). Citation analysis is centered on occurrence and co-occurrence counts of references and citations. In addition, citation analysis also comprises analyses that focus on the function and content of references, as well as the motivation behind giving references in the first place. Citation analysis is a major part of this dissertation, therefore, we elaborate on it in the following two sub-sections.

5.1.1 Citation analysis

Borgman and Furner (2002, p. 10) distinguish between two general purposes for which citation analyses may be conducted ‘evaluative citation analysis’ and ‘relational citation analysis’. A further distinction may also be drawn between:

- Relational and evaluative studies, where quantitative citation analysis is employed as a method for describing, evaluating, explaining, or predicting some aspect of human behaviour other than the act of citing (Borgman & Furner, 2002, p. 10).
- Citation studies in which the act of citing is itself the phenomenon to be understood (Snyder, Cronin, & Davenport, 1995, p. 77).

In evaluative citation analysis, citation counts are used as indicators or measurements of the level of quality, importance, influence, or performance, of individual documents, people, journals, research groups, institutions, research fields, disciplines, or nations. Evaluative citation analysis is particularly used in scientometric studies of research

²⁸ The term *specialty area* denotes, in this respect, the perceived grouping of scientists who are specialized in the same or closely related topics of research (e.g., Small, 1986).

evaluation (Narin, 1976; van Raan, 1988; 1997; 2000; Garfield & Welljams-Dorof, 1992; Glänzel & Schoepflin, 1994; Narin & Hamilton, 1996). The use of citations as indicators of scientific performance makes evaluative citation analysis more controversial than relational citation analysis (e.g., Edge, 1979; MacRoberts & MacRoberts, 1989a; 1989b; 1996; Peritz, 1992; Seglen, 1992; 1994; 1996; 1998).

Our focus is on relational citation analysis. Relational citation analysis is a means to set references and citations into context (Borgman & Furner, 2002). In relational citation analysis, occurrence and co-occurrence counts of references or citations are used as topical or cognitive indicators of the level of connectedness, the strength of relationship, or the direction of flow, between, documents, authors (oeuvres), journals, domains, specialty areas, disciplines or fields (Borgman & Furner, 2002). Maps, graphs, or networks of such entities can be produced in a way that demonstrate their relatedness to one another. For example, maps can be used to visualize the historical and contemporary structure and direction of 'communication activity' in a particular domain through its literature (White & McCain, 1997; Borgman & Furner, 2002).

The availability of the ISI's citation indexes has enabled large-scale relational analysis of citation and reference structures within different research fields²⁹ (e.g., Garfield, 1955; 1979a; Garfield, Sher & Torpie, 1964). The notion behind scientific citation indexes is linked to IR (e.g., Garfield, 1955). Contrary to conventional term indexing methods, citation indexes enable documents to be related and grouped on the basis of their citation links, co-citation links, or reference links (e.g., Garfield, 1955; 1979a; 1990; 1994; Kessler, 1963; 1965; Small, 1973; Kwok; 1985a; 1985b; Pao, 1988; Pao & Worthen, 1989; Shaw, 1990; Ding et al., 2000; 2001; Wolfram, 2000; White, Buzydlowski & Lin, 2000; Lin, White & Buzydlowski, 2003; White et al., 2004). Today, Web search engines have come to utilize the link structure of the World Wide Web in a bibliometric manner for IR purposes (e.g., Brin & Page, 1998; Kleinberg, 1999; Lawrence, Giles & Bollacker, 1999). More broadly, relational citation analysis is conceived as a useful tool for knowledge organization (e.g., Rees Potter, 1987; 1989; Hjørland & Albrechtsen, 1995; Hjørland, 1997; 2002; Schneider & Borlund, 2002; 2004). Rees-Potter (1987; 1989) is especially of interest to the work presented in this dissertation, as it concerns thesaurus construction by use of bibliometric methods.

The following sub-section clarifies the differences between references and citations, and address the terminological confusion that exists in relation to 'citation analysis'.

²⁹ Evaluative citation analysis is obviously also depending on citation indexes.

5.1.2 *References and citations*

An essential part of research papers, particularly in the natural sciences, is the list of references pointing to prior documents. As Ziman (1968, p. 58) observes, "... a scientific paper does not stand alone; it is embedded in the 'literature' of the subject." A reference is the acknowledgement that one document gives to another; a citation is the acknowledgement that one document receives from another (Narin, 1976). In other words, when document *A* appears in the list of references of document *B*, it means that document *A* has been cited by document *B*. In this case, document *A* is a reference of document *B*, but at the same time document *A* has also received a citation from document *B*. Therefore, a citation is a reference to a document given by a more recently published document. The document citing, document *B* in the above example, is the citing document, and the document that receives the citation is the cited document, in the above example, document *A* (King, 1987).

In general, a citation implies a relationship between a part or the whole of the cited document, and a part or the whole of the citing document (Malin, 1968). Citation analysis is the field of bibliometrics that deals with the study of these relationships, and its main tool is a citation index (King, 1987). A citation index is an ordered list of cited documents, each accompanied by a list of citing documents (e.g., Garfield, 1979a; Smith, 1981; Egghe & Rousseau, 1990). Whether to use the term reference or the term citation depends on the focus of investigation, the citing or the cited documents, respectively. The terminological inconsistencies in relation to the use of the term citation are noticeable in the literature on bibliometrics when researchers do not explicitly explain their focus of investigation. Typically, investigations of citing and cited documents are both termed citation analysis, when this in fact should only refer to the study of the cited document(s), the one receiving citations from later published documents. If the citing document(s) is investigated, then we ought to speak of a reference analysis, the one giving references to earlier published documents. Hence, citation analysis is applied as a broader term when it should have been applied along side reference analysis as a narrower term. For example, see Figure 5.1 for the partition of citation analysis. Citations and references do not exhibit the same characteristics, but in order not to bring in more terminological confusion the term citation analysis is maintained in the dissertation, though ideally, it should be divided into two narrower terms, one reflecting reference analysis and one reflecting analysis of documents receiving citations. We will use the terms references and citations according to Price's distinction (1970, p. 7), in order to bring clarity, though, a strict adoption is not possible due to the already widespread convention of using the term citation in both meanings. For example, investigations of citer motivations, content analyses of citation contexts and the classification of citations, essentially belong to the

category of reference analyses, as illustrated in Figure 5.1. That is so, because the focus of analysis is references given in the text of a citing document, and outlined in its bibliography, and not the documents receiving citations. In addition, much of the discussion on whether or not a 'theory of citation' is needed, and what it should comprise, has actually been concerned mainly with references, and the possible motives and intentions of authors for inclusion of particular ones for their bibliographies (Wilson, 1999). In this context, it is more appropriate to use the phrase 'theory of citing' since the focus of investigation is clearly on the citing documents, and the motivation behind their references.

According to Baldi (1998, p. 830), "... past studies have explained the citation process only in terms of the characteristics of the article being cited ... ignoring the role played by characteristics of the *citing* article". As Small (1978) argued, traditional interpretations of citations present references as the sources the author draws upon to give meaning to his or her text, ignoring that the citing author also imparts meaning to the sources by citing them: "[a citation] constitutes the author's interpretation of the cited work. In citing a document an author is creating its meaning ..." (Small, 1978, p. 328). In this dissertation, we focus on the dyadic³⁰ link between cited and citing documents as envisaged by Small (1978).

Similar to automatic thesaurus construction approaches presented in Chapter 3 and 4, co-occurrence analysis plays a major role in the composition of bibliometric methods; especially the methods used for relational citation analysis. Essentially, co-occurrence analysis can map the structure of subject interrelationships in a collection of documents. The following section is centered on an entity-relationship diagram, which is used to introduce, illustrate, and discuss co-occurrence relations of interest to this dissertation.

5.2 Co-occurrence of document entities

Several authors have listed the most commonly used units of bibliometric analysis (Borgman, 1990; Borgman & Furner, 2002; Börner, Chen & Boyack, 2003; Morris & Yen, 2004). Similar to Borgman (1990), Morris and Yen (2004) define units of

³⁰ A *dyad* is a group of two entities that are treated as one. *Dyadic* means the binary relationship between two entities. In the present dissertation, dyad links between cited and citing documents implies a relation that goes both ways and exist at the same time. The opposite of dyadic is either monadic or polyadic (www.webster-dictionary.org/definition/dyad).

interest to bibliometric analysis as *entities* and depict their relations in an entity-relationship diagram for a collection of journal papers (see Figure 5.2 for our modified version). Such a definition of document entities helps identify and explain what can be studied by bibliometric methods.

The generic procedure in a bibliometric analysis is the following, given a set of specific entities, attribute values can be collected about samples drawn from general populations of such entities. These attribute values can be analyzed by use of bibliometric methods and statistical techniques. Conclusions can be drawn from these results about the nature of the populations, about the existence of certain causal processes, or about relational ties between these entities. In bibliometrics, the derived measures are typically frequency counts of observed events that can be considered as probabilities of occurrence and co-occurrence (Borgman & Furner, 2002).

Figure 5.2 illustrates different document entity-types (circles) and their direct pairwise associations (links) in a collection of journal papers³¹ (Morris & Yen, 2004).

FIGURE 5.2. Entity-relationship diagram of co-occurrence relations inspired by Morris and Yen (2004).

³¹ The diagram is not exhaustive; see Morris and Yen (2004) for a more comprehensive diagram.

All links in Figure 5.2 are dyadic, that is, they occur between two entities at the same time. In a dyadic link, the two entities can be: 1) *like* entities, that is, entities of the same entity-type, or 2) *unlike* entities, that is, entities of different entity type. The first entity of interest in a dyad is the primary entity while the other entity is the relative entity. Primary and relative entity-types are the basis for a vector space representation in an $n \times m$ matrix, as introduced in Chapter 3. Occurrence links are between unlike entities that are associated with each other. For example, there is an occurrence link between a paper and a reference if the reference appears in the bibliography of the citing paper. Co-occurrence links, or relations, among entities of the same type, occur when the like entities of the dyad are both associated with an entity of a different entity-type. For example, two references are related when they both appear in the same bibliography of a citing paper. In co-occurrence relations, the like entities of the dyad are primary entities, while the unlike entities, with which they co-occur, are the relative entities.

5.2.1 Co-occurrence relations

On a general level, co-occurrence relations between pairs of entities often imply some meaningful relation between those entities (e.g., Ellis, Furner-Hines & Willett, 1994). In Chapter 4, we discussed several types of potential co-occurrence relations between terms. Similar, in bibliometrics two references that co-occur frequently in bibliographies of citing papers are presumed to treat related subject matter. In reverse order, common references between citing papers imply that pairs of citing papers also treat related subject matter. Consequently, co-occurrence analysis is the main tool used to map automatically the structure of subject relationships within IR and bibliometrics. In bibliometrics, co-occurrence relations are manifested in scientific literatures, in IR, co-occurrence relations are manifested in the natural language text of a document corpus (e.g., White & McCain, 1989; Salton & McGill, 1983)

In Figure 5.2, co-occurrence relations are indicated by labels placed next to their primary entity-type and adjacent to the link that connects the primary entity-type to the relative entity-type. For example, in 'document co-citation', the primary entity is the cited references and the relative entities are the citing papers. Some co-occurrence relations are trivial and not useful (Morris & Yen, 2004). In Figure 5.2, seven co-occurrence relations commonly studied in bibliometrics and IR are marked in bold and given their common names. Other co-occurrence relations in the diagram have no commonly used label; these are marked in grey and placed near their primary entity. The co-occurrence relations of interest to this dissertation include:

- Term co-occurrence analysis (e.g., Salton & McGill, 1983): The relation between two terms based on their co-occurrence in documents. The assumption is that the degree of co-occurrence determines the strength of ‘semantic relatedness’ between two terms. This is the basis of automatic thesaurus construction presented in Chapter 3 and 4. Term co-occurrence is usually considered as an automatic indexing method and not a bibliometric method. The notion of *term* refers to the special procedure of term selection as described in Chapter 3. Nevertheless, word-like co-occurrence analysis does have a special application in bibliometrics. This is the bibliometric method of co-word analysis.
- Co-word analysis (e.g., He, 1999): The relation between two words based on their co-occurrence in documents. In contrast to term co-occurrence analysis, co-word analysis applies specifically selected document entities for its analysis. Such entities can include words or phrases extracted from the title, abstract or summary sections in scientific papers, or keywords, descriptors or classification codes extracted from their respective fields in the bibliographic records of scientific papers. Consequently, co-word analysis does not rely upon term distributions in natural language text to identify its units of analysis. Co-word analysis is applied in a special role in this dissertation. We use noun phrases extracted from citation contexts as the units (i.e. words) in a co-word analysis. The purpose is to create a network of candidate thesaurus terms. Section 5.5 introduces the procedure of, and assumptions behind, co-word analysis.
- Co-citation analysis (Marshakova, 1973; Small, 1973): The relation between two references (cited documents) based on their co-occurrence in bibliographies of citing papers. The assumption is that the degree of co-citation determines the strength of common subject relationship between references (cited documents). Moreover, two frequently co-cited references are ‘symbols’ of similar ‘base knowledge’ (Small, 1978; Morris & Yen, 2004). In addition, co-citation analysis can also be performed on the cited reference author(s) (White & Griffith, 1981) or cited reference journal (McCain, 1991a; 1991b). Besides, the co-citation strength between two documents (or their aggregates) are not fixed, as it may change over time if they accumulate new co-citations. Document co-citation analysis is applied in the present dissertation in order to cluster topically related documents without having to depend on them using similar terminology. Further, the co-citation clusters and their cited documents are investigated for candidate thesaurus terms by use of citation context analysis. Section 5.3 introduces the procedure of, and assumptions behind, document co-citation analysis. Moreover, section 5.6 introduces citation context analysis.

These are the co-occurrence relation of interest to this dissertation. The next subsection describes some of the special characteristics that emerge when primary entities are grouped based on co-occurrence relations with relative entities.

5.2.2 The characteristics of 'entity groups'

The co-occurrence counts between a primary entity-type and a relative entity-type at the opposite end of the relation, together with ordination techniques, creates an 'entity group' of the former. Entity groups are characterized by the possession of communalities (Morris & Yen, 2004). Structural mapping is the process of finding and labelling meaningful entity groups and the relations between them (Morris & Yen, 2004). Conversely, searching is the process of finding important individual entities and relations (Salton & McGill, 1983). The name of such entity groups are indicated in Figure 5.2 within the entity where they emerge. The following entity groups are of interest to this dissertation, 'research fronts', 'intellectual bases', and 'vocabularies'. 'Research fronts' are clusters of citing papers that tend to refer to the same base of references and therefore are assumed to share common research topics (Price, 1965; Cozzens, 1985b; Garfield, 1986; Braam, Moed & van Raan, 1991a; Persson, 1994; Morris et al., 2003). 'Research fronts' can be established both as a result of bibliographic coupling and co-citation analysis (Braam, Moed and van Raan, 1991b; Persson, 1994; Morris et al., 2003). According to Morris et al. (2003), these 'research fronts' can be considered as representing 'Kuhnian puzzles' within a scientific field (Kuhn, 1962). For a successive number of years ISI identified 'research fronts' for specialty areas in the Science Citation Index® (SCI®), and indexed documents according to 'research fronts' (Garfield, 1994).

'Intellectual bases' are clusters of references that tend to be co-cited together, and serve as a 'core' of related base knowledge for a 'research front' and its citing authors (Persson, 1994; Small, 1997; Morris & Yen, 2004). According to Morris & Yen (2004), these 'intellectual base' clusters can be considered as representing 'Kuhnian exemplars' or paradigms. Small (1977; 1978) suggests that the key concepts, methods, or experiments, which researchers build on are represented by cited references located in these clusters. As a result, Small (1978) suggests that the 'intellectual base' references represent knowledge claims in the form of 'concept symbols' in the documents that cite them.

Finally, 'vocabularies' are clusters of terms or words derived from co-occurrence of terms or words in papers. If terms are used as the primary entity-type in a co-occurrence analysis, we then consider the resulting vocabulary a thesaurus as described

in Chapter 3 and 4. If words are used as the primary entity-type in a co-word analysis, then we consider the resulting clusters as specialized vocabularies within a research field (Callon, Courtial & Laville, 1991).

5.2.3 *'Core' and 'scatter' behaviour of 'entity groups'*

White and McCain (1989) describe how bibliometrics is grounded in the patterned behaviour of humans when a document is produced, published and indexed. More specifically, bibliometrics is grounded in the 'linguistic choices' by which humans associate indicators of content with the documents. Humans in this context include authors, editors and indexers (White & McCain, 1989). White and McCain's (1989) 'linguistic choices' are expressed through the different entities of a document, as suggested by Morris and Yen (2004) and illustrated in Figure 5.2. As noted above, the 'linguistic choices', in the document entities, supply the attribute values of bibliographic records and create bibliometric data as they accrue over time (White & McCain, 1989). To White & McCain (1989), different document entities can be seen as indicators of content that comes from a range of different cognitive origins (Ingwersen, 1996). Moreover, the specific instances of individual document entities are to be considered as 'index terms'. These human choices of indication of content take a distinctive pattern of concentration and dispersion within a collection of papers (White & McCain, 1989).

A number of document entities, when collocated in 'entity groups' based on co-occurrence relations between primary and relative entities, exhibit 'core' and 'scatter' relations (White & McCain, 1989; Morris & Yen, 2004). Entity groups tend to possess a small set of 'core' 'linguistic choices' (i.e., 'index terms'), that are strongly related to each other, and a large number of 'scatter' 'choices' (i.e., 'index terms'), which are weakly related (White & McCain, 1989). This pattern of statistical association between a primary entity and co-occurring relative entities is essential for co-occurrence analysis and bibliometrics in general (White & McCain, 1989).

The 'core' and 'scatter' relations among entities in a collection manifest themselves as power-law distributions of entity frequency. Some of these power-law relations are noted in Figure 5.2 in *italic* and underscore at their parent relations. They originate from the founding mathematical models of bibliometrics. For example, Lotka's size-frequency distribution of papers over authors (Lotka, 1926); Bradford's rank-frequency distribution of papers in a discipline over journals (Bradford, 1934); Zipf's rank-frequency distribution of word tokens over types (Zipf, 1949), see Chapter 3; and finally, the reference power law for frequency of references to papers by Naranan (1971). Moreover, the entity groups of 'research fronts' and 'intellectual

bases' are to be seen as 'cores' in a literature of citing papers or cited documents respectively.

According to White and McCain (1989, p. 124) "[w]hat one must grasp is that it is not authors (people) and journals (publications) that are somehow similar but authors' names and journal names as co-occurring indicators of content." What follows from White and McCain's (1989) line of reasoning is that bibliometrics is based on distributions of 'index terms' in a similar way as automatic indexing. Automatic indexing rely on term distributions in natural language text, while bibliometrics most often rely on 'core' and 'scatter' relations of 'index terms' accrued from document entities. As a result, those who study subject term relations, subject citation relations, or some other combination of entities from a bibliographic record, are all undertaking the same kind of thing: studying "clouds of language that form around published works" (White & McCain, 1989, p. 125).

An important result that emerges from bibliometric studies is that these content indicator document entities are associated, that is, they are not statistically independent of each other (White & McCain, 1989). The consequence is that an 'index term' of one kind can be converted into a 'core' and 'scatter' of associated terms (of the same and different kinds) with high interpretability (White & McCain, 1989, p. 124). Some interesting consequences emerge, for IR, literature mapping, and especially for this dissertation. According to White & McCain (1989, p. 127), an 'index term' of a particular kind concentrates associated 'terms' when it produces a 'core' and 'scatter' distribution. Accordingly, in the present dissertation we investigate whether highly cited references, as a kind of 'index terms', concentrate agreed upon 'concept symbols' in citing articles by use of co-citation analysis (Small, 1978; White & McCain, 1989). In Figure 5.2, this notion is indicated by an indirect link path going from the reference entity, through the citing paper entity, ending in the word entity. Consequently, our aim is to use one type of 'index term' to find another type that essentially expresses the same meaning. The reason for this indirect approach is that we can cluster candidate thesaurus terms based on references instead of natural language terms. This way, we may be able to cluster semantically related terms, such as synonymous and near-synonymous terms, which are traditionally difficult to relate in first order co-occurrence analyses, as discussed in Chapter 4.

The following sections introduce the different bibliometric methods we investigate in the present dissertation. The subsequent section 5.3 presents document co-citation analysis, section 5.4 introduces bibliometric ageing methods, section 5.5 describes co-word analysis, and finally section 5.6 presents citation context analysis.

5.3 Document co-citation analysis

Co-citation analysis studies structures of scientific research based upon citations and co-citations (Gmür, 2003). The mere existence of a reference in a citing paper, whether used in a positive or negative sense, is taken as a measure for the significance allocated to the cited reference (Gmür, 2003). A citation is taken to be a valid and reliable indicator of scientific communication (Small, 1978; Garfield, 1979b). A co-citation relationship is interpreted as an ‘intellectual tie’ (semantic or cognitive) between two cited entities, and their co-citation strength is a measure of this ‘intellectual tie’. Co-citation analysis identifies entity groups of cited documents, cited authors or cited journals (Marshakova, 1973; Small, 1973; White & Griffith, 1981; McCain, 1991a; 1991b). Co-citation analysis enables analysis and mapping of the inner subject structure of a literature, its cognitive relationships, schools of research, paradigms, its conceptual networks, and its development over time (e.g., Small, 1973; Small, 1978; Small & Greenlee, 1980; White & Griffith, 1981; Chen et al., 2002). Marshakova (1973) and Small (1973) are credited for independently suggesting the use of co-citations as a measure of inter-document relatedness. Mullins et al. (1977) and McCain (1986) have shown that the co-citation structure of a specialty area is a fair representation of how it is perceived by its members.

Three types of co-citation methods exist: document co-citation analysis of cited documents (Small, 1973; Griffith et al., 1974), author co-citation analysis of cited authors (called *oeuvres*) (White & Griffith, 1981; 1982), and the more rarely used journal co-citation analysis of cited journals (McCain, 1991a; 1991b). In this dissertation, we apply document co-citation analysis to map the subject structure, and its conceptual network, in a research specialty.

5.3.1 *Document co-citation analysis*

Document co-citation analysis is the oldest method of the three. The method was primarily developed to map the structure of scientific literatures, in order to study how they evolve and change over time (e.g., Small, 1973; 1977; 1993; 1997; 1999; Griffith et al., 1974; Small & Griffith, 1974; Garfield, 1979a; Small & Greenlee, 1980; 1989; Nadel, 1981; Small & Sweeney, 1985; Small, Sweeney & Greenlee, 1985). These methodological studies are known as macro-level studies because they focus on the overall structure of fields and disciplines, and, ultimately, on the question of what govern the evolution of science (e.g., Griffith et al., 1974; Small & Griffith, 1974; Small & Greenlee, 1980; Small & Sweeney, 1985; Small, Sweeney & Greenlee, 1985;

Garfield, 1986; 1994). Small and colleagues have successfully achieved the goal of developing such a method for mapping the structure and development of scientific research using document co-citation analysis and different clustering methodologies (e.g., Small & Greenlee, 1980; Small & Sweeney, 1985; Small, Sweeney & Greenlee, 1985). Document co-citation analysis is founded on the premise that the most valid and reliable indicator of a school of research, its scientific assessment or methods, are expressed in its documents available through peer-reviewed publications (Small, 1973; 1978). Moreover, Small's notion with document co-citation analysis is primarily a way to map out in detail the relationships between 'key ideas' as they are conceptualized in cited documents within research specialties (Small, 1977; 1978; 1980; 1986; Small & Greenlee, 1980; Small & Sweeney, 1985; Small, Sweeney & Greenlee, 1985). According to Small (1978), frequently cited documents in a specialty area often come to symbolize 'key ideas' to the current citing authors within that specialty area. Small's (1978) notion is based on Garfield's (1955) original idea of document citation linkages as a tool for citation indexing. Garfield (1955; 1979a) argued that cited documents could be seen as alternative subject headings.

Micro-level co-citation studies seek to describe retrospectively the structure, historical development, and possible interdependencies of individual schools of research, domains, or specialty areas (White & Griffith, 1981; 1982; McCain, 1990). Author co-citation analysis has dominated the micro-level approach to co-citation analysis (Gmür, 2003). Micro-level co-citation studies that apply cited authors as entities essentially map the oeuvres of these authors (McCain, 1990; White & McCain, 1998). Consequently, the structure of a domain, its schools of research, and its line of development, is depicted through its most prominent representatives and not its 'prominent documents' as in document co-citation analysis. Whereas document co-citation analysis depends on the cited reference string within the bibliographic records of ISI's citation indexes, co-citation data of author pairs can be obtained without downloading these expensive bibliographic records (e.g., McCain, 1990; Gmür, 2003). This makes author co-citation analysis less expensive than document co-citation analysis. Perhaps this is a major reason why author co-citation analysis has been used more often in micro-level studies. Today, Web of Science® enables free downloading of a substantial number of ISI records. This should enable more micro-level studies based on document co-citation analyses. The present dissertation is a micro-level study that applies document co-citation analysis.

5.3.2 *Proximity measures and ordination techniques*

The exploration of ‘intellectual’ structures based on co-citation analysis within a literature, for example a specialty area, is commonly known as ‘literature mapping’ or synonymously ‘visualization of literatures’ (e.g., White & McCain, 1989; 1997; Wilson, 1999; Börner, Chen & Boyack, 2003). Co-citation analysis indicates pairwise relations, so in order to map the salient structures within a multivariable matrix of many pairwise relations, ordination techniques are needed (Börner, Chen & Boyack, 2003; Schneider & Borlund, 2004). Ordination is a common name for a number of multivariate statistical techniques that focus on dimensionality reduction and structuring of large matrices. The most commonly used ordination techniques in literature mapping are cluster analysis, multidimensional scaling (MDS), factor analysis, principal components analysis, Singular Value Decomposition, and Pathfinder networks (White & McCain, 1997; Börner, Chen & Boyack, 2003; Schneider & Borlund, 2004). In Chapter 4, we introduced cluster analysis, and briefly mentioned Singular Value Decomposition, as the ordination techniques which are the most widely used for automatic thesaurus construction. Moreover, factor analysis and principal components analysis are mostly applied in author co-citation analysis (Börner, Chen & Boyack, 2003). The ordination techniques mostly used in document co-citation analysis include cluster analysis, MDS, and recently Pathfinder networks (Börner, Chen & Boyack, 2003; Schneider & Borlund, 2004).

Generally, an $n \times m$ matrix of cited entities and citing documents are transformed to an $n \times n$ matrix that represents the pairwise co-citation counts between the cited entities. This is this the classical vector space representation outlined in Chapter 3. The pairwise co-citation count represents the strength of ‘relatedness’ between two entities. Similar to automatic indexing, raw co-citation counts are most often ‘weighted’ or normalized and represented in a proximity matrix. Most often in document co-citation studies, the co-citation count of a pair of co-cited documents is measured in relation to the individual citation counts of the cited documents (Braam, Moed & Van Raan, 1988; Hamers et al., 1989). The aim is to avoid an interpretation of a co-citation density of documents due simply to their wide spread, whilst overlooking proximity relations between more rarely cited entities (Gmür, 2003). Small and Greenlee (1980) introduced the extended Jaccard association measure (12), which satisfies the above-mentioned criteria, see Chapter 4 for the composition of this measure. Small and colleagues at ISI have since applied this measure for co-citation clustering. Braam, Moed & Van Raan (1988) recommend that both the extended Jaccard and the cosine measures should be used in a parallel analysis of a data set in document co-citation studies. This is one way to investigate if the proximity measures and ordination

techniques reveal the salient structures inherent in the data set, instead of imposing a structure on the data set.

The document co-citation method developed by Small and colleagues at ISI include cluster analysis. They apply single-linkage clustering to the full annual files of ISI data in order to derive a macro-level disciplinary structures of nested topical clusters (e.g., Griffith et al., 1974; Small & Griffith, 1974; Small & Greenlee, 1980; 1989; Small, Sweeney & Greenlee, 1985; Small, 1999). Cluster analysis is also applied to micro-level document co-citation studies (e.g., Börner, Chen & Boyack, 2003). Moreover, if proximity measures are converted into distances (see Chapter 4), then the cited entities can be represented as plots in a multidimensional Euclidean space by use of MDS. MDS is the most widely used visualization technique in co-citation analysis (e.g., White & Griffith, 1981; McCain, 1986; McCain, 1990; White & McCain, 1998). The plots in the multidimensional space depict the relations among entities as physical distances. Usually two-dimensional plots are used, but higher dimensions are possible. A long acknowledged weakness of MDS is that only a small set of entities can be represented in the space, which exclude macro-level studies unless they incorporate some form of zooming technique (Small, 1997; 1999). Another weakness of MDS is related to the interpretation of the nature of the resulting dimensions (Börner, Chen & Boyack, 2003). In studies where more local detail and explicit representation of structures are needed, a network solution may be preferred as suggested by Chen (2003). As a result, the use of Pathfinder networks have gained much interest recently for literature mapping based on co-citations (e.g., Chen, 1999; 2003; 2004; Chen & Paul, 2001; Chen, Paul & O'Keefe, 2001; Chen et al., 2002; Chen & Morris, 2003; White, 2003b). Pathfinder networks are introduced and applied in Chapter 8.

5.3.3 Critique of co-citation analysis

Questions have been raised, whether co-citation analysis can identify an entire research specialty, or only a minor subgroup within it (e.g., Sullivan, White & Barboni, 1977; Healey, Rothman & Koch, 1986). Various proposals have been made to improve the alleged low 'recall' of papers in co-citation clusters. For example, Braam, Moed and Van Raan (1991a) suggest the use of co-words in addition to co-citations, and Small (1996) propose of new 'combined linkage measure'.

Small (1977) claims that the citing authors of a cluster constitute a highly relevant subgroup of the current practioners of a specialty. According to Rip (1988), only subgroups with 'shared legitimately tactics' are traced. King (1987), sums up a number of objections against co-citation analysis: loss of relevant papers, inclusion of

non-relevant papers, overrepresentation of theoretical papers, time lag (between emergence of new specialties and capturing of them in a co-citation map), and subjectivity inherent in the setting of threshold levels, as these threshold levels strongly affect size and content of clusters. Others are much more sceptic, and maintain that clusters are mainly artefacts of the applied technique having no further identifiable significance (e.g., Oberski, 1988). As a consequence of this critic, Braam, Moed and Van Raan (1991a; 1991b) investigate whether co-citation analysis indeed provide a useful tool for mapping 'subject matter specialties' of scientific research. The 'cross-examination' method they use is co-word analysis. Their work clarifies a number of issues concerning co-citation analysis. Their results suggest that co-citation analysis indeed depict research specialties and their internal structures, although one specialty may be fragmented across several different clusters. They conclude that co-citation clusters are certainly not artefacts of an applied technique. Yet, their study also suggests that co-citation clusters did not represent the entire body of documents that comprise a specialty. Therefore, they concur with the recommendation of Mullins, Snizek and Oehler (1988) that it would be necessary to analyze different structural aspects of publications to generate significant results in science and literature mapping.

In citation analysis, it is common to study ageing patterns of references and citations within literatures, such as specialty areas. When candidate thesaurus terms are linked to references and citations, it becomes natural to investigate the citation history of these conceptual 'containers' (i.e. cited references) in order to study if bibliometric methods used for ageing studies can give an indication of terminological change over a period of time within a specialty area. The following section presents ageing studies and their related bibliometric methods.

5.4 Ageing studies of literatures and documents

The growth of literatures and their obsolescence represent respectively the initial and final stages of the scientific 'information life cycle' (Wolfram, 2003, p. 56). Growth studies investigate regularities in the creation of literatures or documents over time (Price, 1986; Tabah, 1999). Obsolescence studies investigate the decline in use of literatures or documents over time (Line & Sandison, 1974). Growth and obsolescence studies can focus on individual documents, their aggregates, or more broadly on various literatures (Wolfram, 2003). Growth studies belong to the group of publication analyses. Obsolescence studies belong to the category of citation analyses, as

literature usage is measured through reference and citation data. Our focus is on obsolescence. Several functions are accredited to obsolescence studies. For example, they enable investigation into usage patterns of knowledge by scholars, as well as exploration of the changing intellectual structure of science (Brookes, 1981; White & McCain, 1989).

5.4.1 Obsolescence of literatures

Obsolescence involves the decline in use of documents over time. This is most easily studied by use of document citation counts, where the continued citations received by a document are taken as an indication of its value to a research specialty, discipline etc. (Line & Sandison, 1974). As new documents are published in, for example, a discipline, citation counts for a given document usually decline as the 'research fronts' of the discipline move on. Eventually, a document may no longer be cited. It has become obsolete (Line, 1993). Some documents, on the other hand, continue to see high levels of recognition long after their publication. They become 'citation classics', which authors will continue to cite in recognition of their contributions (e.g., Garfield, 1984). Very few documents fall into this category. Within a literature group, there are significantly more documents that are never cited at all, and although published, they remain unacknowledged (Line, 1993).

For documents that receive sizeable numbers of citations, the distribution of citations over time is generally unimodal³², with the mode dependent on how quickly document findings are cited after publication. White and McCain (1989) point to some common characteristics of unimodal obsolescence distributions, for example, a plotted curve will generally peak in year two to three. Price (1965; 1970) attributes this disproportionate citing of more recent papers to an immediacy effect produced by citing ephemeral papers, as opposed to 'classic papers', at the 'research front'. Price (1970) proposes a measure to describe the recency of citations in a document named the *Price Index*. It is the proportion of citations no more than five years old divided by the total number of citations (Price, 1970; Cozzens, 1985b). Several obsolescence studies of the natural and life sciences support Price's observations (e.g., Braam, Moed & Van Raan, 1991a; 1991b). For example, Braam, Moed, and Van Raan (1991a) suggest that the *Price Index* can be used to identify a 'research front'.

A document begins to receive citations after an initial lag that results from the publication delay of works citing the document and the time needed to carry out research influenced by the initial document (Schubert & Glänzel, 1986). It is also

³² A unimodal distribution is a statistical distribution, such as the normal distribution, that has a single 'peak.'

possible that a document will remain uncited or rarely cited for long periods of time because its contributions are not initially recognized or valued by the research community. Over time, more researchers become aware of the document's value, so the citation count grows. As new documents that contribute to the subject area are published, the citation focus shifts to these new documents. This results in a decline of references being made to the initial document. Citation classics, enjoy continued recognition, with no indication of decline in citations for many years (Line, 1993).

The concept of obsolescence itself is relatively straightforward, but there are two main approaches, synchronous and diachronous, that may be used to collect and analyze obsolescence data. Methods used in obsolescence studies are usually tied to one of these approaches.

5.4.2 *Synchronous and diachronous methods of obsolescence*

Most obsolescence studies have used a synchronous approach, which examines the distribution of the ages of references that appears in a set of citing documents over a specified time, such as a year (Gross & Gross, 1927; Line & Sandison, 1974; Line, 1993). This approach provides a cross-section of citing behaviour within a literature for that point in time. Due to the skewness observed in the distribution of citation ages, the 'median citation age' instead of a 'mean citation age', is used as an indicator of how quickly a literature obsolesces (Line & Sandison, 1974). For example, in some scientific and technological disciplines, the median citation age is quite short because of the rapid turnover in developments, while in the humanities, the median citation age is usually much longer, which indicates a slower rate of obsolescence (Line, 1981; Cunningham & Bockock, 1995; Glänzel & Schoepflin, 1995). As an example, Cunningham and Bockock (1995), found ageing values that ranged from circa 4 years for metallurgical engineering to circa 22 years for Biblical criticism. Wilson (1999) suggests that such results can be obscured by differences in publication types between disciplines. Lindholm-Romantschuk and Warner (1996) shows that the proportion of references to monographs versus references to papers are significantly higher in the humanities than in the social sciences.

Diachronous studies of obsolescence follow the citation behaviours of cited documents over time, usually from the time they are published (Line & Sandison, 1974; Line, 1993). This allows researchers to track the citations received by documents as they age. Obviously, a high proportion of documents is never cited or receives insufficient citations to warrant charting their diachronous usage profiles in this way.

Obsolescence can be thought of as analogous to radioactive decay in physics (e.g., Száva-Kovats, 2002). The term half-life in diachronous obsolescence studies is used as a measure of the time in which half of the currently cited literature is published (e.g., Burton & Kebler 1960; Burton & Green, 1961; Száva-Kovats, 2002)³³.

Obsolescence studies are not without their critics, particularly because it can be argued that they may not measure 'actual' obsolescence but rather 'apparent' obsolescence. Obsolescence studies are often criticised for not taking into account the growth of literatures, which could affect citation counts received by a given paper (e.g., Sandison, 1980; 1987; Motylev, 1989; Line, 1993). However, Brookes (1981) concludes that adjustment for growth is unnecessary if researcher's interest is in the growth rate of knowledge, the rate of revision of scientific literature, and the diminishing utility of previously existing information.

5.4.3 Ageing patterns for individual documents

Ageing studies of individual cited documents are of special interest in the present dissertation, as they are the actual 'containers' of the candidate terms. Aversa (1985) makes the first systematic analysis of generalized diachronous aging patterns for individual documents. From individual citation histories, Aversa (1985) identifies two ageing groups for a set of scientific papers. On average, citations for one group peaked in the third year after publication and declined rapidly thereafter. Citations for the other, more highly cited group peaked in the sixth year and then dropped off slowly.

McCain and Turner (1989), in a diachronous study, found overlapping citation counts for two subsets of papers on molecular genetics, each subset corresponds to one of Aversa's (1985) aging groups. Citation profiles of 11 highly cited papers were traced, four papers peaked at 5-6 years with a slow subsequent decline, and seven papers peaked at 2-3 years with a rapid decline. By use of citation context analysis, McCain and Turner (1989) confirmed previous speculation by Garfield (1975; 1984) that the two aging patterns could be related to the role the papers played in subsequent published research. Papers cited for their theoretical or technical contributions peaked later and aged less rapidly than those cited for specific research findings and experimental results. McCain and Turner (1989) suggest that the former subset of

³³ Recently, Száva-Kováts (2002) demonstrated that Burton and Kebler (1960) did not introduce the notion of half-life in connection with obsolescence studies. Conversely, Burton and Kebler (1960) showed that there is an essential difference between the nature of radioactive half-life and that of literature half-life, and they therefore disapproved of the use of the term half-life in ageing studies of literatures. According to Száva-Kováts (2002), it is therefore unfounded and erroneous to continue to attribute the notion of literature half-life to Burton and Kebler (1960).

papers inspire to new work or offer superior analytical methods of longer duration. The latter subset of papers apparently has immediate and limited application to the field.

These two studies indicate that at a fine level of detail, citation links are content links (White & McCain, 1989). They also reopen the issue of the nature of ‘actual’ obsolescence. In what sense has the content of a document become obsolete when it ceases to be cited? Is it still used in the wider sense of having been learned, ‘obliteration by incorporation’, or has it become outdated? No doubt, that citation and reference counts can broadly measure document usage, thus, the question for this dissertation is whether obsolescence methods, can be used as a tool for identification of conceptual changes in a subject area.

The following section introduces the publication based bibliometric method of co-word analysis.

5.5 Co-word analysis

Co-word analysis is a publication based bibliometric method. Today, the co-word method is a content analysis that uses co-occurrence patterns of identical document entities, extracted from specific fields in bibliographic records or sections in the text of documents. The document entities chosen for analysis is typically words or phrases extracted from the title, abstract or summary sections in scientific papers, or keywords, descriptors or classification codes extracted from their respective fields in the bibliographic records of scientific papers. The procedure of co-word analysis is almost identical to term co-occurrence analysis presented in Chapter 4. But, in contrast to term co-occurrence analysis, co-word analysis does not rely upon term distributions in natural language text to identify units of analysis. The selection of units in a co-word analysis is often much more selective, which usually results in much smaller matrices. Besides, co-word analysis often apply specially developed asymmetrical proximity measures, in contrast to the symmetrical ones traditionally used in long-span term co-occurrence analysis.

Originally, co-word analysis was inspired by the actor-network theory, which fundamental premise is based on scientists’ use of scientific publications as a ‘vehicle’ for research ideas (Latour, 1987). The original purpose of co-word analysis was to identify relationships between ‘ideas’ within a research area in order to map the dynamics of science (Callon, Law & Rip, 1986a; He, 1999). Co-word analysis was developed in opposition to co-citation analysis, as Callon and colleagues did not

approve of citations as an entity by which science could be mapped (Callon, Law & Rip, 1986a). Ironically, Callon, Law and Rip (1986a) returned to a representation of literatures discarded by proponents of citation indexing, that of subject indexing (e.g., Garfield, 1955; 1979a; Wouters, 1999).

In accordance with this view, early co-word analysis used keywords assigned by indexers as units of analysis. Today, this conception of co-word analysis has broadened to encompass almost all document entities in a bibliographic record as outlined above (see for example Leydesdorff, 1987; Hinze, 1994; McCain, 1995; Rotto & Morgan, 1997; He, 1999; Noyons, 1999). In addition, words and phrases extracted from either full text documents or specific positions in full text documents have also recently been applied in co-word analysis (Kostoff, Eberhart & Toothman, 1997).

Law and Whittaker (1992) describe the modern assumptions of co-word analysis. Co-word analysis assumes that words used by authors or indexers to write and index a paper reflect the present stages of the scientific research question. Moreover, co-word analysis assumes that arguments received by other scientists will lead to publication of further scientific papers that are written and indexed by similar sets of words (Law & Whittaker, 1992). If all these are reasonable assumptions, it is then possible for co-word analysis to make use of frequencies of word pairs in the set of papers as a way to map the structure of 'research idea and concepts' embodied in the papers (Law & Whittaker, 1992).

Since its introduction, co-word analysis has been applied to detect the topics in a given research area, the relationship between these topics, the extent to which these topics are central to the whole area, and the degree to which these topics are internally structured (e.g., Callon, Law & Rip, 1986a; Courtial & Law, 1989; Whittaker, 1989; Braam, Moed & van Raan, 1991a; 1991b; Callon, Courtail & Laville, 1991; Coulter, Monarch & Konda, 1998; Noyons & van Raan, 1998; Noyons, Moed & Luwel, 1999). In addition to applications for science mapping, Callon and his colleagues (Callon, Law & Rip, 1986a) has suggested that co-word analysis can be a useful tool in IR and knowledge organization. For example, the method can be employed as a means to classify documents in terms of their evolving centres of interest (He, 1999). From this point of view, it should be useful for the construction and updating of thesauri.

5.5.1 The asymmetric nature of co-word maps

Proximity measures are usually called 'indexes' in the literature on co-word analysis (He, 1999). The most widely used indexes are the *inclusion index*, which is used to detect hierarchies within a subject area, the *proximity index*, which is used to detect 'influential' low frequency words, and the *equivalence index*, which calculates the

association between word pairs. These indexes are asymmetric probabilistic measures (He, 1999), see Chapter 4. The use of symmetric association measures in co-word analysis, such as the extended Jaccard and the 'statistical coefficient', have been studied by Courtial (1986). The extended Jaccard can be used to measure the relative degree of overlap between 'semantic areas' of words within a given collection (Courtial, 1986). However, it cannot handle associations between low frequency and high frequency words very well, because it will have low values even when the low frequency words always appears together with the high-frequency word (Courtial, 1986). Consequently, Courtial (1986) argues, that this measure can only be used to explore overlap between medium-frequency words. Further, according to Courtial (1986), the 'statistical coefficient' is not usually applied in co-word analysis because the strength of association is not an important variable in the produced graphs.

Based on these indexes, entities are displayed, and perhaps grouped, in 'semantic network maps' (He, 1999). Semantic network maps are used to study the dynamic of science, and to understand the cognitive structures of a research area (Bhattacharya & Basu, 1998; Noyons, 1999; Salvador & Lopez-Martinez, 2000). The multivariate statistical techniques applied to determine and map the co-word structures are generally similar to those used in co-citation analysis (Courtial, 1986; Leydesdorff, 1987; Wilson, 1999; Wolfram, 2003). For example, an 'inclusion' map is designed to discover the central themes in a research area and depict their relationship to words that occur less frequently (Callon, Law & Rip, 1986b). A 'proximity' map is designed to discover connections between minor 'ideas' hidden behind the central themes in a research area (Callon, Law & Rip, 1986b). These two kinds of maps correspond to two general types of studies. The first type of study involves getting more information about a certain topic. The second category of study concerns the analysis of the links between topics (He, 1999).

5.5.2 Problems with co-word analysis

It is obvious that the quality of results from co-word analysis depends on a variety of factors, especially the quality of 'words', the scope of the document collection, and the adequacy of statistical methods for simplifying and representing the findings (Healey, Rothman & Hoch, 1986). For example, the use of descriptors in co-word analysis has created much debate in relation to the so-called 'indexer effect' (Callon, Law & Rip, 1986b; Turner et al., 1988; Whittaker, 1989; Law & Whittaker, 1992). The 'indexer effect' resembles the intra-indexer and inter-indexer consistency problems experienced in traditional indexing (Lancaster, 2003). Moreover, the use of title words have also been criticised as authors might choose their title words deliberately in order to address

a particular readership and produce an 'audience effect' (Whittaker, 1989). Finally, the use of full text words can create semantic problems similar to those experienced in automatic indexing; see Chapter 3. In a comparative study, Whittaker (1989) found that descriptor analysis generates a picture similar to, but substantially more detailed than that created by title words. The study does not show that either form of analysis is superior to the other.

Leydesdorff (1997) argue that words and co-words cannot map the development of science, because words change position not only in terms of the dimensional scheme of 'theory', 'methods', and 'observation results', but also change in meaning from one text to another. By using the distribution of words over sections, a clear distinction among 'theoretical', 'observational', and 'methodological' terminology can be made in individual papers but not at the level of the set of papers (Leydesdorff, 1997).

Courtail (1998) has commented on Leydesdorff's (1997) criticism. He claims words are not used as linguistic items to mean something in co-word analysis, but as indicators of links between texts, whatever they mean. In co-word analysis, words are chain indexes, allowing one to compute translation networks. What is important for co-word analysis is not the exact meaning or definition of a word, but the fact that this word is linked to word *X* in one case and word *Y* in another case (Courtail, 1998).

It is important to distinguish between two types of co-word analyses (He, 1999). The first type of analysis uses ordination techniques to assign words into clusters (e.g., Leydesdorff, 1987). The second type of analyses rests more upon the assumption that there is a cluster-type structure and consequently the algorithm is set to build those clusters link by link according to the relative frequencies of words and co-words in the documents (Callon, Law & Rip, 1986a). The goal of the former method is to identify, list, and measure distance between classes to create distinction rather than emphasising connection and continuity. In contrast, the goal of the latter is to describe a network of words and explore the qualitative character of the links between them by concentrating on, and tracing out, connections and crossroads in that network (He, 1999). Therefore, the two methods are actually doing different jobs and are appropriate for different purposes.

In the present dissertation, we explore the use of co-word analysis, with its asymmetrical proximity measures, to emphasize connection and structure in a network of candidate thesaurus terms, comprising 'concept symbols' and noun phrases. In order to reduce the semantic problems related to co-word analysis, we use co-citation analysis to group documents in topically related clusters. A cluster is equal to the 'intellectual base' for a given topic. For each cluster, concept symbols are identified in

the citation contexts of citing documents in the ‘research front’. Moreover, surrounding noun phrases in the citation contexts are extracted as well. Each citation context is then conceived of as a document and the noun phrases are used as units in a co-word analysis. Consequently, we transpose the co-citation cluster into a network of topically related noun phrases, centered on the cluster’s concept symbols, by use of co-word analysis.

The following section presents the bibliometric method of citation context analysis. This method is used to identify ‘concept symbols’ of cited references, in the citation context of citing papers.

5.6 Citation Context analysis

A citation context is that particular textual passage or statement within the citing document that contains a reference to a cited document (Small, 1978). As mentioned in section 5.1.2, we apply the concept of ‘citation context analysis’ albeit we deal with the context of references. A citation context analysis is any attempt to utilize these passages in a systematic fashion (Small, 1982). Small (1982) distinguishes between two methods of analyzing citations on the basis of the context that immediately surrounds them. The first method is to classify abstract features of the relationships between citing and cited documents, such as whether the cited document seems essential or inessential to the argument of the citing document and the citer’s attitude seems positive or negative. The second method is to read references as indicators of concrete topics, as if they stood for subject headings or descriptors (e.g., Garfield, 1974).

Generally, the first method requires the analyst to judge and code some aspect of the reference that is only implicit in the context, while in the second method, the analyst makes use of explicit words or phrases connected with the reference. The first method thus depends to a great deal on the expertise of the analyst, and the second method depends on the text in which the reference occurs. Both are forms of citation context analysis, but the second method, is also called ‘citation content analysis’ (McCain & Turner, 1989) or ‘content analyses of citation contexts’ (Small, 1982, p. 301). A third stream of research, that complements the other two, examines why authors make references (e.g., Liu, 1993). This research is known as studies of citer motivations. Research into citer motivations leads to different conceptions toward a ‘theory of citing’ (Wilson, 1999). An understanding of the motivations behind citing

is fundamental to application of bibliometric methods based on references and citations. Section 5.7 presents the two main positions in the debate about citer motivations. In the following two sections, we present in turn the above mentioned two citation context methods. For a detailed coverage, see Small (1982), Cronin (1984), and Liu (1993).

5.6.1 Classifying citations

Several studies have analyzed the context of citations in order to classify citations into different types (e.g., Chubin & Moitra, 1975; Moravcsik & Murugesan, 1975; Small, 1982; Liu, 1993). These studies are attempts to understand what people are doing when they cite. Over the years, about 20 have appeared from different disciplines with little regard for coordination (e.g., Liu, 1993; Baldi, 1998). Most of the citation classification schemes are idiosyncratic and hard to code³⁴. This means that replications of their use across literatures are scarce, with Cano (1989) and Hooten (1991) as exceptions. However, insofar as their types can be matched, Small (1982) does so for eight of the schemes, all from the 1970s. He thereby presents an integrative view of some of the major functions of citations (Small, 1982). It turns out that although each scheme has its unique nuances, there is a great deal of commonality underlying their different categories.

Small's (1982) results are important because they address the complaint, recurrent in academe, that evaluations based on citation counts are distorted by, for example, 'perfunctory' references (i.e., references not really necessary to the citing paper), negative references, which analysts supposedly do not properly take into account etc. (e.g., MacRoberts & MacRoberts, 1984; Seglen, 1998). As an example, Moravcsik and Murugesan (1975) coded references from papers in high-energy physics and found that 41 percent were 'perfunctory' and 14 percent negative. Small's (1982) results reveal that 'perfunctory' references are indeed quite common, ranging from 20 to 60 percent across the seven studies. Nevertheless, 'perfunctory' citing should not be dismissed as unimportant. It ties many new works to their discourse communities and, universally, holds literatures together (Hargens, 2000a; 2000b).

Small (1982) also found that in all studies the proportion of negative references, that is those in which the citing author criticizes or attempts to overturn the cited work, is relatively low. The result by Moravcsik and Murugesan (1975) seems to be the highest percentage found. This contradicts the myth that high citation counts cannot be trusted as indicators of quality or impact because they are inflated with negative

³⁴ The citation classification schemes in Small (1982) and their later counterparts all require close reading, domain knowledge, and expert judgement to apply (McCain & Turner 1989).

references (e.g., MacRoberts & MacRoberts, 1984). More recent studies (e.g., Vinkler, 1998; Case & Higgins, 2000), continue to report on low percentages of negative references. Even so, it is not at all easy to decide whether a reference is negative or not (Peritz, 1983b; MacRoberts & MacRoberts, 1984; Brooks, 1986; Teuffel, 1999). Cronin (1994) observed, moreover, that a reference could refer to the cited work at different levels of granularity, to an entire oeuvre, a cross-textual motif, a document, a passage, a single sentence, a phrase, a number. Any one of these levels can be criticized, and existing schemes do not take this adequately into account (Cronin, 1994).

Consequently, the great majority of references are confirmative; to some degree the citing work agrees with the cited work in its findings or judgements (White, 2004a). We should add that negative references do not necessarily indicate worthlessness; they often simply show controversy. The negative citers may themselves be wrong and the citees right. In any case, it is something of an achievement to have one's work noticed by others, even if negatively; work deemed substandard or negligible is seldom cited at all (Cole & Cole 1971; White 2001). It is much more common for a work to go uncited than to be negatively cited, and that accounts for the low percentages in the latter category. From the perspective of this dissertation, negative references are as good as positive ones. We are interested in the terminology that surrounds a reference in a citing paper. Citing authors may still refer to the same reference by use of agreed upon noun phrases, whether they agree with its knowledge claim or not. The following section presents the method of identifying 'concept symbols' in citing documents citation contexts.

5.6.2 *Content analysis of citation contexts*

There is a more promising line of analysis than the labour-intensive classification of implicit reference features just described. That is interpretation of explicit words and noun phrases in citation contexts as detected by humans or computers. It has long been known that the distinctive format of references like 'Borlund (2003)' marks a context for recognition. Noun phrases in such contexts subsume many knowledge claims (Budd, 1999). By use of co-citation analysis, Small (1978) established that highly cited documents symbolize concepts to those who cite them. While it has long been known, that when references are turned into citations they can be construed as subject headings (e.g., Garfield, 1974), different people may construe the same cited document differently. Small (1978) showed, however, that citing authors in chemistry tend to be both specific and highly uniform in the meanings they assign to cited documents, as revealed by the contexts of the references. In other words, scientists

tend to give earlier works consensual meaning by 'piling up' identical or similar phrases in the sentences in which their citation markers are embedded (Small, 1978). Consequently, when contexts show that citing authors have used a cited document to stand for a given idea more or less uniformly over many papers, the document has, according to Small (1978) attained the status of a 'concept symbol'. This implies that the cited document communicates a specific topic and resembles a subject heading or descriptor (Small, 1978). Following the notion of White & McCain (1989) presented in section 5.2.3, to a degree, cited references concentrate 'concept symbols', noun phrases, and consensual meaning, not only in the titles of citing papers, but also within their citation contexts (e.g., White & McCain, 1989; Small, 1986).

Consensual results reported in a series of studies by Small (1978; 1980; 1986) and Small and Greenlee (1980; 1986) supports the use of citation context analysis. Moreover, Cozzens (1985a) studied how two well-known papers, one in neuropharmacology and one in the sociology of science, were referred to in subsequent citation contexts. Cozzens (1985a) found that later papers initially cited the neuropharmacology paper for specific details of its experimental procedure and findings. However, over time such citations became less frequent, while general citations of its main knowledge claim became more frequent. Essentially, the knowledge claim became the concept symbol. By contrast, researchers rarely cited the sociology of science paper for its specific results or methods, or even its primary knowledge claim. Instead, authors cited what they believed was the general themes and speculations of the document (Cozzens, 1985a). The results of Cozzens (1985a) indicate that the specificity of concept symbols can vary between disciplines. Hargens (2000a, p. 859) provides another corroborating example. Citations to two foundational papers in economics and cognitive psychology gradually "... became highly standardized, almost as if authors shared a boilerplate for citing them (Hargens, 2000a, p. 859). According to White (2004a), similar examples can be added from almost any scholarly literature.

Small (1979) states that the identification process of concept symbols cannot be done entirely automatically. Since a computer cannot recognize unforeseen synonymy, the noun phrases that show consensus on what documents symbolize must therefore be recognized by a human reader (Small, 1979). O' Conner (1982; 1983) proposes that noun phrases from citation contexts could automatically be gathered as additional subject indexing for the cited document. The size of the recognition window may range from a few words on either side of the reference to multiple sentences (O'Connor, 1982; 1983). The window can even be extended to a specific section of a paper in which the reference appears, for example, the introduction or

methods section (O'Connor, 1982; 1983). Generally, references are most common in the introductory sections of a paper (e.g., Hargens, 2000a; 2000b). Quite a few researchers have linked cited documents to the sections in which they are cited, to see whether the citing works tend to use them in a consistent way (Herlach, 1978; Finney, 1979; Peritz, 1983b; Cano, 1989; McCain & Turner, 1989).

Cozzens (1982) shows that a scientific article may have a somewhat 'split identity'. She found a paper in economics that eventually came to symbolize two distinct concepts over the course of its citation history. Yet, one identity and one 'life' seem much more common (White & McCain, 1989).

Scientific or scholarly concepts occasionally become so ingrained in the scientific discourse that their sources are no longer cited, as when the *uncertainty principle* is discussed without reference to Heisenberg (1927) or *paradigm shift*, without reference to Kuhn (1962) (White, 2004a). In citation analysis, this phenomenon is known as 'obliteration by incorporation' (Garfield, 1977; Merton, 1979). How 'obliteration by incorporation' occurs is a complex process that, according to Cozzens (1989), has not been satisfactorily studied. On another front, there is some evidence that the scientific writings with the highest citation counts and the most longevity over the years tend to convey not theory or empirical results but methods (Garfield, 1977; Peritz, 1983a; McCain & Turner, 1989). Methodological works also tend to have the most consistent meanings as concept symbols (Small, 1982).

Small (1980, p. 194) has suggested that context analysis can be used as a semi-automatic method to produce "... a network whose nodes, lines and regions are labelled according to the consensual use of the [cited] documents". Small (1980; 1986) proposed that this method could be used to summarize and condense knowledge claims in particular research specialities, so-called 'synthetic specialty narratives' (Small, 1980; 1986). In two studies, Small (1985; 1986) showed that when a map of co-cited documents is created, the cited documents can be exchanged with their associated concept symbols, either as noun phrases, or as full-sentence knowledge claims. High co-citation links between documents on the map can also be re-expressed to bring out what citing authors see as the joint meaning of each pair (Small, 1980). Technically, the first process is citation context analysis, which counts recurrences of works as concept symbols in citing passages, while the second is co-citation context analysis, which counts recurrences of paired works as concept symbols. The latter is feasible only if the pairs are cited fairly close together (e.g., in the same paragraph) where their relationship is reasonably clear (White & McCain, 1989). Such maps combine in one network structure the knowledge claims implied by single documents

and by inter-document linkages. Each paper is a node labelled by citation context analysis, and lines that connect the nodes are labelled by co-citation context analysis.

The fact that cited documents can indicate topicality, sometimes very strongly, has obvious implications for IR and knowledge organization. Most important in relation to the present dissertation is the research by Rees-Potter (1987; 1989) on semi-automatic thesaurus maintenance. The objective of Rees-Potter's (1987; 1989) study was primarily to investigate if the bibliometric methods of citation, co-citation, and citation context analysis were able to identify candidate thesaurus terms in two social science domains, sociology and economics. The results indicated that highly cited and co-cited documents did act as concept symbols, thus the terminology used in citing paper's citation contexts reflected concepts of specialty area. But in the case of Rees-Potter (1987; 1989), a high proportion of the highly cited and co-cited documents were monographs. As Rees-Potter (1989, p. 687) states "[i]t may be that there is a difference between citation contexts and citation terminological analysis of monographs and periodical articles in the terminology these different concept symbols would produce". This is supported by Small (1978), who notes that a concept symbol for a monograph is often more complex than one for journal paper. Consequently, we investigate a specialty area where a large majority of citing and cited documents are journal papers. Moreover, Rees-Potter's (1987; 1989) application of the citation context analysis was time consuming as it was done manually. Overall, the research by Rees-Potter (1987; 1989) indicates the value of bibliometric methods for selection of candidate thesaurus terms beyond the traditional term co-occurrence methods. Some of the conditions and problems verified in her study are addressed in the present dissertation.

The following section ends the presentation of bibliometric methods, by looking at the theoretical foundations upon which citation analysis rests, that is, the motivation behind citing.

5.7 Citer motivations

The use of references and citations as units for bibliometric analysis raises the fundamental question of the motivation behind citing – why do authors make references? (e.g., Brooks, 1988; Bonzi & Snyder, 1991; Case & Higgins, 2000) According to White (2001, p. 98), "... a quarter-century of concern with the complexity and variability of citer motivations has managed to obscure a most basic fact: the constant underlying all citing is relevance to a claim, as perceived by the

citer”. White (2001, p. 98) point to results from White and Wang (1997) and Baldi (1998) when he further notes that perceived topical relevance, shared subject matter, is the most common reason for citing in science or scholarship. Consequently, the global reason for citing is the perceived relevance of a document to a work in hand, but motives that are more specific can be elicited through interviews or inferred from the citation context of documents. A number of researchers have listed such motives (e.g., Garfield, 1965; Hodges, 1972; Peritz, 1983b; Brooks, 1988; White & Wang, 1997; Wang & White, 1999).

The oldest list of ‘citer motivations’ comes from Garfield (1965, p. 189). Apparently, this list is inferred from utterances or verbal phrases in citation contexts (White, 2004a). The list comprises 15 reasons for citing, presented in Table 5.1:

TABLE 5.1. *Garfield’s 15 reasons for citing (Garfield, 1965, p. 189).*

1. paying homage to pioneers;	9. alerting researchers to forthcoming work;
2. giving credit for related work;	10. providing leads to poorly disseminated, poorly indexed, or uncited work;
3. identifying methods, equipment, etc.;	11. Authenticating data and classes of fact;
4. providing background reading;	12. identifying original publications in which an idea or concept was discussed;
5. correcting one’s own work;	13. identifying original publication or other work describing an eponymic concept or term;
6. correcting the work of others;	14. disclaiming work or ideas of others;
7. criticizing previous work;	15. disputing priority claims of authors.
8. substantiating claims;	

These reasons imply speech acts on the part of citing authors. This is in contrast to the adjectival categories of classification schemes, which capture the judgement of coders rather than the intentions of citing authors (e.g. Moravcsik & Murugesan, 1975; White, 2004a). Contrary to Garfield (1965), Hodges (1972) derives a list of categories from interviews with citing authors. Hodges’ categories are not cast as verbal phrases that reflect the speech acts of citing authors, but as adjectives that reflect Hodges’s complicated judgements of reasons for citing (Brooks, 1988). Peritz (1983b) found Hodges’s list of categories difficult to apply and simplified it for use on citation contexts in five social science journals. By far the most frequent category found by Peritz (1983b) was ‘setting the stage’, which occurred overwhelmingly in the ‘introduction’ section of scientific articles. This is in accordance with Swales’s (1990, pp. 137-66) account of rhetorical moves in this particular section of physics papers, and Paul’s (2000) study of rhetorical use of references in chaos theory.

The most elaborate continuation of Hodges’s style of research, in which citer motives are learned through interviews and then fitted into a classification scheme, is White and Wang (1997) and Wang and White (1999). Their work commendably links earlier motives for retrieving and reading documents with later motives for citing or

not citing them. White and Wang (1997) devised schemes to cover the remarks of citing authors on: 1) the *contributions* that cited documents made to their new writings; 2) the *criteria* they used in choosing documents to cite; and, 3), their *meta-level documentation concerns* (e.g. completeness of reference lists, propriety of self-citation, and anticipated expectations of referees). The *contributions* of the first scheme are reasons for citing that emerge from contexts in the new document, such as providing an analogy or justifying an idea, whereas the *criteria* of the second scheme are characteristics of the *cited* document, such as its topic or recency. White and Wang's (1997) main conclusion is that most often the content of the earlier cited document is relevantly related to that of the later citing document.

5.7.1 Reward or persuasion

Case and Higgins (2000) point out that research into citer motivations has mainly been focused on validating either the 'reward hypothesis' or the 'persuasion hypothesis'. These are the two main positions that dominate the overall conception of citer motivations (Baldi, 1998; Wilson, 1999).

The 'reward hypothesis' claims, that when authors make a reference, they are mainly rewarding others for the use of their intellectual property. Hence, citing authors are mostly influenced by the worth as well as the cognitive, methodological, or topical content of the cited articles (Garfield, 1965; Kaplan, 1965; Merton, 1968; 1973; Price, 1986). A collateral belief is that citing authors acknowledge their debts fairly (Cole & Cole, 1973). Proponents of this hypothesis adhere to a 'normative theory of citing', which is rooted in Robert Merton's norm-based sociology of science (e.g., Merton, 1973). The assumption is that science is a normative institution governed by internal rewards and sanctions (Merton, 1973). Scientists are believed to exchange information, in the form of publications, for recognition, in the form of awards and citations (Hagstrom, 1965). Hence, the notion is that citing authors are, in effect, conditioned to follow the 'norms of science' in general and the 'norms of citation practice' in their chosen fields in particular (Merton, 1973; Cole & Cole, 1973; Garfield, 1979a; Price, 1986; Cole, 1992). As an example, the first two motives in Garfield's (1965) list presented above capture aspects of reward. To some degree, this conception of citer motivations must be accepted in order to do citation analysis. Otherwise, the connotations about the 'communicative functions' of references and citations, which are studied by bibliometric methods, become irrelevant (e.g., Wilson, 1999). Indeed, empirical evidence suggests that references and citations are valid entities for at least relational citation analysis. Before we present some of this

evidence, we first look at the opposite view, which is highly critical of citation analysis.

The contrasting position, which is much influenced by Gilbert (1977), is the ‘persuasion hypothesis’. The ‘persuasion hypothesis’ accredits citing authors with more egotistical reasons for their citing behaviour (Borgman & Furner, 2002). Authors cite in ways that they believe will most benefit their own personal goals. According to Gilbert (1977, p. 116), “... authors preparing papers will tend to cite the ‘important and correct’ papers, may cite ‘erroneous’ papers in order to challenge them and will avoid citing the ‘trivial’ and ‘irrelevant’ ones. Indeed, respected papers may be cited in order to shine in their reflected glory even if they do not seem closely related to the substantive content of the report.” Consequently, Gilbert (1977) implies that ‘lower-status’ citing authors invoke ‘famous’ earlier works, even when those works are not strictly relevant, to impress readers and make their own new work more convincing. Gilbert (1977) hints, that it may be ‘social attachment’ rather than the ‘scientific truth’ that persuades (White, 2004a). This is a social constructivist position. Social constructivists do not see science as governed by a set of internally sanctioned norms. Instead, they argue that scientific knowledge is socially constructed through the manipulation of political and financial resources and the use of rhetorical devices (Bloor, 1976; Knorr-Cetina, 1981; Latour & Woolgar, 1979; Latour, 1987). References are one rhetorical device that scientists employ to provide support for their papers and convince readers of the validity of their claims (Gilbert, 1977; Latour, 1987). According to the social constructivist position, the factors that influence references have more to do with the location of a cited document’s author within the stratification structure of science than with the intellectual content of the document itself (White, 2004a). Obviously, if this is true citation analysis becomes flawed and pointless.

Latour (1987, pp. 33-34) argues that in order to put up a ‘persuasive front’, citing authors essentially fake their scholarship. According to Latour (1987), citing authors often misrepresent and distort the works they allude to by twisting the meaning to suit the citing author’s own ends. In addition, Latour (1987) argues that citing authors also have a habit of citing documents regardless of their content. This habit resembles Gilbert’s (1977) notion of ‘reflected glory’ and is called name-dropping by White (2004a). White (2004a) questions whether Gilbert (1977) and Latour (1987) actually believe that ‘distorted’ or ‘reflected glory’ references persuade anybody if their true nature is discovered. The real question, to White (2004a) is what persuades? Is it the qualities of the arguments of which the references are part, or is it the reputation of the cited author and his work?

Yet, on one level, Gilbert's (1977) argument about persuasion is true. It is commonly agreed that scientists and scholars write to persuade and that that intention shapes their choices of what to cite, thus citations are used as rhetorical devices (Edge, 1979; Brooks, 1985; 1986; Cozzens, 1989; White, 2004a; 2004b). For example, Cozzens (1989) sensibly argued that citing authors refer both as a reward to others for their intellectual contributions and as a tool of persuasion.

However, this is not what is meant by the 'persuasion hypothesis' outlined above. The respondents in Brooks's (1986) interview study almost certainly understood persuasion as a fully justifiable goal in which references help to build a case. They were not admitting to practices like manipulative name-dropping or distortion of cited authors' meanings as indicated in the 'persuasion hypothesis'. The qualification is important because Brooks' (1986) small survey provides the only hard support for persuasion as a motive in citing (White, 2004b). In fact, there is almost no evidence that a significant number of scientists should seek 'reflected glory' from references that are included 'just for display', Gilbert (1977) and Latour (1987) give no examples themselves.

Moreover, the evidence from statistical studies go against the 'persuasion hypothesis'. In a discussion of the hypothesis, Zuckerman (1987, pp. 334-335) maintains that, if citing authors seek to increase the persuasiveness of their own works merely by referencing 'respected' or 'authoritative' documents whenever possible, then a large fraction of citations overall should go to papers that are already well cited. In Zuckerman's study (1987) this is not the case. If reference by authority significantly guided reference practice, then a much higher proportion of all citations would go to these authoritative papers. Instead, almost two-thirds of all citations went to papers cited only once in five years (Zuckerman, 1987). This result may seem inconsistent with the 'Matthew effect', which predicts that authors who are already well cited will get a disproportionate share of additional citations (Merton, 1968; 1988). Yet, Zuckerman (1987) is not denying the 'Matthew effect'. She is simply noting that citations to 'common' authors far outnumber citations to the 'famous' authors.

As an alternative to interview based studies like Brooks' (1986), researchers have gathered data on citing authors in particular fields and investigated whether they tended to overcite authors whose citation counts are already high (Stewart, 1983; Baldi, 1998; Van Dale & Henkens, 2001). The results from these studies all undermine the 'reflected-glory' argument (Stewart, 1983; Baldi, 1998; Van Dale & Henkens, 2001). In his study on the astrophysics research area, Baldi (1998) included not only the characteristics of cited articles but also the characteristics of the citing

articles. He discovered that authors “... are most likely to cite articles that are relevant to their work in terms of subject, recency of knowledge, theoretical orientation, and seem to have little concern with the characteristics of authors who wrote them” (Baldi, 1998, p. 843).

The data used in Stewart (1983), Baldi (1998), and Van Dalen and Henkens (2001) were gathered unobtrusively from publications, not from interviews with citers. Nevertheless, four recent interview-based studies: Shadish et al. (1995), Vinkler (1998), Wang and White (1999), and Case and Higgins (2000), all tend to support the publication-based findings. The vast majority of reference decisions made by a citing author are motivated by what Vinkler (1998) calls ‘professional’, as opposed to ‘connectional’, concerns. Generally, the perceived relevance and quality of precedent work, its scientific merit, rather than the reputations or personal connections of the citees govern reference practice. Those who believe that successful citing is really a matter of lining up the ‘right’ names have produced no comparable body of evidence (White, 2004a, p. 111).

It seems obvious that scientists refer to prestigious authors and respected works when trying to write persuasively. The effects of author prestige reported in the publication-based research, although small, are probably quite real. The question is, rather, whether use of such authors is warranted or unwarranted. Widespread unwarranted use is what social constructivists have not shown (White, 2004a). If name-dropping or unwarranted use were widespread, author effects in publication-based studies would be stronger than they are, and interview studies “... would likely turn up a few whistle-blowers who want to give the game away” (White, 2004a, p. 111). The phenomenon would also presumably be detectable in the writings of the constructivists themselves (e.g. Latour 1987). Curiously, there is no evidence of empty name-dropping on their part; according to White (2004a, p. 111) they play the citation game straight. According to the statistical studies mentioned above, that is what other scholars and scientists also do most of the time. What persuade are the statements to which reference contribute, not the ‘reflected glory’ of cited names. As White (2004a, p. 111) states “[h]ow else [...] can one account for cases in which the same authority is adduced in the context of opposing arguments? While Luukkonen (1997) cites Latour’s *Science in Action* admiringly, Baldi (1998) cites the same book critically. One could hardly claim that both papers are equally convincing because they both draw lustre from the same famous work”.

As a result, we find it reasonable that a document is most often cited in another document because it provides information relevant to the performance and presentation

of the research, such as positioning the research problem in a broader context, describing methods used, or providing supporting data and arguments (Wilson, 1999). There may be other reasons for citing, but 'subject matter relatedness' is the primary one. Moreover, the dyadic relationship between citing and cited documents shows that references and citations are different phenomena (Small, 1978; Baldi, 1998; Wouters, 1999). Individual references are *not* equal; their functions may differ and references are perhaps given with different motivations in mind. But, when references are transformed into citations they become another entity, as accrued citations are equal (Price, 1970; Smith, 1981; van Raan, 1998; Wouters, 1999). This makes citation counting interesting, as most documents have none, some documents hardly any, and few documents have many citations. The fact that some documents or their aggregates accrue many citations indicates some form of importance acknowledged to these documents. Hence, we find it sensible in relational citation analysis to assume that the more a document is cited from a subsequent body of literature, the more the document influences the reported research. Further, if two documents are jointly cited by another document, they jointly contribute to the content and impact of that research document, and are associated by their role in that research document. Accordingly, the more two documents are co-cited from a body of literature, the greater is the association of their content, in the opinion of the authors of that body of literature (Wilson, 1999).

This validates the use of relational citation analysis and is the assumption behind the bibliometric methods applied in this dissertation. This is not to say that there is no problems inherent to relational citation analysis, there are, but most of these are of a technical nature that can be reflected upon during analysis (e.g., King, 1987; Wilson, 1999).

5.8 Summary

The present chapter has presented the bibliometric methods we apply in the semi-automatic thesaurus construction approach introduced in the following chapter. We have presented the motivation behind their use, and we have discussed their different applications and assumptions.

By use of an entity-relationship diagram, Figure 5.2, we illustrated and discussed the resemblance between bibliometric methods based on co-occurrence analysis and generic term co-occurrence methods, presented in Chapter 3 and 4. The notion of co-occurrence is the same, 'semantic or subject relationship' between two document entities. Both types of methods use vector space representation, weighting, and

ordination techniques. Where they differ is on units of analysis. This means that bibliometric methods depend on probability distributions of document entities, most notably references or citations, and term co-occurrence methods depend on probability distributions of terms in natural language text.

The dyad relationship between cited and citing documents is interesting for thesaurus construction (Rees-Potter, 1987; 1989). We may cluster cited documents that treat similar topics without being dependent on them sharing common terminology. Such clusters are ‘intellectual bases’ to more recent ‘research fronts’. Moreover, the clusters also represent one or several concepts that could be of interest for indexing. The cited documents in these clusters may act as concept symbols to citing authors, and give names to clusters. Knowing that concept symbols emerge from consensus usage, we can follow the indirect link from a primary cited reference entity to its relative citing paper entities, and investigate the latter’s citation contexts for consensus terminology, which may reflect upon the concept(s) in the cluster. This way we may be able to identify recently agreed upon terminology in the ‘research front’, which may be suitable as candidate thesaurus terms. Further, bibliometric methods are used to investigate the ageing patterns of ‘concept symbol containers’ to see if they can indicate terminological changes in the specialty. Finally, terminology is subjected to co-word analysis in order to depict connectivity and structure in a conceptual network.

The following chapter outlines the steps and components of our semi-automatic thesaurus construction approach based on bibliometric methods. Moreover, the chapter also introduces the research specialty we use as a case for investigating the appropriateness of the chosen bibliometric methods for thesaurus construction.

6. A semi-automatic thesaurus construction approach

It is our opinion, as stated in Chapters 1 and 2, that manual thesaurus construction is mandatory due to the complexities and dynamics inherent in languages. However, we fully acknowledge the complexities of construction and maintenance of such controlled vocabularies. We therefore find it beneficial and necessary to combine manual approaches with automatic approaches due to the complementarity of the two approaches (Soergel, 1974; Aitchison, Gilchrist & Bawden, 2000; Lancaster, 2003). Automatic approaches reduce manual workload, but are deficient if used in isolation. Thus, a hybrid of less resource demanding, semi-automatic construction methods, are attractive, as noted by Soergel (1974).

The overall aim of the research project presented in this dissertation is to verify the applicability of bibliometric methods as a supplement to intellectual manual based construction and maintenance of thesauri. The dissertation builds on former research in the area (Rees-Potter, 1987; 1989). We reintroduce, modify, and expand former research approaches in that we develop an *exploratory methodology*. The exploratory methodology is introduced in the present chapter as a semi automatic approach to thesaurus construction and maintenance. For this reason, the focus of the dissertation is the development of the methodology and to a lesser degree its evaluation and validation. The proposed approach is an *exploratory methodology* with emphasis on bibliometric methods, but which also include manual and automatic thesaurus construction aspects. The semi-automatic thesaurus construction approach is conceived of as a tool that may help in manual intellectual thesaurus construction and maintenance. The approach is therefore not a thesaurus construction process in itself, as is the case in automatic thesaurus construction approaches. Consequently, the present dissertation sets out to investigate the extent to which a number of bibliometric methods included in the exploratory methodology can identify candidate thesaurus terms, help identify equivalence, hierarchical, and associative relationships, as well as monitor or identify terminological and conceptual changes in a given specialty area. The object of the bibliometric analyses is the communicative structure within a scientific specialty area. Our focus is therefore on written scientific knowledge structures. As presented in Chapter 2, other knowledge structures can and should be consulted in thesaurus construction.

The chapter serves as a crossover between the theoretical and the empirical based parts of the dissertation. The main function of the chapter is to describe the foundation behind and purposes of the exploratory methodology and its individual components. We do this in relation to the preceding chapters' introduction and discussion of manual and automatic thesaurus construction approaches, as well as bibliometric methods.

6.1 The explorative methodology for semi-automatic thesaurus construction

The Chapters 3 and 4 explained that the relationships that exist between terms based on their meaning result in certain statistical patterns of occurrence and co-occurrence of these terms in text. According to Soergel (1974, p. 449), "... we should be able to conclude from observed statistical patterns of the occurrence and co-occurrence of terms what these conceptual relationships are". But, Soergel (1974, p. 449) also states, "[s]tatistics should not take precedence over human judgement in the evaluation of vocabulary, but these studies and other provide the basis for some useful decisions". This implies that the identification of terms and relationships by automatic methods should be considered as a kind of pre-processing of written open-ended sources (Soergel, 1974). The results of this pre-processing are subsequently used in further steps of manual thesaurus construction.

In Chapter 5, we introduced a number of bibliometric methods and demonstrated that they resemble traditional automatic methods used for thesaurus construction. The key difference between bibliometric and traditional automatic methods is the document entities they use as unit of analysis. Contrary to automatic thesaurus construction methods, bibliometric methods can cluster semantically related documents without being dependent on the terminology used in the individual documents. Chapter 5 also showed that the dyad relationship between cited and later citing documents is a strong indication of a semantic relationship between two documents. Moreover, in Chapter 5 we illustrated how some cited documents may act as concept symbols in later citing documents within a specialty area. Concept symbols and the dyad relationships between cited and citing documents may be exploited for thesaurus constructions purposes. Cited documents are embedded in the document structure of citing papers. They appear as reference markers, and tend to clump together in the text of citing papers. The notion that some cited references act as concept symbols, and the fact that they clump in the structure of citing documents, makes reference markers important for automatic thesaurus construction in the present semi-automatic approach. Reference

markers function as signs of concept symbols and their immediate surrounding contexts becomes the target for term selection. In addition, when terminology is attached to cited documents, it may be possible to monitor the terminological and conceptual changes within a specialty area by studying the citation history of such cited references. Consequently, the terminology in citing papers' citation contexts is explored for semi-automatic thesaurus construction in the present proposed methodology. This means that the methodology does not base its 'vocabulary organisation' on distributions of terms in natural language text, but instead on the distributions of citations and co-citations within a body of citing papers.

This is the basis for the exploratory methodology for semi-automatic thesaurus construction. The conjecture is that the implicit relation between a cited reference and the terminology used in its related citing papers can bring some new aspects to thesaurus construction, different from those of traditional automatic statistical methods. The exploratory methodology includes five components.

- The first component concerns the creation of a text corpus, which implies identification of reliable sources that constitute the basis for thesaurus construction. The *data set isolation method* developed by Ingwersen and Christensen (1997) is used for this purpose.
- The purpose of the second component is to structure and map the text corpus by use of *document co-citation analysis*. This creates clusters of cited documents.
- The third component uses *citation context analysis* to investigate the potential conceptual relationship between these cited documents and their citing papers in the later research front. In addition, *shallow parsing* is applied to extract candidate thesaurus terms from the citing papers' citation contexts.
- The fourth component constructs conceptual networks from the extracted terminology by use of *co-word analysis*. These networks are investigated to see if they can indicate equivalence, hierarchical, or associative relationships.
- Finally, the fifth component explores the use of retrospective concept and citation profiles to monitor possible conceptual and terminological changes over a time period in a specialty area.

The first component is the basis for the other four. The second and third components are closely connected, and both relate to the first research question, presented in Chapter 1. The fourth component is dependent on the preceding ones, not *vice versa*, and relates to the second research question, presented in Chapter 1. The fifth component requires as basis the procedures of the second and third components, and

relates to third research question, presented in Chapter 1. In her study, Rees-Potter (1987; 1989) investigates what might correspond to the second, third, and fifth components, albeit with other methodical means. We should emphasize that the application of methods in the exploratory methodology is not definitive. The composition of the methodology and the application of its components is a result of our interpretations and ideas. As a result, the methodology becomes exploratory in nature.

Below we present the rationale and application of the five components. In addition, section 6.1.6 presents the rationale behind the selection of the specialty area used in the case study. Chapters 7 and 8 explore the methodology in a case study in relation to the research questions posed in Chapter 1.

6.1.1 First component: Creation of a sample text corpus

The first basic component of the methodology is to establish a *sample* text corpus, which is the source of documents used for thesaurus construction. As stated in Chapter 4, automatic thesaurus construction approaches most often use a full text, subject heterogeneous, document collection as the text corpus (Elkalifa, 1991). The ISI approach to macro-level co-citation studies uses the annual file of bibliographic records from SCI® as their text corpus. Rees-Potter (1987; 1989) uses a subset of the annual file of records from Social Science Citation Index® for her co-citation study. The subset corresponds to the individual subject categories of economics and sociology. From this, we can surmise that the decision about text corpus depends on the purpose of analysis. We focus on a micro-level study of a specialty area for thesaurus construction. As discussed in Chapter 4, one way to diminish the terminological problems inherent in a text corpus is to ensure that the text corpus is homogenous in relation to subject matter (e.g., Elkalifa, 1991). Our strategy is, therefore, to create a relatively small, subject specific and homogenous, text corpus. But, contrary to other approaches, the text corpus comprises a 'merged sample' of document representations (Schneider & Borlund, 2002).

The motive behind the establishment of a *sample* text corpus is to utilize the unlike representation of identical documents in *two* different databases, a citation index and a domain dependent bibliographic database. In the present case, we use the SCI® and the MEDLINE® database. MEDLINE® is a domain dependent database for the specialty area of *periodontology* within dentistry, which is used in our case study. The bibliographic records of citing documents from the citation index provide the set of references needed for bibliometric analyses. The MEDLINE® records, for the identical citing documents, provide document entities, such as MeSH® descriptors,

abstract and title noun phrases. MeSH® descriptors is of special interest as this vocabulary is generally considered a state-of-the-art indexing language, which is applied consistently in MEDLINE®. As presented in Chapter 5, White and McCain (1989) argue that certain primary documents entities have the ability to concentrate related relative entities. For example, it is suggested that cited references concentrate descriptors and title terms in citing papers (White & McCain, 1989, p. 127). Conversely, descriptors concentrate amongst other entities cited references. The implication is that any ‘content indicator entity’ can be meaningfully translated into associated entities in order to depict content relationships (White & McCain, 1989). This concentration behaviour in combination with the high regards that pertains to the MeSH® language, enable the application of MeSH® descriptors for a variety of different comparative and evaluative purposes in connection with the exploratory methodology.

Metaphorically speaking, we unite the two different bibliographic records into a single ‘enlarged document representation’. The ‘enlarged document representation’ contains specific document entities, otherwise only located in the records of the citation index, or the records in the domain dependent database. In practice, this procedure is carried out through online searching. ‘Overlapping document sets’ are created by the merger of the identical citing documents’ different bibliographic records in MEDLINE® and SCI® (Ingwersen & Christensen, 1997; Schneider & Borlund, 2002). This is done by use of the data set isolation method described by Ingwersen and Christensen (1997). Besides the characteristics of the data set isolation method, the sample set is also defined by subject and time constraints imposed in online searching. Obviously, the subject is periodontology, and the time interval is set to one-year samples of citing documents.

FIGURE 6.1. ‘Sample of overlapping set of citing documents’.

Figure 6.1 illustrates the *sample* set of overlapping citing documents in the intersection of the Venn-diagram.

A domain dependent bibliographic database usually treats subject matters more thoroughly than a citation index. Primarily this is so because the domain dependent database index more documents on individual subjects. It would therefore seem obvious to use a domain dependent bibliographic database, or a subset of it, as the text corpus for thesaurus construction. However, the merger of records from two databases described above reduces the set of available documents for thesaurus construction. The merger bring about that only documents, which are indexed in both databases are considered for thesaurus construction. Therefore, by merging the two different databases, we only allow documents into the text corpus if they are both indexed in MEDLINE® and SCI®. The merger enables the creation of 'enhanced document representations'.

The required appearance of documents in the citation index excludes a number of documents from the text corpus as they only appear in the domain dependent database. Compared to an approach that applies the domain dependent database as text corpus, the present approach in effect extracts a *sample* from the domain dependent database and uses this as a text corpus for thesaurus construction. The domain dependent database, in our case MEDLINE®, is regarded as the *population* and the 'overlapping document sets' are considered the *sample*. This is illustrated in Figure 6.1, where the domain dependent database contains more documents on a subject than the citation index. As a result, the domain dependent database is treated as the 'population' against which the sample is measured. Moreover, as the criteria for sampling is the document indexing overlap between two databases, it cannot be considered as a random sampling.

In order to validate whether the *sample* text corpus resembles its population as to emphasis of content, corpus similarity methods from NLP are applied (e.g., Biber, 1990; 1993; Dunning, 1993; Kilgariff, 1996; 1997; Kilgariff & Rose, 1998). In corpus linguistics, full text word distributions are mainly used to compare the homogeneity of two text corpora. For the purpose of thesaurus construction, we are interested in comparing the distributions of MeSH® descriptors in the sample to the population. The purpose is to establish whether the sample and population resembles each other in relation to subject matters covered. In other words, to investigate whether the sample text corpus can be said to cover and weight the same subject matters as the population, even though the former contains fewer documents. If the sample and population resemble each other in relation to subject matters covered, then we can expect thesaurus construction results from the sample to be valid

representations of the population. The results are therefore not an outcome of possible subject skewness produced by the sampling procedure. According to White & McCain (1989), we can expect the distribution of MeSH® descriptors to reflect the concentration of cited references. Hence, if the two MeSH® distributions resemble each other in coverage and relative frequency, then we can expect that documents, exclusively appearing in the MEDLINE® population, will not significantly alter the distribution of cited references, if they were incorporated into the sample text corpus. The creation of the sample text corpus is presented in Chapter 7.

6.1.2 *Second component: Creation of 'concept groups' by use of document co-citation analysis*

In Chapters 3 and 4, we comprehensively depict how the document entity of *terms* can be used to 'structure a text corpus' for automatic thesaurus construction. To 'structure' a text corpus in this context, means to utilize the salient semantic and topical structures that can be derived from counts of individual document entities within the corpus, as presented in Chapters 4 and 5, as well as illustrated in Figure 5.2. Besides automatic thesaurus construction, 'corpus structuring' is also the aim of literature mapping. In the exploratory methodology, we apply literature mapping as an indirect approach to thesaurus construction.

Automatic thesaurus construction uses term co-occurrence analysis to establish a 'structure' from the salient associations between terms in a corpus. Structuring is a combination of term association and vocabulary organization as described in Chapter 4. The vocabulary is typically structured or 'organized' into loosely defined concept groups of semantically related terms, sometimes referred to as 'thesaurus classes', by use of ordination techniques. The automatic thesaurus construction process requires that terms be selected *prior* to vocabulary organization. This is apparent, as terms are the primary entity used for vocabulary organization. The major effect of this is that term association and vocabulary organization (structuring), most often depend on direct first order co-occurrence analysis. Figure 6.2 below, is an extract from Figure 5.2, which highlights the primary and relative document entities used in automatic thesaurus construction.

FIGURE 6.2. *Direct link between primary and relative document entities in automatic thesaurus construction.*

Note that automatic thesaurus construction is based on *direct* co-occurrence analysis of terms in the documents of a text corpus (i.e., collection). Such documents correspond to citing papers in the terminology of bibliometrics. Moreover, the mention of *Zipf's Law* indicates that term selection is based on the distribution of terms in the natural language text of these documents. In Chapter 4, we discussed some of the limitations of this generic process. The most noticeable, is the difficulty of direct first order co-occurrence analysis to associate synonymous and near-synonymous terms, as such terms rarely co-occur frequently in the same documents.

In relation to semi-automatic thesaurus construction, it motivates to investigate whether indirect approaches are useful in the creation of concept groups. In this context, an *indirect approach* means not to be dependent on the direct co-occurrence relation between terms in documents for vocabulary organization. The notion of indirect approaches is therefore related to second order co-occurrence analysis and LSI, as described in Chapter 4.

The methodology we explore for semi-automatic thesaurus construction, apply an indirect approach in order to create concept groups. We know from bibliometrics that document co-citation analysis is an appropriate bibliometric method for literature mapping, i.e., clustering of topically related cited documents (Small, 1973). Clusters of co-cited documents serve as the 'intellectual base knowledge' to citing papers in more recent research fronts. The individual cited documents within the clusters are comparable to the notion of 'exemplary documents' suggested by Blair and Kimbrough (2002), in that, they represent the key concepts, methods, or experiments, which researcher build on in a research front of a specialty area (Small, 1978). Some of the cited documents symbolize the same content to a majority of later citing authors, which result in a consensus usage of terminology when citing those documents (Small, 1978). Therefore, highly cited documents may act as concept symbols to citing authors in a research front. Moreover, Rees-Potter (1987: 1989) indicates that clusters

of highly co-cited documents can be treated as concept groups for thesaurus construction purposes, at least in the fields of sociology and economics.

The present second component of concept group creation clusters a number of subject related ‘exemplary documents’ through co-citation analysis. Subsequently, the implicit link that goes from the cited references to their markers in the text of later citing papers is utilized. The marker indicates the citation context, which is the target for term selection. In this way, terminology is attached to a cited reference. It is the assumption that when terminology is related to the usage of references in citing documents, then we can expect the terminology to be contextual and subject specific, similar to the notion of term clumping presented in Chapter 3 (Bookstein, Klein & Raita, 1998). Accordingly, the creation of concept groups is dependent on direct co-occurrence analysis of references in citing documents and *not* direct co-occurrence analysis of terms in citing papers. The *indirect link* between cited references and their conceptual symbol in citing papers is illustrated in Figure 6.3; likewise an extract of Figure 5.2.

FIGURE 6.3. The indirect approach to concept group creation.

Note that the cited reference entity is used as the primary entity for ‘vocabulary organization’. Due to the implicit link between cited references and concept symbols (words entity) in the citation context of citing papers, we can select terminology and assign it to the concept groups (i.e., ‘intellectual base clusters’). This implies that we alter the traditional steps in automatic thesaurus construction, as we do not commence with term selection based on some probability distribution. Instead, we start by establishing a framework of potential concept groups, i.e., clusters of highly co-cited documents. Subsequently, the succeeding term selection component of the methodology will name the concept groups and assign candidate thesaurus terms to these concept groups.

The second component of the exploratory methodology therefore investigates the ability of document co-citation analysis to group related potential concept symbols in corresponding concept groups. The component is comparable to vocabulary organization in traditional automatic thesaurus construction approaches. As noted in Chapter 4, single linkage clustering is often used in automatic thesaurus construction and macro-level co-citation studies. Indications suggest that complete link clustering may be more appropriate for the creation of solid concepts groups (e.g., Lancaster, 2003). Hence, we explore the use of complete link clustering in a micro-level co-citation study, as we want to ensure that the cited documents that represent the individual concept groups are solidly related. In this connection, we explore the usage of different proximity measures to investigate their influence upon clustering. Finally, to obtain a more intuitive understanding of cluster compositions, i.e., their structural relations, the Pathfinder network algorithm (Schvaneveldt, 1990) is imposed on the cluster results. The Pathfinder network visualization enables investigation of the internal structure of the concept groups, as well as their external relations to each other. Pathfinder networks reduce links in a network to the most salient or 'important' ones (Schvaneveldt, 1990).

The following section presents the third component of the exploratory methodology, that of term selection. As indicated in this section, the construction of concept groups is necessary before we commence the term selection process. Yet, the term selection process is important to the present component, as it eventually give names to the concept groups and assign to them candidate thesaurus terms.

6.1.3 Third component: Term selection and identification of 'concept symbols' by use of citation context analysis and shallow parsing

In automatic thesaurus construction, term selection and subsequent co-occurrence analysis is based on probability distributions of terms in natural language text. In Chapter 3, we discuss a number of the most widely used distributions. What is more, we show how weighting schemes based on these distributions adjust the representation and discrimination abilities of index terms. Hitherto, the best trade off is found to be terms with middle-range frequencies. Chapter 3 also points out that there exist an alternative or supplement to term distribution models. Document structures seem to be very appropriate in relation to index term selection.

The term selection component of the exploratory methodology focuses on document structures rather than term distributions in order to identify candidate

thesaurus terms. Nevertheless, simple frequency counts of terms and concepts are applied as a part of the term selection in the methodology.

The part of the ‘document structure’ focused on for term selection, is the loosely defined text window in scientific papers that surrounds reference markers. This is the citation context as presented in Chapter 5. The assumption is that terminology used in the citation contexts of citing papers reflects concepts of a specialty area as outlined in Chapter 5 and above in section 6.1. Note that it is the terminology of the more recent research fronts that is investigated. In addition, it is assumed that the terminology used in the context that surrounds one or several reference markers in citing documents, most often pertains to the same subject matter. In turn, this may lead to disambiguation of terms, and enables subject specific co-occurrence analysis based on citation contexts. Furthermore, in cases where the reference markers signify concept symbols, we may also expect some consensus in the use of terminology. The latter is important, as literary warrant is used as an indication of preferred and non-preferred candidate terms. For example, Sparck Jones (1992, p. 1606) states: “... to function effectively, thesaurus descriptors must be derived from, or at least be strongly motivated by, the particular scientific literature for which they were to be used”. Hence, consensus usage may be a very suitable indication of preferred terminology.

The term selection component of the exploratory methodology serves two closely related purposes. Firstly, to establish whether cited references act as concept symbols, and subsequently, to identify the common concepts expressed by the concept groups in accordance with the concept symbols they contain. Secondly, to select noun phrases from the citation context that surrounds the individual concept symbols within the concept groups. The term selection process is illustrated in Figure 6.4. Figure 6.4 depicts the relationship between citing papers in a research front and their intellectual base of earlier cited documents. If we follow the highlighted citing references in the research front, we are able to illustrate the term selection process.

The citing paper refers to a highly co-cited document (i.e., Gottlow et al. (1986)) in the research fronts’ intellectual base. The citation context for this reference is identified in the citing paper. By comparing a sufficient number of citation contexts for the same cited reference, it is possible to establish whether this reference is a concept symbol to later citing authors. In the present case, the reference symbolizes *guided tissue regeneration*.

FIGURE 6.4. Term selection by use of citation context analysis and shallow parsing. The circles indicate how to read the successive steps illustrated in the figure. (1) Illustrates a citing paper in a research front. One reference in the bibliography is highlighted together with its corresponding citation context in the main document. (2) Illustrates the citation context surrounding the citation marker. (3) A concept symbol analysis identifies an agreed upon concept in the citation context. (4) Shows how a shallow noun phrases parser is able to extract noun phrases from the citation contexts. The dotted lines illustrate that the portfolio of noun phrases is attached to the concept symbol and to the concept group. The concept symbols and their portfolios of noun phrases define the common concept of the group and is the basis for term selection.

In Figure 6.4, the concept symbol is marked in the citation context. Moreover, the concept symbol 'is brought back' to its cited document in the intellectual base cluster, in order to transform the cluster into a group of concepts, where the concept symbols give name to the concept group. In addition, the citation contexts of the potential concept symbols are shallowly parsed in order to extract their attached noun phrases. Eventually, each concept symbol and concept group will have a number of noun phrases attached to them. From this *portfolio* of noun phrases, candidate thesaurus terms are selected based on frequency analysis. Alternatively to Soergel (1974), who suggests that concept identification can be done by computing the sum of frequencies of all terms that designates a concept, we identify concepts based on concept symbols and their parent concept groups.

The term selection method we explore extends the research of Rees-Potter (1987; 1989) in several ways. The most profound extension is that parts of the labour intensive citation context analysis are automated. Traditionally, citation context analysis is done manually, this includes Rees-Potter (1987; 1989). Firstly, citation

contexts have to be identified and selected in citing papers. Next, noun phrases have to be selected and validated. O'Connor (1982; 1983) devised an algorithm for automatic identification of citation contexts in full text documents and subsequent extraction of single word index terms for automatic indexing. We identify and select citation contexts manually. But, noun phrase selection is automatically done by use of a sophisticated shallow noun phrase parser. Hence, we are able to automate the otherwise very labour intensive process of term selection. The specific procedures for context identification, noun phrase parsing, and phrase normalization is presented thoroughly in Chapter 8.

In order to determine whether a cited reference act as a concept symbol in later citing documents, we apply the 'consensus passage' evaluation method used by Small (1986). This entails statistical identification of the citation context that contains the most agreed upon single terms within a sample of citation contexts. The consensus citation context is analysed, in combination with the extracted noun phrases, to determine whether the cited reference act as a concept symbol. Validation of candidate thesaurus terms and their respective concepts groups is done by comparing them to their corresponding MeSH® descriptors, and to their definitions in the *Glossary of Periodontal Terms* (2001).

The noun phrases in a concept group are likely to be related because they are selected from citation contexts. This motivates co-occurrence analyses in order to help the identification of equivalence, hierarchical, and associative relationships of use in thesaurus construction. The use of co-word analysis for this purpose is explored in the fourth component of the methodology, presented in the following section.

6.1.4 *Fourth component: Construction of conceptual networks by use of co-word analysis*

The purpose of the previous component of the exploratory methodology is to identify a potential concept symbol within a concept group, and to extract a number of noun phrases from the citation contexts that surrounds this particular concept symbol in citing papers. The concept symbols and their *portfolio* of noun phrases within a concept group is used in the present component of the methodology to create a conceptual network.

This implies that we essentially have two levels of networks. The first level expresses the relationships between the concept groups, and their individual cited documents, i.e., the concept symbols. This level and its structure are created by the previous two components of the methodology. The second level, expresses the conceptual relationships between noun phrases, including the concept symbols, within

the individual concept groups. This is the conceptual network, which is created with bibliometric methods of the present component. Figure 6.5 below illustrates these two levels of networks.

FIGURE 6.5. *Conceptual network of a concept group.*

The first level includes concept groups and their generic concept symbols. In Figure 6.5, they are illustrated as a network, where the internal structure, i.e., the most salient links, is depicted. Likewise, the external relationship to a neighbouring concept group is also illustrated. The second level illustrates the conceptual network of noun phrases for a concept group. The arrows that point from level one to level two indicate the transformation of concept symbols (reference markers) to noun phrases. The remain of the network is comprised of the portfolios attached to these concept symbols. The creation of the latter is the focus of the fourth component of exploratory methodology. The idea is that such a conceptual network eventually *may* help a thesaurus constructor identify potential associative, equivalence, and hierarchical relationships.

Several approaches exist to the generation of conceptual networks (Soergel, 1974). We apply the bibliometric method of co-word analysis, which in this case is co-occurrence analyses of the extracted noun phrases (i.e., words). In Chapter 4, we described the characteristics of the identified conceptual relationships found in traditional automatic thesaurus construction approaches. As noted above in section 6.1.2, most automatic thesaurus construction approaches use first order term co-occurrence analysis. This form of co-occurrence analysis has its limitations. The most notable is its difficulty in handling synonymous and near-synonymous term associations. From the perspective of co-occurrence analysis, one solution that may alleviate this problem is instead to compare the shared textual contexts of individual index terms. This is the premise of second order co-occurrence analysis and LSI.

The construction of conceptual networks by use of co-word analysis in the present methodology for thesaurus construction is related to the above-mentioned premise of

‘shared textual context’. The selected set of noun phrases assigned to a concept group is *not* clustered due to their frequent co-occurrence in full text documents, as in traditional term association analyses. Conversely, these noun phrases appear together because they frequently share the same textual context, that is, the context that surrounds a specific concept symbol in citing papers. The citation context is likely to be contextual in relation to subject matter. It is therefore assumed that the selected set of noun phrases (candidate thesaurus terms) are important and in some way semantically related. The conjecture is that some of these noun phrases will co-occur directly with each other in the short-span citation contexts. While other noun phrases will not co-occur at all with each other in these contexts, except that they do co-occur with the concept symbol, or rather, with its reference marker. Thereby, all noun phrases become related to each other, either directly by occurring in the same context, or indirectly through their common co-occurrence with the concept symbol. For example, noun phrase *A* and noun phrase *B* are indirectly related if they both co-occur with concept symbol *X*, but not directly with each other. Theoretically, this opens up the possibility of bringing synonymous or near-synonymous terms into the analysis. These terms rarely co-occur together, but they often share common textual contexts. This important basis is the major difference between traditional term association analyses, and the term association analyses applied in the present methodology. There are other minor differences, for example in relation to proximity measures, but the ‘shared textual context’ approach to the co-occurrence analysis is the major difference.

As stated above, the construction of a conceptual network from the noun phrases in a concept group is based on co-word analysis. We explore different applications of the co-word analysis in order to investigate the possibilities of disclosing potential associative, equivalence, and hierarchical relationships.

The basis for the co-word analysis is a vector space representation of noun phrases and their citation contexts in an $n \times m$ matrix. The special thesaurus construction procedures of the second and third components imply that we can pursue what corresponds to first and second order co-occurrence analysis from the same basic $n \times m$ matrix. The matrix is not a traditional co-occurrence matrix. The matrix depicts co-occurrences of noun phrases in ‘shared textual contexts’ of concept symbols within a concept group. This implies that the matrix can contain semantically related noun phrases with low or no direct co-occurrence counts, since such phrases may have similar ‘association profiles’ of other noun phrases in the matrix. The latter corresponds to second order co-occurrence analysis, where association profiles of noun phrases are investigated instead of direct co-occurrences.

First order associations can mean any type of relationship (Soergel, 1974). We therefore apply different proximity measures in order to investigate their potential to disclose specific relationships between the noun phrases. For example, Soergel (1974) suggests that a hierarchical relationship can be surmised if there is a one-sided overlap in the association between two terms.

Second order co-occurrence analysis is also explored for the construction of conceptual networks. Due to the special composition of the matrix, we are able to compare 'association profiles' of two noun phrases by use of second order co-occurrence analysis. The association profile constitutes the co-occurring terms of a noun phrase. This is an investigation of the 'shared textual contexts' of two noun phrases. According to Soergel (1974, p. 38), two noun phrases that have a high second-order association is likely to be in an equivalence relationship. In addition, Soergel (1974, p. 38) suggests, that pairs of terms with high second order associations can be subtracted from the list of term pairs with high first-order associations. Then, according to Soergel (1974, p. 38), the remaining pairs of first order associations should be in an associative relationship. Second order co-occurrence analysis can also be applied to investigate the 'external' relations between the individual concepts groups. Accordingly, the 'association profile' will thus correspond to all chosen noun phrases in the concept group.

In combination with the co-occurrence analyses, we also apply the technique of subsumption as presented in Chapter 4. Traditionally, subsumption is applied in order to derive what corresponds to hierarchical relationships. Further, the subsumption technique can be used in a way similar to factoring of compound terms as presented in Chapter 2. In Chapter 2, 3 and 4, we describe how essentially similar noun phrases, when parsed for indexing purposes, can emerge in different variants. It is therefore necessary to normalize some of the noun phrases in order to do co-occurrence analyses. As an example, normalization can be done by reducing the number of modifying constituents in a noun phrase through subsumption, or by applying manual factoring techniques. Note, that it is important to adjust the frequency count if two phrases are subsumed or otherwise joined (Soergel, 1974).

The final aspect of the construction of a conceptual network is that of visualization. In Chapter 2, we demonstrate the appropriateness of representing the relational nature of thesaurus terms in a graphical display. Likewise, in Chapter 4, we discuss the application of graph theoretical methods to vocabulary organization (Sparck Jones, 1971). Equally, Soergel (1974) stresses the suitability of graph theoretical methods for visualizing the relational nature inherent in thesaurus data. Hence, we interpret the different proximity matrices from a graph theoretical point of view, and consequently

visualize the resulting networks. This provides the possibility of visual inspection and validation of the relationships. It is very likely that some relationships cannot be determined specifically and as a result, they are only indicated empirically. Yet, the visualization brings such relationships into context. Therefore, visualization of conceptual networks may help in the intellectual interpretation of the thesaurus relationships.

In brief, the fourth component of the semi-automatic approach to thesaurus construction investigates the creation of a conceptual network of noun phrases within a concept group, by use of co-word analysis. Different proximity measures and co-occurrence methods are explored to investigate their ability to depict specific relationships. Finally, graph theoretical methods are used as a visual aid for the interpretation of the relational network structures. Consequently, this component goes beyond the former research of Rees-Potter (1987; 1989).

The following section presents the fifth and final component of the exploratory methodology. This component investigates to what extent bibliometric methods can monitor and identify terminological and conceptual changes in a specialty area over a given time period.

6.1.5 *Fifth component: Investigation of terminological and conceptual changes by use of bibliometric methods*

The scientific language in a specialty area, its terms and concepts, may change over a period of time as research progresses. Terms used to describe concepts in documents may change over time. New concepts for which there is no accepted term arise and new terms then appear. Terms fall into disuse and disappear, as the concepts they represent change meaning or become obsolete. New documents will be added to the collection, which describe new or varying concepts. In other words, language is dynamic. This means that thesaurus terms need to be updated in order to reflect their corresponding concepts. Overall, thesauri are time consuming and can be difficult to maintain and update. As a result, there is a need for the development of methodologies by which thesauri can be more easily updated, to account for change and variation in the terminology of a given specialty area. One such maintenance approach is to look for terminological changes within the citation contexts of citing papers. The assumption is that terminology used in the citation contexts of citing papers reflects the concepts represented by frequently cited documents in a specialty area. An analysis of these citation contexts can be used to detect terminological change over time. Terminological change, to some degree reflects conceptual change in a specialty area (Rees-Potter, 1987; 1989).

The research by Rees-Potter (1987; 1989) is the first attempt to apply bibliometric methods as tool for monitoring terminological changes within a specialty area for the purpose of thesaurus maintenance. The results, however, are disappointing (Rees-Potter, 1989). Nevertheless, we also want to investigate whether bibliometric methods can be used as a tool for monitoring the dynamics of language within a scientific specialty area. This is the focus of the present fifth component of the exploratory methodology for semi-automatic thesaurus construction and maintenance.

Similar to Rees-Potter (1987; 1989), we assume that terminological change within a subject area, to some degree also reflects conceptual changes within that area. We assume that concept symbols within a concept group, as well as their attached portfolios of noun phrases, reflect upon a common concept. The attachment of concept symbols and their portfolio of noun phrases to a cited reference facilitate a retrospective bibliometric analysis of the cited references and their parent concept groups. We can go back in time and examine possible changes in the terminology used to convey the concept symbol in citing papers. Hence, we use the cited reference and its citation history as a guide to the monitoring of changes in terminology and concepts over time. The purpose of this analysis is to investigate if bibliometric ageing methods can be used as an indicator that may forecast new or changing terminology in a specialty area.

Two types of bibliometric methods are included in this component in order to investigate potential terminological and conceptual changes over time in a specialty area. The first method is based on quantitative research methods used to map dynamical aspects of science and technology (Braam, Moed & Van Raan, 1991a; 1991b). The second method is a descriptive bibliometric ageing study of concept symbols and concept groups.

The first method is basically a replication of the procedures of the second and third components presented above. To be exact, creation of concept groups, identification of concept symbols, and selection of noun phrase portfolios. However, where the basic case study focuses on a sample text corpus of citing documents published in 2001, the present component investigates four one-year samples of citing documents with an interval of four years between them going back to 1989 (i.e., 2001, 1997, 1993, and 1989). A concept group and its concept symbols from the 2001 sample are selected and their citation and co-citation history is traced by focusing on the selected one-year samples as devised by Small (1977). The retrospective analysis compares *concept profiles* of the different samples with the purpose of investigating continuity or change in these profiles (Braam, Moed & van Raan; 1991a; 1991b). A concept profile is based on the portfolios of noun phrases related to a concept group. In essence, the

concept profile analysis is a second order co-occurrence analysis. In order to study the dynamics of the terminology in more detail a plot of the temporal development of individual noun phrases within the concept group is made. Hereafter we identify ‘core cited references’ for the concept group in order to study the changes in the intellectual base over time. This leads to the second method, the descriptive bibliometric ageing study.

The second method supports the concept profile analysis. Traditional bibliometric ageing methods, presented in Chapter 5, are applied in order to characterize the citation profile of individual cited references (concept symbols). As illustrated by McCain and Turner (1989), such studies can indicate the role played by the concept symbol in later citing documents.

It is likely that these two bibliometric ageing methods may only be suggestive in their findings. However, we explore them to see whether knowledge about the retrospective concept and citation profiles of concept symbols and concept groups, can in some way be used as an indication of terminological and conceptual changes within a specialty area over a period of time.

This concludes the exploratory methodology. The following section presents the rationale behind the selection of the specialty area chosen for the case study.

6.1.6 *The rationale behind the selection of a specialty area for the case study*

The methodology for semi-automatic thesaurus construction presented in this chapter is explored in a case study within the specialty area of periodontology. The criteria for the selection of the specialty area are mainly formulated in relation to the study of Rees-Potter (1987). She investigates two disciplines within the social sciences “... because its terminology may be ‘softer’ i.e. – less precise – than terminology in a physical science ...” (Rees-Potter, 1987, p. 17). Conversely, we want to investigate the opposite, a specialty area that comes from the physical or life sciences, as the behaviour of terminology here may be different to that of the social sciences. We have chosen the specialty area of *periodontology* within dentistry, coming from the life sciences.

Citing authors from the physical and life sciences tend to be more specific in what they refer to in older cited documents. In contrast, authors from the social sciences, tend to cite what they believe to be the general themes of the cited documents (e.g., Cozzens, 1985; Bazerman, 1988; Seglen, 1998; Hargens, 2000a; 2000b). This is interesting, as it may indicate differences in the specificity of citing terminology used between the two types of sciences.

Rees-Potter (1987; 1989) found that most of the concept symbols identified in her study came from cited monographs. Monographs and papers reveal different patterns in citation context analyses (e.g., Small, 1978). Citations to monographs resemble the pattern mentioned above from the social sciences, where it is mainly the general theme of a document that is referred to (Small, 1978; White & McCain, 1989). As Rees-Potter (1987; 1989) did her analysis in the social sciences, and since her concept symbols were mostly cited monographs, it is of no surprise that these concept symbols signified more general concepts in the investigated specialty areas.

Consequently, we want to apply the exploratory methodology in a specialty area, where the primary dissemination of knowledge claims is mediated through journal papers. In addition, the proportion of cited monographs to cited papers most likely affects the composition and changes in the research fronts of a specialty area (Hargens, 2000a; 2000b). Hargens (2000a) suggest that specific citations to a paper are more likely to decline as it ages than citations for general themes. This equally affects the composition and changes in the research fronts of a specialty area (Hargens, 2000a). Thus, we may expect more rapidly changing research fronts than Rees-Potter (1987; 1989), as we focus on a specialty area in the life sciences, where the primary means of written communication are journal papers. The specialty area of periodontology is presented in the following chapter.

Finally, the matrix below, Table 6.1, summarizes the exploratory methodology for semi-automatic thesaurus construction and maintenance.

TABLE 6.1. *Matrix of the exploratory methodology.*

Component	Research question	Purpose	Bibliometric methods investigated	Case study: sample
1: Creation of text corpus	-	To create 4 text corpus samples of overlapping documents from MEDLINE® and SCI®	Data set isolation method	2001; 1997; 1993; 1989
2: Creation of concept groups (Vocabulary organisation)	1	To establish intellectual base clusters of cited references from the 2001 sample. Subsequently, these clusters are transposed into concept groups.	Document co-citation analysis [Pathfinder Networks]	2001
3: Term selection	1	Selection of candidate thesaurus terms through identification of concept symbols, extraction of noun phrases, and naming of concept groups.	Citation context analysis [noun phrase parsing]	2001
4: Conceptual network (Term association)	2	Creation of a conceptual network from a concept group in order to disclose equivalence, hierarchical, and associative relationships.	Co-word analysis [Network visualization]	2001
5: Terminological and conceptual changes	3	Retrospective investigation of a concept group and its concept symbols in order to detect terminological and conceptual changes.	Basis: Components 2 and 3 + 'Concept profile analysis'; Bibliometric ageing studies	2001; 1997; 1993; 1989

Let us stress that we use the case study to explore the different components of the methodology. The aim is not thesaurus construction in itself, but an investigation of the applicability of different bibliometric methods for such purposes. Consequently, the case study is not exhaustive but merely a tool for this investigation.

7. Data analysis: Creation of sample text corpora

The present and following chapters explore the five components of the proposed methodology for semi-automatic thesaurus construction. Chapter 7 concerns the first component of the proposed methodology, that is, the creation of reliable text corpora, whereas Chapter 8 reports on the four remaining components of the proposed methodology. As introduced in Chapter 6, the creation of text corpora constitutes the basis for thesaurus construction and maintenance. The first component therefore is the basis for the subsequent four components of the methodology; where the latter explicitly focus on thesaurus construction and maintenance.

The exploration of the methodology for semi-automatic thesaurus construction and maintenance is based on a case study of a scientific specialty area. The chosen specialty area is periodontology, a specialty area within dentistry. The aim of the first component, therefore is to create reliable sample text corpora based on documents from the chosen specialty area of periodontology. The sample text corpora from periodontology thus constitute the basis for the case study. Subsequently, the case study is used to explore the four remaining components of the methodology, in order to investigate their applicability in relation to thesaurus construction and maintenance.

As the first component is the basis for the methodology, it is treated separately in the present chapter. In brief, the chapter serves two purposes, first to introduce the specialty area of periodontology chosen for the case study; and secondly, to create reliable sample text corpora of documents that concern periodontology. The sample text corpus is generated by use of the *data set isolation method* described by Ingwersen and Christensen (1997). The data set isolation method is based on online searching in bibliographic databases. In the present application, the method creates ‘overlapping’ sets of identical documents from two different databases. An overlapping set of identical documents is defined as the ‘sample text corpus’.

Section 7.1 introduces the specialty area of periodontology. Section 7.2 describes the actual creation of sample text corpora. Finally, section 7.3 describes the validation procedures applied to the created sample text corpora.

7.1 Case study: Periodontology

In the following three sub-sections, we introduce the specialty area of periodontology and some of its most important concepts. As stated above, the introduction serves as background knowledge that can be used in connection with the analyses presented in Chapter 8. The introduction is based on two textbooks, *Periodontics – A Synopsis* (Jenkins & Allen, 2001) and *Carranza's Clinical Periodontology* (Newman, Takei & Carranza, 2002). The textbooks are supplemented by conceptual definitions from *Glossary of Periodontal Terms* (2001) and MeSH® (www.nlm.nih.gov/cgi/mesh).

7.1.1 *The etiology and pathogenesis of periodontal diseases*

According to *Glossary of Periodontal Terms*³⁵ (2001, pp. 38-40), *periodontology* is the scientific study of the periodontium in health and disease. The *periodontium* is the tissues that invest and support the teeth, which include the gingiva (gums), periodontal ligament, cementum and alveolar bone. The *periodontal ligament* is the connective tissue that surrounds and attaches roots of teeth to the alveolar bone. The term *periodontics* refer to the specialty of dentistry concerned with the histology, physiology, and pathology of the tissues that support, attach, and surround the teeth, and of the treatment and prevention of disease affecting these tissues (www.nlm.nih.gov/cgi/mesh).

The term *periodontal* means something that is situated or occurring around a tooth, pertaining to the periodontium (GPT, 2001, p. 38). Consequently, *periodontal disease* is those pathologic processes that affect the periodontium, most often gingivitis and periodontitis (GPT, 2001, p. 39).

The etiology of periodontal disease is bacterial plaque (Jenkins & Allan, 2001, p. 20). The mouth is full of bacteria. These bacteria constantly form a plaque on the teeth. The longer plaque resides on teeth, the more harmful it becomes. Thus, plaque-induced periodontal diseases are those inflammatory conditions of the periodontium, with bacterial plaque as the primary etiological agent (Jenkins & Allan, 2001). Early inflammation is called *gingivitis* (Jenkins & Allan, 2001, p. 9). Gingivitis is a mild form of periodontal disease, and does not include any loss of bone and tissue that hold teeth in place. If gingivitis is not treated, it can advance to *periodontitis* (Jenkins & Allan, 2001, p. 9). Periodontitis is a more serious inflammation of the supporting tissues of the teeth. Periodontitis leads to a progressively destructive change, which causes loss of bone and periodontal ligament, and in the worst cases, the loss of teeth.

³⁵ For short we refer to *Glossary of Periodontal Terms* (2001) as GPT (2001).

Periodontal disease classifications are useful for the purpose of diagnosis, prognosis, and treatment planning (Newman, Takei & Carranza, 2002, p. 63). Several classification systems have been developed over the years (Armitage, 2002). They have usually been changed as new knowledge has improved the understanding of the etiology and pathology of the diseases of the periodontium, see sub-section 7.1.3 below.

According to Jenkins and Allan (2001, pp. 20-22), the microbiology of periodontal diseases includes a number of different bacterial species; the most renowned are *Actinobacillus actinomycetemcomitans* and *Porphyromonas gingivalis*.

The immune system of the body fights the bacteria as the plaque spreads. The extent of tissue damage is dependent on the interaction between plaque bacteria and host defence mechanisms such as inflammatory and immune responses (Jenkins & Allan, 2001). The bacterial toxins and the body's enzymes that fight the infection, actually start to break down the bone and connective tissue that hold teeth in place (Jenkins & Allan, 2001, pp. 21-23). If not treated, the periodontium that support the teeth will be destroyed.

Systemic or local modifying factors act by altering the host response to bacterial plaque (Jenkins & Allan, 2001, pp. 25-30). Local environmental factors may favour the accumulation of bacterial plaque, whereas, a number of systemic conditions have shown to increase the severity of plaque-induced periodontal diseases. The systemic conditions are often named as risk factors. Some of the most significant risk factors associated with the development of periodontal disease are smoking, hormonal changes in women, diabetes, stress, medications, illnesses like cancer or AIDS, and genetic susceptibilities (Jenkins & Allan, 2001). Moreover, recent studies have indicated that periodontal diseases can cause health problems beyond the mouth (Newman, Takei & Carranza, 2002). So far, the research is inconclusive, though, studies are ongoing to try to determine whether there is a cause-and-effect relationship between periodontal disease and an increased risk of heart attack or stroke; an increased risk of delivering preterm, low birth weight babies; and difficulty with controlling blood sugar levels in people with diabetes (Newman, Takei & Carranza, 2002, pp. 229-244).

Epidemiological studies have been valuable in the investigation of the etiology and pathology of periodontal diseases. Epidemiological studies have been carried out to determine population trends in the occurrence and distribution of periodontal diseases (Jenkins & Allan, 2001, p. 38). When large populations are compared, differences in gingivitis levels may be revealed. Differences in gingivitis levels are largely attributable to differences in oral hygiene. Differences in levels of periodontitis

between populations, however, are usually much less pronounced. This is a reflection of the important role of *host response factors*, which are often present in equal measure in different populations (Jenkins & Allan, 2001).

7.1.2 The treatment of periodontal diseases

Oral hygiene is the most important factor in relation to periodontal diseases (Jenkins & Allan, 2001). The most important aspect of oral hygiene in periodontal therapy is plaque control (Jenkins & Allan, 2001). The groundbreaking research on 'experimental gingivitis' in the 1960s, established that plaque control is vital to the monitoring of periodontal diseases (e.g., Löe & Silness, 1963; Silness & Löe, 1964; Löe, Theilade & Jensen, 1965; Theilade et al., 1966; Löe et al., 1967; Jensen et al., 1968). As a result, different periodontal indices were developed, in order to rate the periodontal status of a person or population, for use in periodontal therapy (Löe & Silness, 1963; Silness & Löe, 1964).

The main goal of periodontal disease treatment is to control the infection. The number and types of treatments vary depending on the extent of the periodontal disease. Treatment of periodontal diseases is divided into non-surgical periodontal treatment and periodontal surgery. Non-surgical periodontal treatment include, for example, scaling and root planning, which removes the plaque from the deep periodontal pockets created by the infection (Jenkins & Allan, 2001, p. 73). Medications can be used with scaling and root planning, but medications cannot always take the place of surgery. Periodontal surgery may be necessary if inflammation and deep pockets remain following treatment with scaling and root planning, as well as medications (Jenkins & Allan, 2001). Surgical treatment techniques include, for example, gingivectomy surgery, flap procedures, root separation, mucogingival surgery, surgical crown lengthening, and different regenerative techniques. The latter include, for example, bone and tissue grafts, as well as guided tissue regeneration (Jenkins & Allan, 2001, p. 82).

7.1.3 Paradigms in periodontology

In the past 130 years, classification systems for periodontal diseases have evolved based on the understanding of the nature of these diseases at the time the classifications were proposed (Newman, Takei & Carranza, 2002, p. 63). One consistent feature of the development of classification systems is the guaranteed controversy surrounding any suggested revisions to the previously accepted system of nomenclature (see Armitage (2002) for a review of the classification systems).

Revisions of existing systems have largely been influenced by three dominant paradigms that reflect thinking at the time the classifications were proposed. These are the ‘clinical characteristics paradigm’ (~1870 – 1920), the ‘classical pathology paradigm’ (~1920 – 1970), and the ‘infection/host response paradigm’ (~1970 – present) (Armitage, 2002, p. 9).

For the period from approximately 1870 to 1920 very little was known about the etiology and pathogenesis of periodontal diseases. Accordingly, the diseases were classified almost entirely based on their clinical characteristics. This period is characterized by the ‘clinical characteristics paradigm’ (Armitage, 2002).

As the field of periodontology began to mature scientifically in the first half of the 20th century, many clinical scholars started to develop nomenclature and classification systems for periodontal diseases. Most importantly, it was discovered that periodontal diseases were either inflammatory or non-inflammatory, and that they were based on the principles of general pathology (Løe, 1993). As a result, the classification systems of this period were dominated by what is known today as the ‘classical pathology paradigm’ (Armitage, 2002).

Soon after Robert Koch (1876) provided experimental proof of the germ theory of disease, some dentists began to suggest that bacteria might cause periodontal diseases. Miller (1890), in particular, was an early proponent of the infectious nature of periodontal diseases. Although he spent most of his life studying the oral microflora associated with caries and periodontal disease, his work had very little impact on convincing his contemporaries that periodontal diseases were infections (Løe, 1993). Miller was, nevertheless, an early advocate of the ‘infection/host response paradigm’ that would come to dominate the field nearly a hundred years later. Despite an extensive amount of work on the microbiology of periodontal diseases from approximately 1880 to 1963 very little headway was made in establishing bacterial infections as the foundation upon which periodontal diseases should be classified (Løe, 1993). Part of the reluctance of the profession to accept the notion that most periodontal diseases were infections was the domination of the ‘classical pathology paradigm’ (Løe, 1993; Armitage, 2002). It was not until the classical ‘experimental gingivitis’ studies published by Harald Løe and his colleagues from 1963 to 1968 that the *infection/host response paradigm* began to move in the direction of becoming the dominant paradigm within periodontology (Løe & Silness, 1963; Silness & Løe, 1964; Løe, Theilade & Jensen, 1965; Løe et al., 1967; Jensen et al., 1968). These studies were significant because they provided convincing data that specific changes occurred in the dental plaque flora during the development of gingivitis (Armitage, 2002). This marked the beginning of the dominance of the ‘infection/host response paradigm’.

Although classification systems for periodontal diseases currently in use are firmly based on, and dominated by, the 'infection/host response paradigm', some features of the older paradigms are still valid and have been retained (Armitage, 2002). This concludes the brief introduction to the specialty area of periodontology.

The following section presents the creation and validation of sample text corpora by use of the data set isolation method (Ingwersen & Christensen, 1997). This is the first component of the exploratory methodology for semi-automatic thesaurus construction.

7.2 Creation of 'overlapping document sets' by use of the data set isolation method

As stated in Chapter 6, the motive behind the establishment of a 'sample text corpus' is to utilize the unlike representations of identical documents in *two* different databases, a citation index, and a domain dependent bibliographic database.

By use of cross-file searching and duplicate removal procedures, the data set isolation method identifies the intersection of documents between two files³⁶ (Ingwersen & Christensen, 1997). The intersection consists of documents that are indexed in both files, thus, the intersection corresponds to an 'overlapping document set'. This implies that we have two slightly different bibliographic records for the same document, because the bibliographic representation of the document varies between the files. Then we 'merge' the two different bibliographic records into a single 'enlarged document representation'. The resulting 'enlarged document representation' contains specific document entities, otherwise only located in the records of the citation index, or the records in the domain dependent database.

As presented in section 7.1, the specialty area is periodontology. Bibliographic records from MEDLINE® and SCI® are chosen as objects for the data set isolation method. MEDLINE® is the domain dependent database of periodontology, whereas SCI® is this corresponding citation database. The bibliographic records of 'citing documents' from SCI® provide the set of references needed for the subsequent bibliometric analyses of the methodology. The MEDLINE® records, for the identical 'citing documents', provide document entities, such as MeSH® descriptors, as well as abstract and title noun phrases, which are used for different validation purposes.

³⁶ We use the terms *database* and *file* synonymously.

The application of the data set isolation method in the present dissertation encompasses two successive steps. First, a search string is developed in order to retrieve documents concerning periodontology through a cross-file search. Secondly, the special online procedures of duplicate removal and reverse duplicate removal are applied to isolate the overlapping document sets. Subsequently, the bibliographic records of the isolated document sets can be the target for online data analyses or downloading.

In order to perform data set isolation, some requirements are needed by the online database host. Obviously, the online database host must provide the files relevant to the study. Further, the online database host must support a number of data processing tools, which enable automatic distribution of search results across the files, duplicate removal, as well as frequency analyses (Ingwersen & Christensen, 1997, p. 207). These are the basic elements needed for the online data set isolation method. Dialog® is chosen for the present case study, as it is the online database host of MEDLINE® (file 155) and SCI® (files 34, 434). Furthermore, Dialog® also supports the needed online data processing tools.

The application of the data set isolation method, as the first component in the exploratory methodology, is presented in following two sub-sections. Section 7.2.1 describes the composition of the search string and its application in cross-file searching. Section 7.2.2 describes the process of isolating overlapping document sets. Sub-section 7.2.2 also presents the resulting sample text corpora to be used in the case study.

7.2.1 Cross-file searching

The first step in the data set isolation method concerns the development of an appropriate search string, which is able to retrieve two sets of 'associated' documents from the chosen files in a cross-file search (Ingwersen & Christensen, 1997).

The basis for creation of a document set is a file of bibliographic records. The document set constitutes either the entire file of bibliographic records, or a subset of this file. The criteria by which bibliographic records are 'associated' in a subset of a file are determined by the valid search parameters for the file. Besides the special search commands, which the online database host makes available, search parameters also include the searchable fields of bibliographic records. On a general level, the entity-relationship diagram in Figure 5.2 illustrates a number of document entities, which are searchable in the fields of bibliographic records. If the document set is not defined as an entire file of bibliographic records, then a subset of this file is created from a search string, which comprises one or several document entities, such as

journal, author, descriptor, or words from titles or abstracts. In fact, the cross-file search facility makes it possible to simultaneously observe the number of documents held in different files associated with one or several of such document entities (Ingwersen & Christensen, 1997, p. 207). Ingwersen and Christensen (1997, p. 207) suggest that form-based entities such as publication year, language, and document type, can be used to 'tune' the search string.

Consequently, a search string submitted simultaneously to two files by use of the cross-file search facility produces two sets of 'associated' documents. These 'associated' document sets are characterized by a common share of attributes searched for in explicit document entities. Notice that some of the 'associated' documents are duplicates, as they are indexed in both files. This is the 'overlapping' documents of interest to present study. In order to establish the intersection of documents between the two retrieved sets, data set isolation is applied. As a result, the intersection constitutes the 'overlapping document set', this is elaborated in section 7.2.2.

Ingwersen and Christensen (1997, p. 207) draw attention to the 'comprehensiveness' of the final document set when documents concerning a topic or specialty area is retrieved. If the objective is to isolate and analyze as many documents as possible, including those that incorporate the topic as a minor aspect, the search should include all full-text fields (Ingwersen & Christensen, 1997). Conversely, if the concern is documents that treat the topic as a major aspect, the data set isolation might profit from searching the descriptor and title fields only (Ingwersen & Christensen, 1997). The intention with the search strategy pursued in the present work is to incorporate into the overlapping document set as many documents as possible concerning periodontology. As a result, a search string is created that enables searching in full text document entities in both files simultaneously, such as title, abstract, and descriptor entities.

Clearly, we need to devise a search string that reflects upon the concept of periodontology. It would seem obvious to utilize the supposed high standards of indexing in MEDLINE® for this purpose. Unfortunately, the term periodontology is not a descriptor in the MeSH® vocabulary. Instead, MeSH® uses the term *periodontics* to refer to the specialty area of dentistry that concerns the histology, physiology, and pathology of the tissues that support, attach, and surround the teeth, and of the treatment and prevention of disease affecting these tissues (www.nlm.nih.gov/cgi/mesh). This is a broader definition that includes the aspects of periodontology as defined in *Glossary of Periodontal Terms* (2001, p. 40).

The term periodontics is therefore chosen for the search string. Moreover, in order to exploit the MEDLINE® indexing, the ‘explode’³⁷ search feature in Dialog® is utilized together with the term periodontics. An ‘explode’ search traverses the MeSH® hierarchy of a chosen descriptor and retrieves its narrower terms. This broadens the search, as narrower terms to periodontics will be included in the search result. However, cross-file searching impose restrictions on a search string due to differences in search parameters between files (Ingwersen & Christensen, 1997). This implies that the application of the ‘explode’ search feature is only applicable in the MEDLINE® database. So, to be able to retrieve documents from SCI® an additional search term is needed.

It turns out that the word stem ‘*periodont*’, with a suffix truncation marker, is a significant search term. The truncation ensures that both periodontology, periodontics, periodontium, and periodontal is covered by the search term. The search term ‘periodont’ is not only suitable for searching SCI®, it also ensures that terms, such as periodontium, which is not a part of the periodontics hierarchy in MeSH® is search for in the MEDLINE® file.

Accordingly, the first half of the search string comprises two term variants, which both reflect upon the concept of periodontology. ‘Periodontics’ is applied in an ‘explode’ search, combined in a parenthesis by a Boolean ‘or’ with the truncated word stem ‘periodont’, in order to secure a high recall. The second half of the search string constitutes form-based entities, as we want to delimit the search results by publication year, language, and document type.

The fifth component of the methodology sets out to investigate the applicability of bibliometric methods for monitoring terminological and conceptual changes over a specified period of time. In order to investigate such changes, the sample text corpora used in the case study needs to be delimited to annual ‘overlapping document sets’.

Annual publication windows are therefore applied in the search string, in order to create annual (1-year) sets of overlapping documents. In addition, one overlapping document set is chosen as the basis for the case study. The publication year for the chosen basis set is 2001³⁸, the most recent of the overlapping document sets. The 2001 set is the central text corpus investigated by the four subsequent components of the exploratory methodology. The former document sets are used to investigate possible changes in terminology of the citation contexts. A total of four annual overlapping

³⁷ The ‘explode’ feature works in files that have an online thesaurus with narrower terms (Dialog® Pocket Guide, 2001, p. 14).

³⁸ The year 2001 is chosen due to the commencement of the empirical investigation in 2002.

document sets, with an interval of four years between them, are created by the data set isolation method, these are, 2001, 1997, 1993, and 1989.

The reason why the case study does not investigate successive annual document sets from 1989 to 2001 is due to the high costs associated with downloading bibliographic records from SCI®³⁹. Further, the research by Rees-Potter (1987; 1989) indicates that it is valid to use document sets with intervals between them for thesaurus maintenance purposes based on bibliometric analyses. This means that the fifth component of the methodology covers possible changes in terminology over a publication period of 13 years within periodontology. The investigation is not carried out by focusing on each successive year, but by focusing on four different years within the specified publication period.

The annual document sets concerning periodontology are further restricted to contain two documents types: articles or review articles. Also, the language of the documents is restricted to English, as this is the primary scientific language used in periodontology. The specific document types are chosen because we are only interested in annual 'citing documents' for the investigation. A 'citing document' in this context means a scientific journal paper that propagates knowledge claims and includes references (see Figure 5.2). The above considerations leads to the following cross-file search string:

```
(periodontics! OR periodont?) AND py=2001 AND la=english AND dt=(article OR journal article OR review)
```

In Dialog®, the OneSearch® feature allows cross-files searching. The file numbers of the chosen files are entered in the desired order. Subsequently, the search string is entered once to search all files chosen (Ingwersen & Christensen, 1997). The file order is very important. Eventually, the file order determines where the overlapping documents are contained, and thus from which file the documents can be downloaded (Ingwersen & Christensen, 1997). The order of the files can be altered at will by the 'set files' command.

The following presentation of the data set isolation method is illustrated by the isolation of the 2001 overlapping document set. The results from the other three iterations, i.e., 1989, 1993, and 1997, are presented in Table 7.4. The OneSearch® feature is chosen and the file order is set to 155, 34, which means that the current file is MEDLINE®. The search string is submitted, and the initial search result is presented in Table 7.1.

³⁹ During the dissertation work, it has become possible to download larger sets of bibliographic records free of charge from the Web of Science®. However, this possibility came too late to alter the present search strategy.

TABLE 7.1. Results from cross-file search of 2001 search string.

	<i>File order: 155, 34</i>
	?set detail on
	DETAIL set on
	?s (periodontics! or periodont?) and py=2001 and la=english and dt=(article or journal article or review)
	155: MEDLINE® 1966-2004/Jun W2
	<i>1387 items</i>
	34: SciSearch® 1990-2004/Jun W4 - The term PERIODONTICS has no related terms in File 34.
	<i>1061 items</i>
	TOTAL FILES: 155, 34
S1	2448 items (PERIODONTICS! OR PERIODONT?) AND PY=2001 AND LA=ENGLISH AND DT=(ARTICLE OR JOURNAL ARTICLE OR REVIEW)

The search result presented in Table 7.1 is an adapted version of the search output from OneSearch®. The cross-file retrieval result of the processed 2001 search string has a total of 2448 retrieved documents. The command ‘set detail on’ makes it possible to observe the distribution of the search result between the individual files. The document set retrieved from MEDLINE® contains 1387 documents, and the set retrieved from SCI® contains 1061 documents. The cross-file searching concludes the first step of the data set isolation method. The following sub-section describes the duplicate removal, and reverse duplicate removal procedures used to isolate the overlapping documents from the cross-file search results.

7.2.2 *Isolation of overlapping documents*

In a traditional application of the data set isolation method, the primary focus is usually the ‘removal’ of duplicate records from the retrieved document sets. Ingwersen and Christensen (1997, p. 208) do, however, point out that an alternative to the traditional application, is to focus the study on the duplicate records instead. As stated above, the set of duplicate records between two document sets correspond to an ‘overlapping document set’ in the present study. This following sub-section describes how the overlapping document set is created online by use of the duplicate removal and reversed duplicate removal procedures outlined by Ingwersen and Christensen (1997).

Operationally, the duplicate and reversed duplicate removal procedures are simple. First, it is important to underline that we operate with two different document sets, i.e.

the cross-file search results from MEDLINE® and SCI® respectively. The aim is to isolate the overlapping documents from these search results. The data set isolation method never physically creates a third independent overlapping document set. Instead, depending on the file order, the data set isolation method *isolates* the overlapping documents in the current file. For example, in the above case depicted in Table 7.1. where the file order is 155, 34 (MEDLINE®, SCI®), the current file is MEDLINE®. This implies that the overlapping document set will be isolated among the MEDLINE® records. As a consequence, the file order is very important because it determines where the overlapping document set is isolated (Ingwersen & Christensen, 1997). Figure 7.1 below uses Venn-diagrams to illustrate the influence of the file order in isolating the overlapping document set.

FIGURE 7.1. The influence of the file order on data set isolation. The file order determines where the overlapping document set is isolated.

The isolation process of duplicate records proceeds in two tempi. First, duplicates from the current file are 'removed'. Subsequently, the file order is changed, and the duplicates from the new current file are likewise 'removed'. These are the duplicate removal and reverse duplicate removal procedures. Secondly, the subsequent *isolation* of the document sets is carried out with the 'from' command in OneSearch®.

The duplicate and reversed duplicate removal procedures are carried out with the 'rd' (remove duplicate) and 'set file' commands in Dialog®. Table 7.2 below illustrates the removal and reverse duplicate removal processes for the 2001 cross-file search results.

The search results outlined in Table 7.2 (next page) is a continuation of the cross-file search presented above in Table 7.1. We proceed with search number 2 (S2), and the file order is still 155, 34. First, the 'rd' command is invoked on the MEDLINE® file, which results in 1646 unique documents. The 'display search' command reveals the distribution of unique documents between the files: 1386 unique documents in file 155 (MEDLINE®), and 260 unique documents in file 34 (SCI®). Currently, the

duplicate records reside in file 155 (MEDLINE®). Note that the result of S1 is 1387 documents from MEDLINE®.

TABLE 7.2. Removal and reverse duplicate removal of the 2001 cross-file search result.

<i>File order: 155, 34 (from S1, the cross-file search)</i>	
	?rd s1
S2	1646 RD S1 (unique items)
	?ds (display search)
	File 155: 1386 (unique items)
	File 34: 260 (unique items)
	TOTAL: 1646 RD S1 (unique items)
	?set file 34, 155
	New file order: 34, 155
	?rd s1
S3	1646 RD S1 (unique items)
	?ds (display search)
	File 155: 585 (unique items)
	File 34: 1061 (unique items)
	TOTAL: 1646 RD S1 (unique items)

The result of 1386 unique documents after invoking duplicate removal in S2, most likely means that one document has been indexed twice in MEDLINE®. The duplicate for this document is automatically removed from the set by the 'rd' command.

The 'set file' command switches the file order, so that file 34 (SCI®) can be subjected to the reverse duplicate removal procedure. Obviously, we obtain the same total number of 1646 unique documents as in S2, but this time the duplicate records are placed in file 34 (SCI®). The altered internal distribution of unique documents between the files is observable when we display the search result of S3. File 155 (MEDLINE®) now contains 585 unique documents, whereas file 34 (SCI®) contains 1061. The alteration in the number of documents within the files from S2 to S3 is due to the shifting of duplicate records by the removal and reverse duplicate removal procedures. At this stage, we are able to calculate the actual number of duplicate records, or overlapping documents, by subtracting the two instances of a file after the removal of duplicates (S2) and again after the reverse duplicate removal process (S3). For example, file 155 has 1386 unique documents after S2, and 585 unique documents after S3. Subtracting, 585 from 1386 gives 801 overlapping documents. The same

result is obtained if we subtract the two instances of file 34 (SCI®) from each other. However, in order to exploit the data set isolation method for either online analyses, or downloading of documents, we still need to *isolate* the overlapping documents.

What follows is the isolation of the duplicate records in MEDLINE® and SCI® respectively, by use of the 'from' command as illustrated in Table 7.3 below.

TABLE 7.3. *The isolation of overlapping document sets in MEDLINE® and SCI®.*

	?s s2 from 155
	155: MEDLINE®_1966-2004/Jun W2
S4	1386 S2 FROM 155
	?s s3 from 155
	155: MEDLINE®_1966-2004/Jun W2
S5	585 S3 FROM 155
	?s s4 not s5
	155: MEDLINE®_1966-2004/Jun W2
	1386 S4
	585 S5
S6	801 S4 NOT S5
	?s s3 from 34
	34: SciSearch® Cited Ref Sci_1990-2004/Jun W3
S7	1061 S3 FROM 34
	?s s2 from 34
	34: SciSearch® Cited Ref Sci_1990-2004/Jun W3
S8	260 S2 FROM 34
	?s s7 not s8
	34: SciSearch® Cited Ref Sci_1990-2004/Jun W3
	1061 S7
	260 S8
S9	801 S7 NOT S8

As an example, the isolation of the 2001 overlapping document set within MEDLINE® is carried out on the two instances of the MEDLINE® file identified in S2 and S3. This is done by use of the commands 'S2 from 155' and 'S3 from 155'. Subsequently, the former search result S4 is subtracted from the latter search result S5 by the 'not' operator. The command sequence results in an overlap set of documents from MEDLINE®, which is ready for online data analyses or downloading. The overlap set of the identical documents from SCI® is obtained by a similar command sequence, illustrated in S7, S8, and S9 in Table 7.3.

We have now created an overlapping set of identical documents from MEDLINE® and SCI®, which is ready for online data analyses or downloading. As mentioned above we need the SCI® records for bibliometric analyses. As a result, the file order is set to 34, and the 801 bibliographic records for the overlapping documents are downloaded in their SCI® format from Dialog®⁴⁰. The MEDLINE® records are needed for validation purposes, thus the file order is set to 155 and the identical 801 overlapping documents are downloaded in their MEDLINE® format from Dialog®. The bibliographic records of the two overlapping document sets are subsequently manipulated offline, in order to create the ‘enhanced document representations’. The compilation of such enhanced document representations is carried out by use of the Bibexcel software (www.umu.se/inforsk/). An alternative to downloading the expensive records from Dialog® is to save the ID-numbers for documents in the retrieved overlapping sets created in Dialog®, and then subsequently retrieve and download the identical records from Web of Science® and PubMed®; the web-based editions of SCI® and MEDLINE®. Although some restrictions are imposed on downloading records from Web of Science®, with some ingenuity, all overlapping documents can be downloaded free of charge from these databases. While the latter process is free of charge, it is also cumbersome and time consuming.

As mentioned above the online database host must support a number of online data processing tools in order to facilitate data analyses. One such tool is the ‘rank’ command in Dialog®. ‘Rank’ enables frequency analyses of a number of different document entities within the individual files⁴¹. As presented in Chapter 6, we consider an overlapping document set as a sample text corpus, which is extracted from the population of MEDLINE® documents. The population is defined as the total number of documents retrieved by the search string in the MEDLINE® file. From Table 7.1 we can observe that the total number of documents retrieved from MEDLINE® is 1387, which is adjusted to 1386 due to the removal of one duplicate. This corresponds to the population, and the 801 overlapping documents correspond to the sample. The question of interest to the present study is whether the subject coverage of the sample resembles that of the population. In other words, do the 585 documents, which only appear in the population set and therefore are not incorporated into the case study, alter the subject coverage markedly between the population and the sample? It is important to underline that the search string may produce inexpedient search results, however, this is not what we investigate. We do not consider the relevance of the search result

⁴⁰ The download of MEDLINE® and SCI® records from Dialog® used in the dissertation work, is partly funded by the Danish Medical Research Council, and partly by the Department of Information Studies at the Royal School of Library and Information Science, Denmark.

⁴¹ Frequency analyses can of course also be carried out on the downloaded document sets.

in relation to the composition of the population. We are merely interested in the resemblance between sample and population. In order to investigate this aspect, parallel online frequency analyses are carried out to establish the distribution of major descriptors in the isolated overlapping document set from MEDLINE®, and the population of MEDLINE® documents from where the sample is extracted. The following section 7.3 describes the validation procedures undertaken to investigate the validity of the samples.

In the preceding two sub-sections, we have presented the data set isolation method as it is applied in the first component of the exploratory methodology. The creation of the 2001 basis set of overlapping documents is used as an illustration of the applied method. An additional three annual overlapping document sets are needed for the investigation of the fifth component of the exploratory methodology. As mentioned above, the annual overlapping document sets have an interval of 4 years between them. Therefore, we need to create overlapping document sets of MEDLINE® and SCI® records of the years 1989, 1993, and 1997. The creation of the three extra document sets is similar to the above presented method. The only difference is that the publication year in the search string is changed for each iteration to 1989, 1993, and 1997, respectively. The results are presented in Table 7.4, where the sample text corpus of overlapping documents and the population size of the retrieved MEDLINE® records are indicated.

TABLE 7.4. *Results of the data set isolation method for four 1-year document sets - MEDLINE®.*

Year of document set	1989	1993	1997	2001
Number of overlapping documents (<i>sample text corpus</i>)	322	595	752	801
Number of retrieved documents from MEDLINE® (<i>population</i>)	923	1206	1309	1386
Percentage of sample overlap from the population	34.8%	49.3%	57.4%	57.8%

From Table 7.4 it can be inferred that sample sizes in relation to the populations, almost doubles during the period of investigation. In 1989 the sample constitutes 34.8% of the population, whereas in 2001 the sample constitutes 57.8% of the population. The increasing size of samples indicates that more documents on periodontology are indexed in SCI® during this period. Remember that the overlapping document sets are defined by those documents that appear in both files.

Eventually, it is the journal indexing exhaustivity in SCI® that determines the sample size, as we can expect the domain dependent database MEDLINE® to be more exhaustive in its indexing of journals covering the specialty area of periodontology.

In retrospect, we acknowledge that the applied search string does not fully take into consideration the nature of the MeSH® vocabulary. The search string could have been more exhaustive in relation to periodontology. The question is whether this affects the final sampling results besides their actual size. Are central topics in periodontology missed? In Appendix 21 we further discuss this problem. Evidence suggest that in the present case, a more elaborate search string would make no significant difference.

The following section presents the validation of the samples extracted by the data set isolation method.

7.3 Validation of sample text corpora by use of descriptor frequency profiles

This section describes the validation measures applied to the sample text corpora. The purpose of validation is to establish whether a sample text corpus, and the population corpus which it is extracted from, resemble each other in relation to subject matters covered. As presented in Chapter 6, MeSH® descriptor distributions from the sample and the population are compared in order to establish their degree of common subject coverage. This is the notion of corpus homogeneity (e.g., Kilgarriff & Rose, 1998). A certain degree of homogeneity between the two corpora is desirable because in such cases we can expect thesaurus construction results from the sample to be valid representations of the population. Moreover, such results are most likely *not* a product that has come about due to possible skewness in subject coverage produced by the sampling procedure.

Validation of the sample text corpora is pursued by use of ‘MeSH® descriptor frequency profiles’. Two aspects are considered, the overall resemblance between the ranked MeSH® descriptor frequency profiles, and subsequent identification of significant descriptors in the corpora, that is, the descriptors which differentiate one corpus from another.

7.3.1 *Measures of corpus homogeneity – the problem of hypothesis testing*

The development of measures for comparison of text corpora comes from corpus linguistics (e.g., Johansson & Hofland, 1989; Biber, 1990; 1993; Dunning, 1993;

Kilgariff, 1997; Kilgariff & Rose, 1998). In corpus linguistics, frequency distributions of words (word senses, part-of-speech etc.) are used to compare two or more text corpora (e.g., Biber, 1990; 1993). Several types of corpus comparison exist (Kilgariff & Rose, 1998). The present study focuses on comparison of a sample corpus to a larger population corpus.

There are a number of issues that need to be considered when comparing two text corpora, such as random sampling, reliability of statistical tests, and discovery of features that distinguish one corpus from another (Kilgariff, 1997; Kilgariff & Rose, 1998). In automatic text analysis, statistically based measures are usually based on inferential test statistics. Inferential statistics are useful because, given certain assumptions, they show a known distribution, usually either a normal distribution or a chi-squared distribution (Dunning, 1993). Such distributions can be used to accurately assess significance in a number of different settings (Dunning, 1993).

It therefore seems evident to compare two corpora by use of inferential test statistics. Nonparametric statistics are used in corpus linguistics as word frequency distributions are not normally distributed (Wolfram, 2003). For example, in a pilot study, which precedes the present dissertation work, Schneider and Borlund (2002) test the homogeneity between pairs of samples and populations by use of the chi-square test for homogeneity (Siegel & Castellan, 1988, p. 111). Similar to the sampling procedure presented in the previous section, the pilot study investigated and tested annual overlapping document sets concerning periodontology (i.e., publication year 1989, 1993, and 1997). The test statistics of the three pairs significantly confirmed that the samples and populations were homogenous.

However, several features make hypothesis testing unreliable for textual analyses including corpus comparison (Dunning, 1993; Kilgariff, 1997). The most critical feature is the assumption of random sampling. This is an assumption that pertains to all inferential statistics (e.g., Miller, 1984; Siegel & Castellan, 1988). In relation to corpus comparison, Kilgariff (1997) emphasizes the well-known fact that words in natural language text are not randomly selected (see Chapter 3 on term burstiness). This also includes the assignment of controlled index terms (Wolfram, 2003). The reliability of hypothesis tests is therefore questionable in the absence of random sampling (Siegel & Castellan, 1988). Consequently, the comparison of two supposedly randomly selected text corpora becomes untrustworthy due to the violation of random sampling (Dunning, 1993; Church & Gale, 1995; Kilgariff, 1996; 1998; Wolfram, 2003).

As a consequence, the sampling procedures presented in section 7.2 and in Schneider and Borlund (2002), do not produce random samples. The assignment of

descriptors to MEDLINE® documents is not a random process, and the applied sampling procedures do not produce random samples of the overall populations being studied. The samples are rather incomplete representations of the populations⁴² (Chen, 1989). This does not mean that nonparametric statistics, such as chi-square values, cannot be used for comparative purposes. Indeed they can, but without the power of hypothesis testing, see section 7.3.4.

As a result, we do not apply hypothesis tests of statistical significance to determine the homogeneity between samples and populations in the present dissertation work. Instead, we apply measures derived from nonparametric descriptive statistics as indicators of resemblance and differences between the two corpora.

The following sub-section briefly describes the process of MeSH® descriptor frequency profiling, as this is the basis for the subsequent comparisons by use of nonparametric descriptive statistics.

7.3.2 *Descriptor frequency profiles*

Descriptor frequency profiles from the samples and populations are applied as the basis for corpus comparison. A frequency profile is simply an ordered frequency list of MeSH® descriptors. Due to the isolation of overlapping documents in MEDLINE®, as described in section 7.2, we are able to use the ‘rank’ online frequency tool in Dialog® to generate such frequency lists for the samples and populations respectively. Four pairs of frequency profiles are generated, one for each annual document set (1989, 1993, 1997, and 2001).

Table 7.5 illustrates an excerpt from the MeSH® descriptor frequency list for the 2001 sample.

TABLE 7.5. *Excerpt from the ranked results of MeSH® descriptors in the 2001 sample.*

RANK: S6/1-801 Field: /DE File(s): 155, 34 (Rank fields found in 801 records – 4701 unique terms)		
RANK No.	Items	Term
24	45	WOUND HEALING
455	6	WOUND HEALING –DRUG EFFECTS –DE
4689	1	WOUND HEALING –GENETICS –GE
380	7	WOUND HEALING –PHYSIOLOGY –PH

The descriptors are ordered alphabetically. The alphabetic ordering demonstrates that major subject headings (descriptors) can be subdivided by means of topical

⁴² In fact, Wolfram (2003, p. 91) questions whether data sets used in informetric research truly represent random samples of a larger population.

subheadings⁴³, exemplified in Table 7.5 by WOUND HEALING. As can be inferred from the example, the major descriptor is WOUND HEALING, and the subheading strings are specifications of the major descriptor. Subheadings enlarge the number of unique descriptor entries in a set, while they usually have low frequency counts. Conversely, the frequency count of major descriptors will be scattered among the more specific subheadings, when the latter are incorporated into the frequency profiles. The purpose of validation in the present study is to compare the *general* subject coverage between the two corpora. It is therefore desirable to dissolve the subheadings and pass their frequency counts to the parent major descriptor. Table 7.6 below, illustrates the dissolving procedure for the WOUND HEALING example from above.

TABLE 7.6. *Dissolution of subject strings.*

45	WOUND HEALING	}	59	WOUND HEALING
7	WOUND HEALING –PHYSIOLOGY –PH			
1	WOUND HEALING –GENETICS –GE			
6	WOUND HEALING –DRUG EFFECTS –DE			

Consequently, the ranked MeSH® descriptor frequency lists are reduced, so that they only contain major descriptors with adjusted frequencies for the dissolved subheadings. The dissolution procedure is carried out offline in Microsoft Access®. The resulting adjusted frequency lists constitute the frequency profiles for the four pairs of samples and populations. The frequency profiles are subjected to the following nonparametric descriptive statistics for corpus comparison purposes:

- Spearman rank order correlation statistics, which may indicate the general resemblance in subject coverage between the two corpora, based on their frequency rankings of major descriptors.
- Log-likelihood statistics, which can indicate the most significant relative descriptor frequency differences between the two corpora.

The following sub-section presents the results of the four comparisons of sample and population pairs. This constitutes the validation of the sample text corpora by use of descriptor frequency profiles. Appendix 1 contains the frequency profiles and calculations for the four pairs of samples and populations. The focus in the following sub-section is mainly on the 2001 pair, because the 2001 text corpus sample is the

⁴³ Subheadings are used to restrict retrieval to precisely the aspect of the topic desired, such as WOUND HEALING –PHYSIOLOGY –PH.

basis of study in the four remaining components of the exploratory methodology presented in Chapter 8.

7.3.3 *Spearman rank order correlation statistics*

The Spearman rank order correlation coefficient (Spearman's r_s) is a nonparametric measure of association between two variables. Spearman's r_s requires that both variables are measured in at least an ordinal scale so that the attributes under study may be ranked in two ordered lists (Siegel & Castellan, 1988, p. 235). It is specifically designed to measure the degree of a monotonic relationship between two variables (Miller, 1984, p. 139). A monotonic relationship between two variables exists if one variable tends to increase or decrease as the other variable also increases or decreases⁴⁴. Spearman's r_s is based on a simplified formula of the Pearson product moment correlation coefficient.

The rationale behind Spearman rank order correlation is simple. If two variables are correlated, we can expect the attributes who obtain the lower scores on one variable to obtain the lower scores on the other, and those who have high scores on one variable to obtain high scores on the other (Miller, 1984). Spearman's r_s is based on ranks and not the values of the attributes. This implies that Spearman's r_s does not take into account the difference in magnitude between pairs of attributes in the two variables.

The use of Spearman rank correlation statistics for corpus comparison originates with Johansson and Hofland (1989). In the present study, the pair of variables constitutes the sample and the population. The attributes under study are the frequency rankings of the major descriptors in the two variables. The frequency profiles for the two variables are ranked from high to low, and for each of the N most common descriptors, the difference (d) in rank order between the two variables is taken. The statistic is then the normalized sum of the squares of these differences:

$$r_s = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)} \quad (14)$$

The formula is so constructed that r_s will be +1 when the differences are all zero; that is, when two variables are perfectly correlated. When there is a perfect negative correlation the differences will tend to be very large and r_s becomes -1. When there is no relationship between the variables, the differences will be intermediate in value, and r_s has the value of zero.

⁴⁴ A linear relationship is a special case of the more general monotonic relation between two variables.

While it is possible to make inferences about r_s , we refrain from such hypothesis testing due to problems mentioned above concerning random sampling; this obviously weakens the interpretation of the statistics.

Two correlation coefficients are obtained in the present analyses; one for the top 10 percent ranked descriptors in the population, and another for the top 25 percent ranked descriptors in the population. The choice of 10 and 25 percent rankings is inspired by comparable studies of Kilgarrieff and Rose (1998). Notice that it is the percentage of descriptor rankings from the population that are used, because the aim is to compare the sample rankings to the population rankings.

The percentages are transposed to a numerical threshold ranking position. The threshold ranking position is adjusted to incorporate all ranking values that correspond to the value of the chosen threshold ranking position. The adjustment ensures that tied rankings are incorporated into the correlation analysis. For example, the 2001 population has 4495 unique descriptors (see Table 7.7 below); the top 10 percent of these ranked descriptors include rank position 1 to rank position 449. However, the 449th rank position is tied between several descriptors with equal values. As a result, the 449th rank position is extended to the 458th to incorporate all the tied ranking values. As a result, the N uppermost ranked descriptors in the population are chosen for analysis. In the above example, N corresponds to 458.

We obtain two coefficients in order to observe how the correlation coefficient behaves when more descriptor rankings are included into the analysis. The thresholds of 10 and 25 percent are chosen because we can expect a large majority of descriptor frequencies to appear among the uppermost rankings. This is due to the skewed descriptor distribution as illustrated for the 2001 population in Figure 7.2.

FIGURE 7.2. *Ranked-log-frequency distribution for the 2001 population.*

From Figure 7.2 we can observe that the majority of descriptor frequencies does reside in the uppermost rankings, which corresponds to the left half of the distribution. Further, it is noticeable that approximately half of the unique descriptors in the 2001 population have a frequency count of one, i.e., the right half of the distribution. The top 10 and 25 percent thresholds are indicated in Figure 7.2 as dotted lines. The lines indicate the total descriptor frequencies in the population accounted for by the Spearman rank order correlation analyses at the different percentage thresholds.

It is interesting to establish the total population descriptor frequencies accounted for in the correlation analyses. This is interesting because it is assumed that if a considerable number of such frequencies appear among the uppermost rankings chosen for analyses, then we should be able to state that these descriptors reflect the general subject coverage in the population.

The percentage of descriptor frequencies accounted for by the Spearman rank order correlation analyses is calculate by dividing the sum of frequencies obtained for the N descriptors with the total number of descriptor frequencies in the population and sample respectively. It is assumed that the more population descriptor frequencies accounted for in the analyses, the more reliable the statistical result in relation to resemblance in subject coverage. Likewise, it is assumed that the lower frequency descriptors reflect specific subject matter. These descriptors are therefore not incorporated into the correlation analyses, as the latter concerns general subject coverage.

Consequently, we compare the rankings of descriptors in the population to their rankings in the sample, in order to determine whether the sample resembles the population. We assume that related rankings imply homogeneity in subject coverage between sample and population. We are interested in general subject coverage, which means that we only compare the N uppermost ranked descriptors. Consequently, high correlation coefficients therefore indicate homogeneity between sample and population, that is, the sample reflects the general subject coverage in the population. The results are presented in Table 7.7 on the following page.

Overall, the results from the Spearman rank order correlation analyses indicate a solid correlation between the uppermost rankings of major descriptors in the populations and samples for all document sets and percentage thresholds.

TABLE 7.7. *Results of the Spearman rank order correlation analyses.*

Document sets	1989	1993	1997	2001
Spearman's r_s for the top 10 percent descriptors	$N = 151$ $r_s = 0.781$	$N = 198$ $r_s = 0.761$	$N = 247$ $r_s = 0.806$	$N = 458$ $r_s = 0.870$
Percentage of frequencies covered by the correlation analysis (top 10 percent)	Pop: 67% Sam: 71%	Pop: 66% Sam: 66%	Pop: 67% Sam: 67%	Pop: 71% Sam: 78%
Spearman's r_s for the top 25 percent descriptors	$N = 417$ $r_s = 0.737$	$N = 524$ $r_s = 0.763$	$N = 635$ $r_s = 0.841$	$N = 1062$ $r_s = 0.789$
Percentage of frequencies covered by the correlation analysis (top 25 percent)	Pop: 84% Sam: 86%	Pop: 83% Sam: 83%	Pop: 84% Sam: 84%	Pop: 83% Sam: 89%
Frequency of descriptors in population (pop) and sample (sam)	Pop: 9286 Sam: 3931	Pop: 15473 Sam: 8291	Pop: 21328 Sam: 13478	Pop: 27780 Sam: 14401
Percentage overlap	42%	54%	63%	52%
Unique descriptors in population (pop) and sample (sam)	Pop: 1402 Sam: 785	Pop: 1967 Sam: 1414	Pop: 2363 Sam: 1854	Pop: 4495 Sam: 1995
Percentage overlap	56%	72%	78%	44%

Even though the samples are not randomly selected; we would expect the correlation coefficient to be highest among the top 10 percent ranked descriptors due to the core phenomenon introduced in Chapter 5. The core phenomenon implies that a minor set of descriptors concentrate a significant number of the total frequencies. Likewise, we would expect the correlation coefficient to drop, as more descriptors are included in the analyses, due to the scatter phenomenon. The scatter phenomenon implies that frequencies are dispersed among many descriptors. In brief, it would seem more difficult to match frequency rankings of 500 different descriptors instead of 200 descriptors.

Nevertheless, the core and scatter phenomena is not so apparent in the present correlation analyses. The 1993 document set has almost no change between the two thresholds, whereas the 1997 document set has a rise from the top 10 percent threshold to the top 25 percent threshold. In contrast, the 2001 document set has a marked difference between the two thresholds, which do indicate the core and scatter phenomena. In addition, the top 10 percent threshold in the 2001 document set is the highest correlation coefficient in the analyses. The graphs in Figure 7.3 below compare the changes in the coefficients between the threshold percentages.

FIGURE 7.3. Comparison of annual Spearman rank order correlation coefficients.

In Figure 7.3, the annual changes in the coefficients between the threshold percentages may seem considerable. Likewise, the deflection of coefficients between the threshold percentages within a document set, the core-scatter phenomena, may also seem considerable, especially in the 1997 document set. But, they are most likely not considerable changes or deflections, as the interval of the observed coefficients is in fact rather small. The deflection within the 1997 document set is a reflection of minor differences of rank order between corresponding descriptors. For example, if a successive string of ten descriptors in one corpus are tied at one rank, and the corresponding descriptors in the other corpus have individual rankings, then this difference will cause some minor deflections.

Notice the uniformity in the percentage of population descriptor frequencies accounted for by the correlation analyses. For the top 10 percent threshold, the percentages accounted for in the four analyses are between 66 to 71%, while the percentages accounted for at the top 25 percent threshold are very stable at 83 to 84%. As a result, both thresholds indicate sufficiently high correlation coefficients, and at the same time account for a considerable number of the total population descriptor frequencies. It is therefore fair to conclude that the samples resemble the populations in relation to general subject coverage as defined above. In fact, the uniformity observed at the top 25 percent threshold suggests that the first quartile (25th percentile) generally is able to concentrate around 80% of the total population descriptor frequencies. This is a core phenomenon related to the Bradford (1934) or Lotka (1926) distributions. The question in this context is, whether the correlation coefficient is sufficiently high at the first quartile, for this to be a general threshold in similar studies. In the present study, we do find the first quartile threshold sufficiently high.

From a validation perspective, it is interesting to monitor the proportionate growth in total descriptor frequencies, as well as the proportionate growth in the number of unique descriptors, between populations and samples. Obviously, there exist a proportionate gap in frequencies and descriptors between the populations and samples. What is interesting is whether such gaps are stable and continuous during the period of study. If deviations occur, they need to be investigated, in order to make sure what they imply for the validation of the overlapping document sets.

The graphs in Figure 7.4 below, project the total frequency counts of descriptors in the populations and the samples for the four document sets (data from Table 7.7 above).

FIGURE 7.4. *Frequency of descriptors in populations and sample.*

It is observable that the population growth is almost linear, and that that the population and samples go together in their proportionate growth until 2001. The population frequency counts continue to increase at the approximate growth rate from 1997 to 2001, while the sample proportionate growth rate, for the same period, stalls. From table 7.7, we can observe that the percentage overlap of descriptor frequencies between population and sample drops from 63% in 1997 to 52% in 2001.

In comparison, Figure 7.5 below shows a similar growth pattern as the one in Figure 7.4, where populations and samples go together, until the marked difference in 2001. However, Figure 7.5 concerns the number of unique descriptors in the populations and samples of the document sets during the period under study (data from Table 7.7 above).

FIGURE 7.5. *Unique descriptors in population and sample.*

From 1997 to 2001, the number of unique descriptors in the population increases considerably, seen in relation to the relatively small rise in the sample. From Table 7.7 we can notice that the overlap of unique descriptors between the populations and samples rise from 56% in 1989 to 78% in 1997. In 2001, however, the overlap of unique descriptors then suddenly drops to 44%. Thus, a continuous growth can be observed in the period under study until the 2001 document set.

On the one hand, the total descriptor frequencies in the 2001 population continue to grow almost linear. But, at the same time there is a marked rise in the number of unique descriptors in the population. On the other hand, compared to population, the proportionate growth of both the total descriptor frequencies and the number of unique descriptors in the 2001 sample flatten off. The question is what this difference implies for the resemblance in general subject coverage between the sample and population.

From the results of the correlation analyses for the 2001 document set, presented in Table 7.7, it can be observed that both the top 10 and 25 percentage thresholds yield sufficiently high correlation coefficients, and account for a considerable number of population descriptor frequencies compared to the other documents sets. This means that the 2001 sample do in fact reflect the general subject coverage of the population as defined in this study.

As a consequence, the deviation in the proportionate growth in the 2001 document set must be accredited to a considerable number of unique descriptors with low frequencies (the scatter phenomenon), which is not incorporated into the sample of overlapping documents. The reason why these unique descriptors are not incorporated into the sample is most likely that they come from journal papers not indexed in SCI®.

Obviously, these descriptors account for more than half of all the descriptors in the population, but still the large quanta of frequencies reside among the uppermost rankings due to the core phenomenon. Therefore, the deviation does not influence the validation of general subject coverage investigated by the Spearman rank order correlation coefficient. We thus conclude that all created sample text corpora resemble the populations in relation to general subject coverage.

7.3.4 Identification of relative descriptor frequency differences

The Spearman rank order correlation coefficient indicates the degree of resemblance between two descriptor frequency profiles. But, the correlation coefficients do not specify the significant similarities and differences between the descriptor frequency profiles, because ranks are compared and not the magnitudes of the descriptor frequencies. Knowledge concerning the similarities and especially the differences in the descriptor profiles between samples and populations are important in order to understand the descriptor composition of the samples. Eventually, knowledge about the descriptor composition of a sample, i.e., the subject coverage of a text corpus, is valuable for the analyses of the subsequent four components of the methodology.

We apply the log-likelihood statistic to measure the difference in a descriptor's frequency between the sample and population profiles used for the correlation analyses. Again, we refrain from hypothesis testing. An alternative is the chi-square statistic, but Dunning (1993) points out that the chi-square statistic becomes unreliable when frequencies are less than five. Likewise, the chi-square also overestimates high frequencies (Dunning, 1993). Instead, Dunning (1993) proposes the use of the log-likelihood statistics. The log-likelihood statistic is composed from a contingency table of descriptor frequencies illustrated in Table 7.8.

TABLE 7.8. *Contingency table for descriptor frequencies.*

	Population	Sample	Total
Frequency of Descriptor	a	b	a+b
Frequency of other Descriptors	c-a	d-b	c+d-a-b
Total	c	d	c+d

Notice that the value ‘c’ corresponds to the descriptor frequencies in the population, and ‘d’ corresponds to the descriptor frequencies in the sample (n values). The values ‘a’ and ‘b’ are the observed frequencies (O). Consequently, the expected frequencies (E) need to be calculated according to the following formula:

$$E_i = \frac{n_i \cdot \sum_i O_i}{\sum_i n_i} \quad (15)$$

In the present study, the population ($n1$) = c, and the sample ($n2$) = d. So, for descriptor X , the two expected values are calculated in the following manner: $E1 = c \times (a+b)/(c+d)$ and $E2 = d \times (a+b)/(c+d)$. The calculation for the expected values takes account of the total frequencies of the two profiles, thus the figures do not need to be normalized before applying the log-likelihood formula. The log-likelihood value is calculated according to the following formula (Dunning, 1993):

$$-2 \ln \lambda = 2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right) \quad (16)$$

This equates to calculating the log-likelihood as follows: $2 \times ((a \times \ln(a/E1)) + (b \times \ln(b/E2)))$. The resulting relative descriptor frequency list is sorted by the log-likelihood values. This gives the effect of placing the largest log-likelihood value at the top of the list. The top of the list therefore shows the descriptors that have the most significant relative frequency differences in the sample compared to the population. Descriptors that appear with roughly similar relative frequencies in the two corpora appear lower down the list. Table 7.9 on the following page illustrates the top and bottom of the relative descriptor frequency list of the 2001 document set (see Appendix 1) for the complete lists for 1989, 1993, 1997, and 2001).

The log-likelihood statistic does not consider zero frequencies. This is inexpedient, as we also want to monitor the frequently occurring descriptors in the population, which is not included in the sample. As a result, these descriptors must be obtained in another manner. Notice however, that the correlation coefficient in a general manner indicates missing descriptors in the sample, as the coefficient drops for each missing descriptor in the sample.

TABLE 7.9. *Relative frequency similarities and differences between descriptors in the population and sample.*

Top 20 log-likelihood values		Bottom 20 log-likelihood values	
Log-likelihood value	Descriptor	Log-likelihood value	Descriptor
18,948	GINGIVA		
15,702	DENTAL IMPLANTS	0,001	ADIPOSE TISSUE
13,356	PERIODONTAL ATTACHMENT LOSS	0,001	ADOPTIVE TRANSFER
13,345	DENTAL IMPLANTATION	0,001	ALANINE
12,662	PATIENT CARE PLANNING	0,001	ALCOHOLISM
12,299	ALVEOLAR RIDGE AUGMENTATION	0,001	ALGORITHMS
12,237	PORPHYROMONAS GINGIVALIS	0,001	AMELOBLASTS
12,153	PERIODONTAL POCKET	0,001	ANGIOSPERMS
11,326	PERIODONTAL LIGAMENT	0,001	ASIAN AMERICANS
10,967	RECEPTORS	0,001	AUTOCRINE COMMUNICATION
10,760	INFLAMMATION	0,001	BENZENESULFONATES
10,585	CELLS	0,001	BLADDER NEOPLASMS
10,080	EPIDEMIOLOGY	0,001	BONE CEMENTS
10,046	ANTIGENS	0,001	BRADYKININ
9,635	RNA	0,001	BRAIN ISCHEMIA
9,591	FIBROBLASTS	0,001	BRONCHITIS
8,508	RATS	0,001	CARBON RADIOISOTOPES
8,486	PERIODONTAL DISEASES	0,001	CHELATING AGENTS
8,463	MICE	0,001	CLUSTER ANALYSIS
8,138	ANIMALS	0,001	COFFEE
.....		0,001	COLLAGEN TYPE IV

The results of the log-likelihood statistics give an indication of the relative similarity and difference between descriptors in the population and sample. Notice Spearman's r_s compare ranks, whereas log-likelihood investigates the relative magnitudes of the frequencies between population and sample. Where Spearman's r_s are used to validate the samples, the log-likelihood statistics are used to elaborate on the composition of the descriptor frequency profiles compared in the correlation analysis. Eventually, this knowledge may come in handy, in the subsequent analyses of the exploratory methodology.

This concludes the exploration of the first component of the proposed methodology for semi-automatic thesaurus construction. This component has created four annual sample text corpora and validated them. Moreover, we devised a statistic that can characterize the similarities and differences between the frequency profiles used for validation. We now proceed to the components of the methodology that concerns actual thesaurus construction. This is presented in Chapter 8.

8. Data analysis: Exploration of components 2 to 5 of the proposed methodology

The present Chapter continues the exploration of the methodology's four remaining components. Chapter 7 presented the first component, which concerned the creation of sample text corpora from the specialty area of periodontology. Four annual set of sample text corpora were created: 1989, 1993, 1997, and 2001. The sample text corpora are used as a case study for exploration of the components two to five of the methodology presented in the present chapter. These remaining components concern semi-automatic thesaurus construction and maintenance based on bibliometric methods. The 2001 sample text corpus is chosen as the basis corpus for the case study. This implies that the 2001 sample text corpus is the only corpus applied in the second component concerning vocabulary organization, the third component concerning term selection, and the fourth component concerning term association. Conversely, all four sample text corpora are applied in the fifth component, which concerns the monitoring of terminological and conceptual changes over a period of time.

The purpose of the present chapter is to explore the applicability of bibliometric methods for semi-automatic thesaurus construction and maintenance as envisaged in the components of the methodology. It is important to underline that the focus of the dissertation work is the development of a methodology for semi-automatic thesaurus construction and maintenance based on bibliometric methods. This means that the envisaged methodology is explored through a case study to acquire an initial impression of the efficiency of the methodology and its components in relation to a specific specialty area. This also means that no large-scale experimental evaluation of the methodology and its components is carried out. However, the present exploratory study presented in this chapter can be seen as a verificative study, which examines the initial applicability of bibliometric methods for semi-automatic thesaurus construction and maintenance. The results of the present study may therefore constitute the basis for further experimental evaluation studies.

The present chapter comprises four sections, where each section concerns the exploration of a component from the methodology. As outlined in Table 6.1, the second and third components set out to investigate the first research question. The fourth component investigates the second research question, and finally, the fifth component investigates the third research question. Thus, the exploration of

components two to five successively follows the three research questions presented in Chapter 1. This succession defines the composition of the present chapter.

Chapter 6 describes the foundation and purposes of the exploratory methodology and its individual components. Further, Table 6.1 summarizes the purposes of the individual components, the bibliometric methods they investigate, and finally attach the components to the research questions. We elaborate on Table 6.1, by recapitulating the three research questions, and relating them to the components of the methodology.

The first and most central research question investigates the basic characteristic of thesaurus construction, which is the identification of candidate thesaurus terms. The question reads:

1. Is it possible, by use of bibliometric methods, to detect candidate thesaurus terms in a specialty area within the life sciences given its disciplinary, publication, and terminological conditions?

As stated in Chapter 1, the first question will clarify whether candidate thesaurus terms can be identified at a lower domain level than hitherto done, and in a scientific domain where scholarly knowledge is primarily mediated through journal papers.

The second and third components of the methodology explore the first research question. The purpose of the second component is 'vocabulary organization'. In the present case, it implies the structuring and mapping of the 2001 sample text corpus by use of *document co-citation analysis* and *Pathfinder Networks*. Eventually, this creates concept groups further explored in the third component.

The purpose of the third component is 'term selection'. Based on the concept groups for the 2001 sample established in the previous component, the third component investigates the ability of *citation context analysis* and *noun phrase parsing* for the selection of candidate thesaurus terms. This implies identification of concept symbols, extraction of noun phrases, and naming of concept groups.

The second research question investigates to what degree the applied methodology is able to help identify basic thesaural relationships between individual terms and concepts. This leads to the following research question:

2. To what extent can the applied bibliometric methods, outlined in the exploratory methodology of the semi-automatic approach, help identify equivalence,

hierarchical and associative relationships between individual terms and concepts within the concept groups?

As stated in Chapter 1, this will clarify whether it is possible to identify, and to what extent, thesaural relationships by use of the tools applied in the proposed methodology.

The fourth component explores the second research question. The purpose of the fourth component is ‘term association’. In the actual case, this implies the creation of a conceptual network from the selected terminology of one concept group by use of *co-word analyses* and *network visualization techniques*. In this context, terminology comprises of concept symbols and their portfolio of noun phrases, identified and selected in the second and third components of methodology. The conceptual network is investigated to see whether the present application of co-word analysis can indicate equivalence, hierarchical, or associative relationships.

The third research question concerns the applicability of bibliometric methods to monitor possible terminological and conceptual changes over a period of time in a specialty area. This is an important aspect of thesaurus maintenance because thesauri need to be updated as a result of such changes. This leads to the third and last research question:

3. To what extent can the applied bibliometric methods, outlined in the exploratory methodology of the semi-automatic approach, monitor and identify terminological and conceptual changes in a given subject specialty area over a given time period?

As pointed out in Chapter 1, it is assumed that terminological changes within a specialty area, to some degree also reflect conceptual change within that area.

The fifth component of the methodology explores the third research question. The purpose of the fifth component is to monitor terminological and conceptual changes within a specialty area for thesaurus maintenance. Two types of bibliometric methods are investigated for their ability to monitor potential terminological change or continuity over a period of time in the specialty area. The first type of method is a retrospective *concept profile analysis* carried out for a chosen concept group, with the purpose of investigating continuity or change in these profiles. One concept group in the 2001 sample text corpus is chosen and monitored retrospectively. The retrospective analysis is carried out by use of the remaining three sample text corpora, i.e., 1989, 1993, and 1997. This means that the remaining three sample text corpora are only used in the case study for the fifth component. The three remaining sample

text corpora need to be subjected to the procedures of the second and third components, that is, vocabulary organization and term selection, in order to be used in the retrospective analyses.

The second type of method applied in the fifth component utilizes *bibliometric ageing methods* in order to support the concept profile analyses. Even though the exploration is retrospective, it may indicate whether changes in concept profiles and the citation history of highly cited documents can be used as an indication of conceptual change in a specialty area?

Statistical test are used to validate the initial co-citation results. Further, a baseline of MeSH® descriptors and term definitions from the *Glossary of Periodontal Terms* (2001) are used in the quantitative evaluation of the three research questions.

The following four sections of the chapter investigate the three research questions by applying the case study of periodontology. Each section briefly restates the purpose of investigation and introduces the bibliometric methods investigated. Subsequently, the methodical steps are presented and discussed in conjunction with results and validation procedures. Finally, each section concludes with a brief summary and discussion of results. Consequently, section 8.1 reports on the creation of concept groups (vocabulary organization); section 8.2 concerns term selection; section 8.3 deals with conceptual networks; and finally section 8.4 concerns the monitoring of terminological and conceptual changes.

A note should be made about the format and organization of appendices related to the performed analyses in this chapter. Due to the extensive amount of data used in the analyses, several appendices are enclosed as xls- or pdf-files on a CD-rom. A reference is made in the text to an appendix number that appears in the appendix section of the printed dissertation. The individual appendix, in the appendix section, will explain where to locate and how to read the data on the CD-rom.

We thus proceed with the presentation and discussion of the second component of the methodology, which together with the third component, investigates the first research question stated above.

8.1 Second component: Creation of concept groups (vocabulary organization)

The second component of the exploratory methodology investigates the ability of document co-citation analysis to group semantically related cited references, that is,

potential concept symbols, in corresponding concept groups. The component is comparable to vocabulary organization in traditional automatic thesaurus construction approaches as discussed in Chapter 6. This implies that the traditional steps in automatic thesaurus construction are altered, as we do not commence with term selection. Instead, we commence by establishing a framework of potential concept groups, that is, clusters of highly co-cited documents. Subsequently, the succeeding term selection component of the methodology will identify concept symbols, name the concept groups, and assign candidate thesaurus terms to these concept groups.

As a result, the second component serves as the initial step for term selection, and refers to the first research question, in conjunction with the third component.

Clusters of co-cited documents serve as the ‘intellectual base knowledge’ to citing papers in more recent research fronts. The individual cited documents within these clusters are comparable to the notion of ‘exemplary documents’ suggested by Blair and Kimbrough (2002), in that they represent the key concepts, methods, or experiments, which researcher build on in a research front of a specialty area (Small, 1978). Some of these cited documents symbolize the same content to a majority of later citing authors, which result in a consensus usage of terminology when citing those documents (Small, 1978). Hence, highly cited documents may act as concept symbols to citing authors in a research front. Moreover, Rees-Potter (1987: 1989) indicates that such clusters of highly co-cited documents can be treated as concept groups for thesaurus construction purposes. As a result, the cited reference is used as the primary entity for ‘vocabulary organization’, and co-citation analysis is used as the principal method of organization. Due to the implicit link between cited references and concept symbols in the citation context of citing papers, we can select terminology from citation contexts and assign it to the concept groups. It is assumed that terminology related to the usage of references in citing documents is contextual and subject specific. As presented in Chapter 6, this is an indirect approach to vocabulary organization.

In order to investigate the ability of document co-citation analysis for the above-mentioned purposes, we explore the usage of two different proximity measures to investigate their influence upon the applied complete-link clustering algorithm.

Finally, to obtain a more intuitive understanding of cluster compositions, i.e., their internal and external structural relations, the Pathfinder network algorithm (Schvaneveldt, 1990) is imposed on the proximity results.

The following section presents, discusses, and validates, the methodical steps of the second component.

8.1.1 Methodical steps of the second component

The primary methodical steps and decisions of the second component is outlined in Figure 8.1, which serves as a general view of the following presentation and discussion.

FIGURE 8.1. *Methodical steps of the second component.*

Figure 8.1 illustrates the process of concept group creation by use of document co-citation analysis. The rectangular boxes denote the results of the transformation process of cited references to concept groups. Chapters 4, 5, and 6 discussed the potential value of such a transformation in relation to thesaurus construction. The results, which these boxes illustrate, are presented either below or as appendices. The circular boxes denote the bibliometric methods, multivariate statistical techniques, and network analyses, applied to the cited references. The generic procedures of citation

and document co-citation analysis are thoroughly presented in Chapter 5. Likewise, the generic procedure of cluster analysis is presented in Chapter 4, and its application in document co-citation analysis is discussed in Chapter 5. The Pathfinder network analysis tool applied in the present dissertation is introduced in the present sub-section 8.1.1.4. The diamond boxes illustrate some crucial choices in the process of document co-citation analysis. The choices concern threshold values and proximity measures. The composition, characteristics, and meanings, of proximity measures are discussed in Chapter 4. Further, the application of proximity measures for document co-citation analysis is presented in Chapter 5. As stated above, two different proximity measures are investigated in order to investigate their influence upon clustering. Finally, the arrows in Figure 8.1 indicate the successive steps of the second component. The presentation and discussion of the present component follows the successive methodical steps outlined in Figure 8.1. The sub-sections are centred on the circular boxes, as indicated in Figure 8.1. Thus, we commence with document citation analysis of the 2001 text corpus created in the first component of the methodology.

8.1.1.1 Document citation analysis

The focus of study is highly cited documents from the 2001 sample text corpus. The present component identifies and clusters highly cited references. The subsequent components investigate if the highly cited references act as concept symbols in the citing documents published in 2001 concerning periodontology. Note, the records of the 2001 sample text corpus comprise of cited references from ISI records, as well as MeSH® descriptors and abstracts from MEDLINE®. The 2001 sample comprises of ‘enhanced document representations’ due the merger of records of identical documents, made possible by the data set isolation method.

A major criticism often levelled at citation analyses is the problem of erroneous cited reference strings in ISI citation indices (e.g., King, 1987). Obviously, such errors affect a frequency analysis, as frequencies get scattered among deviations of essentially identical cited references. The records in ISI are automatically indexed. Cited reference strings are automatically extracted from the bibliographies of citing papers. The most common error is therefore misspelling of references in the citing documents. In addition, problems such as different abbreviations of cited works, due to the lexical analysis applied in indexing at ISI, are also commonly observed. Consequently, we need to correct the erroneous indexing of identical cited references to a standard form in order to do frequency analysis.

The 2001 sample consists of 801 documents and 31750 cited reference strings. The strings are sorted alphabetically, and manually scanned to identify cited references

with a frequency of four or more. The cited reference strings are subsequently checked for erroneous indexing. The standardised cited references are thereby ready for citation analysis.

The document citation analysis is performed by use of the Bibexcel software (www.umu.se/inforsk/). The result is presented in Appendix 2. From Appendix 2 it appears that a threshold value is imposed for the present study. At this early stage, we could subject the highly cited documents for the 2001 sample to citation context analysis, in order to investigate whether they act as concept symbols. But, the purpose of the second component is to create concept groups of highly cited documents through co-citation analysis. A threshold value is therefore needed to delimit the co-citation study. Clearly, the choice of threshold value influences the document co-citation results. In the present case study, a relatively high threshold value most likely reduces the number of concept groups. However, the aim of our dissertation work is not an exhaustive mapping of a specialty area, but rather a probing study that explores the general applicability of the investigated bibliometric methods. In addition, concept symbols are usually identified among highly cited documents, so anyhow a certain threshold value is necessary, the difficulty will always be how to define such a value.

For pragmatic reasons the threshold value is set to 13 citations, which is a relatively high value for the specialty area of periodontology in 2001. The investigations of the third component will reveal if the citation threshold causes problems in the process of identifying concept symbols.

As mentioned above the 2001 sample contains 801 citing documents with 31750 references, distributed among 21288 unique references. The mean number of references per citing document in periodontology in 2001 is 39.0, whereas, the mean number of citations received per cited document is 1.5. The latter is of no surprise, as distributions of citations conform to the traditional highly skewed rank-frequency distributions, illustrated in Figure 8.2 below.

FIGURE 8.2. *Frequency distribution of citations in the 2001 sample.*

The threshold value of 13 citations results in the incorporation of 64 cited references into the document co-citation analysis. These 64 cited documents constitute 1224 citations, which covers approximately 4% of the total number of citations in the sample. Nevertheless, it is the highly cited references; the ones we can expect to act as concept symbols, and the ones that are most likely to form part of significant co-citation relationships. Appendix 2 presents the top 64 cited documents chosen for the subsequent document co-citation analysis.

This concludes the first initial methodical step in the creation of concept groups. The following sub-section concerns the procedures of document co-citation analysis.

8.1.1.2 Document co-citation analysis

In the previous section, we selected 64 highly cited documents using an integer citation threshold. In the present sub-section, the 64 most cited documents within the 2001 sample text corpus concerning periodontology are subjected to a document co-citation analysis. This implies that the *relative* strength of co-citation relations among the 64 cited documents has to be established in this present sub-section. Relative co-citation strengths are determined by proximity values and not integers. A proximity value is appropriate for co-citation clustering as it eliminates some of the size effect involved in linking documents with widely different citation and co-citation counts. This leads to a selection of ‘significant’ and ‘non-significant’ pairs of co-cited documents for the present study.

As described in Chapters 4 and 5, a co-occurrence analysis commences with an $n \times m$ vector component matrix. References that have been cited together or individually by m different citing papers are ordered in some fixed arbitrary way. The vector that represents a citing document has a *one* or a *zero* in the k^{th} component according to whether or not it cites the k^{th} cited reference. This means that the basis matrix for document co-citation analysis is a binary $n \times m$ matrix. The aim is to transpose the $n \times m$ binary matrix into a symmetric $n \times n$ proximity matrix that depicts the relative strength of co-citation relations among the cited reference pairs. The relative strength of co-citation relations indicates the ‘significant’ and ‘non-significant’ pairs of co-cited documents. This leads to a modification of the matrix. The modified proximity matrix is subsequently used for clustering and link reduction, which is presented in sub-section 8.1.1.3 and 8.1.1.4 respectively.

As discussed in Chapters 4 and 5, a major issue in co-occurrence analyses is the choice of proximity measures. The relative strength value of a co-citation relation between

two cited documents may differ depending on the proximity measure used to calculate this value. As discussed in Chapter 4, proximity measures have different focus and composition, which leads to differences in strength values for identical co-occurrence relations. For example, it has been argued that in co-citation studies the cosine measure deals more effectively with relations between high- and low-cited papers than does the Jaccard measure (Small & Sweeney, 1985).

The question is whether this difference in strength value is monotonic or not. Monotonic proximity measures differ in magnitude (strength value) but *not* in the rank ordering of the objects (e.g., Braam, Moed and Van Raan, 1988). As a consequence, monotonic proximity measures should produce comparable clustering results, provided that threshold values are adjusted for the proportionate difference in co-occurrence strength values produced by the different measures. Conversely, non-monotonic proximity measures very likely produces different clustering results.

It has been shown that the cosine and Jaccard measures can be monotonic to each other in their ranking of co-cited documents (Braam, Moed & van Raan, 1988; Hamers et al., 1989). Both studies by Braam, Moed and Van Raan (1988) and Hamers et al. (1989) conclude that identical rankings occur with the cosine strength value twice the strength value of the Jaccard measure. However, proximity measures can behave differently from data set to data set (Hubálek, 1982; Ellis, Furner-Hines & Willett, 1994). It is therefore sensible to investigate whether different proximity measures produce similar or different rankings to the same set of co-citation pairs. Similar rankings indicate monotonicity of the proximity measures investigated. As a result, monotonicity validates the use of one of these measures for subsequent clustering. As mentioned in Chapter 4, Hubálek (1982) suggests that empirical studies would do well to select at least two different proximity measures in order to compare the results. Hubálek (1982) suggests a linear and a non-linear measure as defined in Chapter 4.

The Jaccard association measure is non-linear and the cosine angular association measure is linear as presented in Chapter 4. The Jaccard and cosine proximity measures are the most widely used measures in document co-citation studies (e.g., Small, 1973; Salton & Bergmark, 1979; Gmür, 2003). Moreover, the Jaccard is the preferred association measure by Small and colleagues at ISI (e.g., Small, 1973; Small & Greenlee, 1980; Gmür, 2003). Jaccard is also applied as the primary proximity measure in the present case study of periodontology. In order to validate the application of the non-linear Jaccard association measure, the co-citation results produced by this measure is compared to those of the linear cosine measure. The degree of monotonicity between the Jaccard and cosine measures is tested by use of

the non-parametric Mantel test for correlation between two proximity matrices (Mantel, 1967; Mantel & Valand, 1970).

The proximity measures are thoroughly discussed in Chapters 3 and 4, and related to document co-citation analysis in Chapter 5. Nevertheless, we briefly recapitulate their major characteristics below. The binary composition of the two proximity measures applied in the present co-citation study is illustrated by use of a $p \times p$ contingency table, where p is equal to the number of possible values that the components of a vector can take (Anderberg, 1973; Ellis, Furner-Hines & Willett, 1994). In the present binary case, where component values are either 0 or 1, p is equal to 2.

TABLE 8.1. 2×2 contingency table from Anderberg (1973, p. 83).

Vector A \ Vector B	1	0	Totals
	1	0	
1	a	b	$a + b$
0	b	d	$c + d$
Totals	$a + c$	$b + d$	n

A letter (a , b , c , and d) in Table 8.1 indicates the combination of *present-absence* values (0, 1) for a component between the pair of vectors. For example, the letter a indicates that the component is present in both vectors, whereas b indicates that the component is present in vector A but absent in vector B .

Table 8.2 below illustrates the binary forms of the two proximity measures investigated for their monotonicity in the present case study. The binary forms are deduced from the contingency table (Table 8.1) presented above (Ellis, Furner-Hines & Willett, 1994, pp. 137-138).

TABLE 8.2. Binary formulas of the Jaccard and cosine proximity measures.

ID	Common name	Binary formula	Range
A	Binary Jaccard	$\frac{a}{a + b + c}$	0 to 1 (17)
A	Ochiai (cosine)	$\frac{a}{\sqrt{(a + b)(a + c)}}$	0 to 1 (18)

The non-linear Jaccard and the linear cosine belong to the *association* type of proximity measures. In Chapter 4, we demonstrated that association measures are based on the inner product (denoted by a in Table 8.1). The inner product concerns the *co-occurrence* of components between two vectors.

When choosing a proximity measure for binary data, the choice generally depends on how to handle co-absences (Everitt & Rabe-Hesketh, 1997). Co-absences are denoted by d in Table 8.1. In the case of co-citations, it gives no semantic meaning to incorporate co-absences in the formula. The binary $n \times m$ matrix is sparse, thus if co-absences are incorporated into the study most co-citation pairs will become very similar. This makes it impossible to discriminate between the co-citation pairs. Thus, the type of measure we are interested in is one that only account for asymmetric binary values in the $n \times m$ matrix. In other words a measure that only counts co-occurrences in the numerator, such as the inner product (a) in the Jaccard and cosine association measures.

Proximity measures that exclusively focus on co-occurrences in the nominator differ in the way they 'normalize' these co-occurrence counts in the denominator. The most pronounced normalization component used is individual occurrences of the pair of components. The letters b and c in Table 8.1 concerns the *occurrence* of a component in one of the two vectors. As can be observed from Table 8.2, b and c are handled very differently between the Jaccard and cosine measures.

The Jaccard measure in its binary form is defined as the conditional probability that a randomly chosen component will score 1 on both vectors, given that components with 0–0 matches (co-absences) are discarded first. The co-absences are treated as totally irrelevant (Anderberg, 1973 p. 89). The Jaccard measure is non-linear as it is based on set theory (van Rijsbergen, 1979). As described in Chapter 3, the cosine measure is linear as it is defined as the inner product of two vectors, divided by the product of their Euclidean lengths, i.e., the square root of the sum of squares of all components in the two vectors. In its binary form, the sum of squares corresponds to $a + b$ and $a + c$ in Table 8.1.

Surely, the Jaccard and cosine measures will not derive equal strength values due to their different normalization procedures. More importantly, the test for monotonicity will determine if the different normalization procedures of the two measures yields different ranking of the 2001 cited documents. Consequently, in the context of the document co-citation analysis, the Jaccard association measure is defined as follows:

$$S_J(i, j) = \frac{\text{coc}(i, j)}{\text{cit}(i) + \text{cit}(j) - \text{coc}(i, j)} \quad (19)$$

Where $S_J(i, j)$ denotes the co-citation strength between the cited documents i and j , as calculated according to Jaccard's formula, $\text{cit}(k)$ denotes the number of citations received by the cited document k ($k = i$ or j), and $\text{coc}(i, j)$ is the number of co-citations received by i and j . This number can also be described as the number of cited documents in the intersection of the set of all citations to i and the set of all citations to j , divided by the number of cited documents in the union of these two sets.

Our case study commences with a binary matrix of 801 citing documents and 64 cited documents. The matrix generation is carried out by use of the Bibexcel software (www.umu.se/inforsk/). The 'enhanced document representations' are stored in a text file, and loaded into Bibexcel, where the co-occurrence module creates an $n \times m$ matrix. Subsequently, the $n \times m$ matrix is transformed into two $n \times n$ symmetric proximity matrices (64×64) by use of the statistical package XLSTAT® (www.xlstat.com). The results are two co-occurrence matrices of relative co-citation counts between the 64 cited documents chosen for the study. Accordingly, the two co-citation matrices are a Jaccard matrix and a cosine matrix presented together in Appendix 3.

8.1.1.2.1 Mantel correlation statistic for test of monotonicity between two proximity matrices

The purpose of the Mantel statistic in its present application is to test whether the relative co-citation values in the produced proximity matrices are monotonic to each other. This is done by investigating if the values are rank correlated. The basis is the values from the Jaccard matrix. The Mantel statistic investigates whether the results from the alternative cosine matrix are rank correlated with the results from the Jaccard matrix. Thus, we test the monotonicity between relative co-citation strength values by comparing the two proximity matrices.

It is problematic to apply a traditional correlation analysis, moment or rank correlation, to test the correlation between two matrices, even if the matrices are reformatted into vectors (Aldenderfer & Blashfield, 1984, p. 64). Since the proximity values in each matrix are not independent from each other, the application conditions of the traditional correlation tests are violated, whether they are parametric or non-parametric tests (Everitt & Dunn, 2001). There is a need therefore for specific tests of correlations between two matrices; the best known is the Mantel test (Mantel, 1967).

The standardized Mantel statistic finds the correlation between two alike proximity matrices (i.e., similarity, dissimilarity, or correlation matrices), and evaluates the significance of the statistic by permuting rows and columns of the proximity matrix (Sokal & Rohlf, 1995). The statistic is evaluated either as a moment correlation or as a rank correlation.

As the components of proximity matrices are not independent, Mantel's test of significance is evaluated via permutation procedures⁴⁵ (Sokal & Rohlf, 1995). A correlation coefficient (the Mantel statistic) is calculated for the original matrices. Subsequently, the rows and columns within the matrices are randomly rearranged, and new correlation coefficients are recalculated. The distribution of values for the statistic is generated via thousands of iterations (Sokal & Rohlf, 1995). If the original matrices are correlated, the disruption caused by the permutations should reduce the correlation coefficient (Sokal & Rohlf, 1995).

When the dimension of the matrix is too high ($n > 10$), it is almost impossible to compute the statistic for all possible permutations. However, it is possible to sample randomly in the permutations to obtain a precise estimate of the p -value (the precision increases with the number of random permutations).

Mantel's statistic is based on simple cross products:

$$Z = \sum_{i=1}^n \sum_{j=1}^n x_{ij} \cdot y_{ij} ; \quad (20)$$

and the standardized correlation formula:

$$r = \frac{1}{(n-1)} \sum_{i=1}^n \sum_{j=1}^n \frac{(x_{ij} - \bar{x})}{s_x} \cdot \frac{(y_{ij} - \bar{y})}{s_y} \quad (21)$$

Where x and y are proximity values measured at locations i and j and n is the number of components in the symmetric proximity matrices [$n(n-1)/2$], and the s_x and s_y are standard deviations for proximity value x and y .

Notice that the Mantel statistic is based on linear correlation and hence is subject to the same assumptions that beset a common product moment correlation test. Alternatively, the Mantel statistic can be performed as a Spearman rank order correlation test instead. As reported in Chapter 7, a linear relationship is a special case of the more general monotonic relationship between two variables. Rank correlation

⁴⁵ Permutation corresponds to Monte Carlo procedures, where random number generators are used to create sample data sets from an original data set (Aldenderfer & Blashfield, 1984).

investigates monotonic relationships, which is also the aim of the present investigation. Consequently, as the proximity values do not conform to the normal distribution, and since we are interested in monotonicity, the Mantel statistic is evaluated as a rank correlation.

The Mantel test is carried out in XLSTAT⁴⁶ (www.xlstat.com) on the Jaccard and cosine matrices, which are of the same type, that is, similarity matrices. The results are presented below in Table 8.3; the null hypothesis is that no monotonicity exists between the two compared matrices.

TABLE 8.3. *Mantel test of monotonicity: Jaccard versus cosine similarity matrix.*

Mantel test of monotonicity: Jaccard matrix (A) versus cosine matrix (B)	
Rank correlation r_s (Jaccard, cosine)	1.000
Number of random permutations	100000
Two-tailed p-value	< 0,0001
Alpha	0,05

At the level of significance $\alpha^2=0.05$, the decision is to reject the null hypothesis of absence of rank correlation between matrices A and B. In other words, the rank correlation between the Jaccard matrix and the cosine matrix is significant. The Mantel test indicate joint monotonicity between the Jaccard and cosine proximity values. The result substantiates former results by Braam, Moed and Van Raan (1988) and Hamers et al. (1989).

Joint monotonicity indicates a certain degree of robustness in the choice of proximity measure (Sneath & Sokal, 1973). Because of the joint monotonicity, we can expect comparable clustering results if the cosine measure is applied instead of the Jaccard measure. Conversely, the absence of joint monotonicity means that clusters formed by the same clustering algorithm will have different components depending on the nature of the proximity measures.

The monotonicity results in the exclusive application of the Jaccard proximity for document co-citation clustering.

⁴⁶ The Mantel test procedure is available in the XLSTAT package (www.xlstat.com).

8.1.1.2.2 Adjustment of the Jaccard proximity matrix

As mentioned above, the proximity matrix can be used to differentiate between 'significant' and 'non-significant' pairs of co-cited documents. Obviously, the determination of significance is highly subjective and depends on the study at hand.

Griffith et al. (1974) show that an initial integer citation frequency threshold, at least in the sciences, most likely selects a number of very highly cited documents reporting on methods and procedures. Some of these highly cited documents can cause problems to the subsequent co-citation clustering (Small & Greenlee, 1980). According to Small and Greenlee (1980) 'method' papers simply do not cluster at most levels since they are not cited with any other papers a sufficiently high percentage of the time. For example, a cited reference that is included in a co-citation analysis because it has a high citation count, does not necessarily have a high co-citation count with any of the other selected cited references. In such cases, it has low co-citation counts with many of the other selected cited references. If such references are included in the analysis, they most likely create arbitrarily clustering results, as they have problems 'finding a cluster where they naturally belong'. As a result, the initial citation count of a select cited reference has to be compared to its distribution of co-citations among the other selected cited references in a set. The Jaccard association value is a reflection of this relation. The normalization of the co-citation count by individual citation counts (b and c in Table 8.1) estimates the 'significance' of the cited reference in relation to co-citation clustering.

The purpose of the second component of the proposed methodology is to create concept groups. Concept groups are solid clusters of significantly co-cited reference pairs. Thus, so-called 'method papers', that is, cited references that are not co-cited a significant number of times with other selected cited references, are inappropriate for the present cluster analysis. Hence, such cited references are identified and removed from the proximity matrix. The removal of such 'method papers' from the set of cited references is dependent on a threshold value. 'Method papers' will have low Jaccard co-citation values due to the composition of the Jaccard measure described above.

Determination of a threshold value is arbitrary. In the present study, 'trial and error' has been used to establish a threshold value that yields clusters containing at a minimum two components. Small and Greenlee (1980) report on a Jaccard threshold value of 0.18 in their studies. For the 2001 sample, we apply a threshold value 0.16. This implies that the Jaccard proximity matrix generated above is reduced, so that it only contains cited references that have at least one co-citation link of 0.16 or above.

The discarded highly cited references are not considered further in the present study. But, as Small (1978) and Rees-Potter (1987; 1989) has shown, such 'method papers' most likely act as concept symbols when subjected to citation context analysis.

Of course, this makes them interesting for thesaurus construction purposes, as investigated by Rees-Potter (1987; 1989). Nevertheless, the fact that the papers have high citation counts and low relative co-citation values means that the use of these papers are scattered among several research fronts in the citing literature. As a result, the cited references, while surely a part of the intellectual base, have problems finding a significant cluster membership. As mentioned above, in the present study we are only interested in the creation of solid concept groups with a minimum of two cited references, thus cited references violating the threshold value are discarded.

However, nothing prevents these highly cited references from being incorporated into a citation context analysis for thesaurus construction purposes. They will, however, most surely disrupt the creation of concept groups, if grouped by complete-link clustering. The result is either arbitrarily composed clusters, or a number of singleton clusters, that is, a string of concept groups containing one 'method paper' each. A more suitable approach would be to apply single-link clustering instead. Although, this algorithm does not produce the desired densely packed and closely defined clusters as in complete-link clustering (Small, 1986).

Enforcing the threshold value upon the initial 64×64 Jaccard matrix, results in a new reduced 45×45 Jaccard matrix shown in Appendix 4. Appendix 2 shows the top 64 cited documents initially chosen for the document co-citation analysis, the appendix is similar to Appendix 5. However, in Appendix 5, the cited references marked in grey are the 19 highly cited references removed before the clustering analysis. The new reduced matrix is the basis for the document co-citation clustering, which is presented in the following sub-section. The full bibliographic descriptions of the 45 individual cited references are presented in Appendix 12.

8.1.1.3 Document co-citation cluster analysis

The present and the following sub-sections describe the ordination techniques of dimensionality and link reduction applied to the set of 45 cited references in order to create a number of concept groups for thesaurus construction purposes. This corresponds to literature mapping, where the basic aim is to improve simplicity and clarity of the hidden structures embedded in the proximity matrix. Dimensionality reduction algorithms, such as Multidimensional Scaling, Principal Components Analysis, and Cluster Analysis, can reduce implicit links (dimensions), whereas, link reduction algorithms, such as Minimum Spanning Trees (MST) and Pathfinder Networks (PFNET) can reduce explicit links (connections). Dimensionality and link reduction can therefore be an effective way to derive useful representations of high-dimensional data.

Cluster analysis is a special case of dimensionality reduction. In essence, clusters should be internally homogenous (components are similar to one another) and externally heterogeneous (components are not like members of other clusters).

The purpose of the present cluster analysis is to divide the set of 45 cited references into a number of clusters that best match their co-citation structure. This is a reduction of the initial 45 dimensions. As a result, we structure the intellectual base for the 2001 sample text corpus in order to exploit the dyad relationship between the cited and citing documents in the following third component of the methodology.

As presented in Chapter 4, we apply agglomerative hierarchical clustering to produce the nested data set in which pairs of cited references are successively joined. The clustering algorithm chosen for the study is complete linkage due to its strong cluster criterion, and its ability to create 'clique-like' clusters. This is deemed advantageous in connection with concept group creation, where the aim is to maximize intra-cluster similarity and minimize inter-cluster similarity (Sparck Jones, 1971).

Appendix 6 shows a dendrogram of the 2001 cluster result. Iterative clustering resulted in a decision to truncate the partition at 13 clusters. Table 8.4, on the following page, presents the 13 clusters, their size and abbreviated name of cited references, as well as the median age of publication for the cited documents. Further, a cited reference that is a review paper is marked with [R].

A few comments can be made about the derived clusters of interest for the subsequent analyses. Most importantly, *all* cited references are journal papers, in contrast to the study of Rees-Potter (1987; 1989). As a result, we are able to investigate a specialty area where the primary mediator of knowledge claims is journal papers. Of 45 cited references, 10 are review papers. Further, the 13 clusters have a mean size of 3.6 cited documents. The median citation age of all clusters is 7 (1994). Notice the basis of the analysis is 2001. Median publication ages for the individual clusters are given in Table 8.4. Median ages are very important in the investigation of conceptual origins and developments. Likewise, they give an indication of age composition of the intellectual base. For example, cluster 12 has a median age of 37.5 years. This is by far the oldest intellectual base cluster for the 2001 sample.

As introduced in Chapter 7, the two cited documents (LOE_63 and SILNESS_64) belong to the string of documents published in the mid-1960s on experimental gingivitis, which eventually came to define the infection/host response paradigm in periodontology (Armitage, 2002). Some of these documents are still cited in 2001 and comprise a central part of the intellectual base; in fact the two documents are the most cited documents overall in the 2001 sample.

TABLE 8.4. Cluster result of 2001 sample containing 45 cited references. Median ages of the publications in the clusters are indicated in the right column.

Cluster	Size	Cited documents	Median age of publication
1	4	HAMMARSTROM_97a [R]; HAMMARSTROM_97b; HEIJL_97a; HEIJL_97b	4 years
2	3	GOTTLOW_86; NYMAN_82a; NYMAN_82b	19 years
3	6	BECK_96; DESTEFANO_93; JOSHIPURA_96; MATTILA_89; OFFENBACHER_96a; [R] OFFENBACHER_96b	6.5 years
4	5	GARRETT_96 [R]; HAMP_75; HIRSCHFELD_78; MCFALL_82; OLEARY_72	21 years
5	5	GENCO_96 [R]; GROSSI_94; GROSSI_95; KORNMAN_97; PAPAPANOU_96 [R]	5.5 years
6	2	HOLT_88; PIKE_94	10 years
7	2	HOLT_99 [R]; SLOTS_99 [R]	2 years
8	2	MASADA_90; STASHENKO_91	10.5 years
9	5	HAFFAJEE_94 [R]; LOESCHE_85; MOORE_94 [R]	7 years
10	2	SOCRANSKY_94; SOCRANSKY_98	5 years
11	3	BRAGD_87; SLOTS_82; SYED_72	19 years
12	2	LOE_63; SILNESS_64	37.5 years
13	6	BERGSTROM_89; BERGSTROM_94 [R]; HABER_93; MACHTEI_97; STOLTENBERG_93; ZAMBON_96b	7.5 years

The complete-link clustering criterion states that any candidate for inclusion into an existing cluster must be within a certain level of proximity to *all* component members of that cluster (Aldenderfer & Blashfield, 1984). A cluster in which all components are linked to all others is said to be complete. The extent to which a cluster is complete is indicated by its density, which is defined as the number of links divided by the maximum number possible; i.e., $n(n - 1)/2$ for a symmetric matrix. Complete clusters are also called *cliques*. A clique is defined in graph theory as a maximal complete sub-graph. A sub-graph of a graph G is a graph whose nodes and links are

contained in G^{47} . A complete sub-graph is a sub-graph of G that is *complete* and is maximal in the sense that no other node of G could be added to the sub-graph without losing the completeness property.

As all components in the present study are linked due to the complete-link clustering criterion, the density of clusters is one (maximum). Consequently, the result of the document co-citation analysis of the 2001 sample of cited references is 13 complete-link clusters of the prototypical clique type as presented in Chapter 4. In relation to document co-citation analysis, cliques type clusters are "...a clear indicator of the formation of a school of research" (Gmür, 2003, p. 33). Likewise, in relation to thesaurus construction, Sparck Jones (1971, p. 56) argue that cliques indicate conceptual formations. Thus, the individual clusters in the intellectual base for the 2001 sample text corpus most likely reflects salient 'schools of research' or 'research topics'. The dyad relationship between the cited and citing documents makes it possible to identify terminology in citing papers that characterize and conceptualize the 'school of research' or 'research topics' identified through the clusters. In the following section 8.2, clique type clusters are transformed into concept groups, which contain terminology that reflect the cited references comprising the clusters and define the potential 'school of research' or 'research topics'.

8.1.1.3.1 Cophenetic correlation statistic for test of match between cluster result and original proximity values between co-cited pairs of references

Hierarchical clustering techniques impose a hierarchical structure on the data set. It is advisable to consider whether the structure is merited or whether it introduces unacceptable distortions of the original relationships among the components as implied by their proximity values (Everitt & Dunn, 2001). The method most commonly used for assessing the match between the derived dendrogram (tree) and the original proximity values is the 'cophenetic' correlation coefficient (Sokal & Rohlf, 1962: Sneath & Sokal, 1973). Traditionally, the cophenetic correlation coefficient is simply the product moment correlation (Sokal & Rohlf, 1962)

By use of a dendrogram, it is possible to create an implied proximity matrix (cophenetic matrix) that shows the similarities between all pairs of entities as suggested by the hierarchical solution.

Consult the dendrogram in Appendix 6, which shows the complete linkage clustering solution using Jaccard's coefficient to the 45 cited references from the 2001 sample. Each similarity value marked in the tree represents the Jaccard proximity

⁴⁷ In graph theory, a matrix is defined as a graph G ; a sub-graph is thus a part of the matrix such as a cluster.

value at which the respective pair of cited references was merged into a common cluster. An important point to note regarding the implied cophenetic matrix is that there are at most only $n - 1$ (i.e., 44) unique values in this matrix since hierarchical agglomerative methods always require $n - 1$ merger steps. In contrast, the original proximity matrix contains $n(n - 1)/2$ unique values, excluding ties.

The cophenetic correlation is the correlation between the values in the original proximity matrix and the values in the implied cophenetic matrix. The original Jaccard proximity matrix of the 45 cited references is illustrated in Appendix 4, and the corresponding cophenetic proximity matrix is presented in Appendix 7.

Despite its rather frequent use, the cophenetic correlation does have distinct problems (Aldenderfer & Blashfield, 1984). Firstly, the use of the product moment correlation assumes normal distributions of the values in the two matrices being correlated. This assumption is generally violated for the values in the implied cophenetic matrix since the clustering method used largely determines the distribution of the proximity values in this matrix. Thus, the use of the correlation coefficient is not an optimal estimator of the degree of correlation between the values in the two matrices. Secondly, since the number of unique values in the implied proximity matrix is much smaller than the number of unique values in the original proximity matrix, the amount of information contained in the two matrices is quite different.

Nevertheless, Rohlf and Fisher (1968) studied the distribution of this type of correlation under the hypothesis that the individuals are randomly chosen from a single multivariate normal distribution. They found that the average value of the coefficients tends to decrease with n . They also suggested that values of the cophenetic correlation above 0.8 were usually sufficient to reject the null hypothesis. However, in a later paper Rohlf (1970) warns that even high cophenetic correlation coefficients may not be a guarantee that the dendrogram serves as a sufficiently good summary of the original matrix. Holgersson (1978) suggests that a more reliable cophenetic correlation coefficient can be obtained by use of random permutations. This is similar to the Mantel statistic introduced in section 8.1.1.2.1. Consequently, we apply random permutations in order to perform the cophenetic correlation test of the match between the derived cluster results and the original proximity values in the Jaccard matrix. The results are presented in Table 8.5 on the next page.

At the level of significance $\alpha=0.05$, the decision is to reject the null hypothesis of absence of correlation between the Jaccard and cophenetic matrices. In other words, the correlation between the two matrices is significant.

TABLE 8.5. *Cophenetic correlation test.*

Cophenetic correlation test: Jaccard matrix (A) versus cophenetic matrix (B)	
Cophenetic correlation r_c (Jaccard, cophenetic)	0.844
Number of values in the cophenetic matrix (maximum = 44)	27
Number of random permutations	10000
One-tailed p-value	0.0001
Alpha	0.05

With the above-mentioned criticisms in mind, we do conclude that the cluster results are a good representation of the original proximity values. Section 8.2, concerning citation context analysis, elaborates on the results from the cophenetic correlation test. As the result of the cophenetic test indicates good cluster solutions, we expect that the cited references in the clusters are semantically related. Due to dyad relationship between cited and citing documents, we also expect this similarity to be reflected in the terminology used to describe these references in the citing papers. If this is the case, then we have created a solid concept group.

8.1.1.4 *Pathfinder network scaling*

In Chapter 5, the notion of literature mapping is introduced. Usually, a map is a spatial representation of how disciplines, fields, specialties, and individual papers or authors are related to one another as shown by their physical proximity and relative locations (Small, 1999). Lin (1997) has provided a useful typology of the various styles of representation, including hierarchical, network, scatter, and map displays. Maps are used to impose a structure on a collection of objects, as we find arranging information in space a natural and useful heuristic tool (e.g., Small, 1999; Chen, 1999; 2003). It therefore seems reasonable that creating spatial environments with information objects distributed in a stable and meaningful fashion has the potential of enhancing information usability and retrieval.

In the case of scientific literature, a spatial representation can facilitate our understanding of conceptual relationships and developments (Small, 1978; 1986; Schvaneveldt, 1990; Chen, 2003; White et al., 2004). Hence, the purpose of applying Pathfinder network scaling in the present analysis is to facilitate the creation and analysis of concept groups for thesaurus construction purposes.

Hierarchical cluster analysis creates potential concept groups by focusing on the most salient relations in the matrix through dimensionality reduction. The proximity matrix, however, still contains important relations not directly expressed in the cluster results. From the point of view of concept group creation, these relations are of importance, as they can visualize the internal and external cluster structures. We are therefore interested in the extraction and mapping of the most important co-citation relations. Pathfinder network scaling is an appropriate algorithm for such purposes.

Pathfinder network scaling is a structural modelling technique originally developed for conceptual analysis based on proximity data in psychology by Schvaneveldt (1990). Pathfinder network scaling extracts underlying patterns in a matrix of proximity values and represents them spatially in a class of networks called Pathfinder networks (PFNETs) (Schvaneveldt, 1990).

The Pathfinder algorithm defines a network given estimates of proximities between pairs of objects, and preserves only the most ‘important links’ as defined in the proximity matrix. The resulting PFNET consists of the objects as nodes and a set of links⁴⁸, which connect the nodes. As a result, the PFNET retains the same set of nodes as the original matrix, but the number of links in a PFNET is greatly reduced. This means that Pathfinder network scaling is an effective link-reduction algorithm that prevents a network from being cluttered by too many links (Chen, 2003). The Pathfinder algorithm treats proximity values as weights on the paths that can be drawn between nodes representing any two cited references in the input set. The Pathfinder algorithm selects the lowest-weight paths (also called minimum-distance or minimum cost paths) between nodes to render the most salient ties. In our matrix, the closest connections are signalled by the Jaccard similarity measure, where high values signal high co-occurrence counts. The similarity values therefore require a transformation (subtraction from a constant) to convert them to a distance measure before PFNETs are actually plotted.

The PFNET algorithm compares path weights over both direct (one link) and indirect (multilink) paths between nodes. It retains just those links that constitute minimum-weight paths. It is required that such paths do not violate the triangle inequality. Triangular inequality is an important property of a Euclidean space, which specifies that the distance between two objects is less than or equal to the distance or path connecting the two objects via a third point. This implies that minimum-weight paths in a PFNET will be direct unless an indirect path is computed to be shorter. Thus, the triangle inequality eliminates ‘redundant’ or ‘counter-intuitive’ links

⁴⁸ Links can be directed or undirected for symmetrical or non-symmetrical proximity values.

(Schvaneveldt, 1990). Triangle inequality and its role in defining distance metrics is introduced in Chapter 4.

Given two links or paths in a network that connect two nodes, the link/path is preserved that has a greater weight⁴⁹ defined via the Minkowski distance metric (9). It is assumed that the link/path with the greater weight better captures the interrelationship between the two nodes and that the alternative link/path with less weight is 'redundant' or perhaps 'counter-intuitive' and should be pruned from the network (Schvaneveldt, 1990). Two parameters r and q influence the topology of a PFNET (Schvaneveldt, 1990). The r -parameter influences the weight of a link/path based on the Minkowski distance metric. For $r = 1$, the path weight is the sum of the link weights along the path (the Manhattan distance); for $r = 2$, the path weight is computed as the Euclidean distance; and for $r = \infty$, the path weight is the same as the maximum weight associated with any link along the path (Schvaneveldt, 1990). An advantage of $r = \infty$ is that one only needs to assume that the original distance estimates have ordinal properties. Another advantage is that the link structure will be preserved for any monotonic transformation of the data (Fowler, Wilson & Fowler, 1992).

The parameter q sets the range within which all paths of length q will be examined in the test of the triangle inequality and removed if they violate it. The larger the value of q , the more extensive the triangle inequality constraint; therefore, links are more likely on a path that violates the rule. Increasing the value of the parameter r or q can reduce the number of links in a network. A network of n nodes can have a maximum path length of $q = n - 1$. With $q = n - 1$ the triangle inequality is maintained throughout the entire network. A PFNET with parameters set at $r = \infty$, and $q = n - 1$ has the least number of links. These settings are widely used in Pathfinder research because they tend to produce networks that are highly intelligible simplifications of the data (White et al., 2004).

Figure 8.3 illustrates how a link is removed if it violates the triangle inequality. The example is from cluster 2 of the 2001 sample set. Interestingly, the original Jaccard similarity values are transformed into dissimilarities in order to apply the Minkowski distance metric.

⁴⁹ Notice that the Minkowski measure is a distance metric, which means that 'greater weights' are closer to zero and 'lower weight' are closer to 1.

FIGURE 8.3. Link reduction procedure due to the triangle inequality. Two paths connect GOTTLLOW_86 and NYMAN_82a. If $r = \infty$ path 2 is longer than path 1, violating the triangle inequality, so it needs to be removed (W =dissimilarity weight).

A force-directed graph-drawing algorithm ('spring-embedder') determines the spatial layout of a Pathfinder network, where link distance is uniformly rendered, and void space has no semantics in its own right (Kamada & Kwai, 1989; Fruchterman & Reingold, 1991). As a result, connectivity and paths are the predominant objects for interpretation. This makes PFNETs scale-free networks, as opposed to for example MDS.

According to Chen (1999, p. 408), Pathfinder network scaling provides "...a fuller representation of the salient semantic structures than minimal spanning trees, but also a more accurate representation of local structures than multidimensional scaling techniques". The most fundamental problem with MDS-maps is that they do not necessarily group explicit information together, so that patterns must be judged carefully to identify the underlying structure (Chen, 2003; White, 2003).

Consequently, the value of Pathfinder network scaling is its ability to reduce the number of links in a matrix in meaningful ways, which often results in a concise structural representation of clarified proximity patterns, relatively easy for our perception to detect.

We apply Pathfinder network scaling in the present second component in order to facilitate the identification of important relationships between pairs of co-cited references, within and between their parent clusters, as well as to visualize these structural patterns in a meaningful way. The pairs are linked as output only if their co-occurrence counts are the highest (or tied-highest) in their respective vectors.

In section 8.2, clusters are transformed into concept groups and cited references are transformed into concept symbols. In relation to the investigation and analysis in section 8.2, knowledge concerning central nodes within clusters and nodes that connect one cluster with another is important. Such knowledge can facilitate the interpretation

of concept symbols and concept groups, and help reveal the structural relations between the concept groups. Thus, the salient links exposed in the PFNET may indicate the centrality of concepts within clusters, as well as the structural relation between concepts and concept groups in the entire network. Further, by emphasizing only the most prominent links, PFNETs reduce the user's cognitive load in interpreting the most important relationships depicted in the map.

8.1.1.4.1 PFNET solution for the 2001 sample of cited references in periodontology

Figure 8.4 below presents the PFNET solution to the Jaccard matrix of 45 cited documents from 2001 sample. The PFNET parameters are set to $r = \infty$, and $q = n - 1$.

FIGURE 8.4. PFNET of the 2001 Jaccard matrix visualized with Pajek (Batagelj & Mrvar, 1998).

The colours of nodes, as well as bracket numbers in Figure 8.4 indicate the partition of the network based of the complete-link clustering result presented above. Please note that there is a perfect match between the clusters derived from the cluster analysis and

the placement and linking of nodes in the PFNET. Some $45(44)/2^{50} = 990$ possible links are reduced to the 45 most salient links. Two types of structural information, inferred from the PFNET, are of interest to the present study: 1) the internal structure within the individual clusters; and 2) the external structure or connectivity between the clusters.

As can be observed from Figure 8.4, the original clique type structure of each cluster is broken down so that the internal cluster structure resembles strings, except cluster 5 and cluster 11. The former resembles a star and the latter a clump. Initial indications of centrality of nodes within clusters and the connectivity of nodes between clusters, are given by visual inspection of the PFNET in Figure 8.4 and presented in Table 8.6.

TABLE 8.6. *Central and connective nodes deduced by visual inspection of the 2001 periodontology intellectual base PFNET.*

Cluster	Central cited references	Connective cited reference
1	HAMMARSTROM_97b	HEIJL_97a → Cluster 4 HEIJL_97b → Cluster 2
2	NYMAN_82b	GOTTLOW_86 → Cluster 1
3	BECK_96	BECK_96 → Cluster 6 JOSHIPURA_96 → Cluster 5
4	<i>n/a</i>	GARRETT_96 → Cluster 1 HIRSCHFELD_78 → Cluster 5
5	GROSSI_95	GENCO_96 → Cluster 3 PAPAPANOU_96 → Cluster 4 KORNMAN_97 → Cluster 9 GROSSI_95 → Cluster 13
6	HOLT_88	HOLT_88 → Cluster 3 and 8
7	<i>n/a</i>	HOLT_99 → Cluster 9
8	<i>n/a</i>	MASADA_90 → Cluster 6
9	MOORE_94	HAFFAJEE_94 → Cluster 5, 7 and 11
10	<i>n/a</i>	SOCRANSKY_94 → Cluster 13
11	SLOTS_82	SLOTS_82 → Cluster 9 and 12
12	<i>n/a</i>	SILNESS_64 → Cluster 11
13	STOLTENBERG_93; ZAMBON_96b	ZAMBON_96b → Cluster 5 and 10

⁵⁰i.e., $n(n-1)/2$; n = number of cited references in the matrix; the number is divided by 2 as the matrix is symmetrical.

An *n/a* indicates that no central node could immediately be located by visual inspection. The arrows indicate a salient link from a cited reference to a neighbouring cluster. Only the target cluster is mentioned in relation to the connective nodes.

Besides visual inspection of the network, one common way to identify important nodes in a network is by use of centrality measures (e.g., Roberts, 1997; Popping, 2000; Scott, 2000). Centrality is a structural attribute of nodes in a network. Thus, the nodes are compared to each other beyond the scope of individual clusters. Three measures of centrality are commonly used in network analysis: *degree*, *closeness*, and *betweenness* (Freeman, 1979). Degree is the number of direct links to others; closeness is the graph-theoretic distance (geodesic), i.e. the sum of all shortest paths, of a given node to all other nodes in the network; and finally, betweenness is the number of geodesic paths that pass through a node. Table 8.7 gives the scores of top-ranked cited references on normed versions of these measures as computed by UCInet⁵¹ (Borgatti, Everett & Freeman, 2002). Appendix 8 contains scores for all the cited references in the network.

TABLE 8.7. *Top ranks on three normed centrality measures (degree, closeness, and betweenness) for the 2001 periodontology PFNET.*

Degree	Score	Closeness	Score	Betweenness	Score
HAFFAJEE_94	6.107	GROSSI_95	20.465	GROSSI_95	76.004
SLOTS_82	5.723	PAPAPANOU_96	18.803	PAPAPANOU_96	40.592
GROSSI_95	4.823	GROSSI_94	18.644	HIRSCHFELD_78	38.372
HOLT_88	4.595	GENCO_96	18.487	GROSSI_94	38.372
ZAMBON_96b	3.895	ZAMBON_96b	18.033	MCFALL_82	35.941
STOLTENBERG_93	3.427	HIRSCHFELD_78	17.255	KORNMAN_97	35.941
BECK_96	3.130	KORNMAN_97	16.988	GENCO_96	35.941
PAPAPANOU_96	3.016	JOSHIPURA_96	16.730	HAFFAJEE_94	35.835
MASADA_90	3.014	MCFALL_82	15.827	JOSHIPURA_96	33.298
KORNMAN_97	2.975	STOLTENBERG_93	15.771	HAMP_75	33.298

The results of the normed centrality measures are used in the naming of concept groups investigated in section 8.2. Nevertheless, we give some examples here of what the results indicate.

At $r = \infty$, and $q = n - 1$, some references are linked to one other reference at most, while other references are linked to more than one. A reference must be relatively highly cited overall to be highly co-cited with others and perhaps linked to more than one other cited reference. The references with several links to others appear in the

⁵¹ Norming divides the score for each node by the maximum possible value for degree, closeness, or betweenness.

PFNETs as stars with entourages; they are high on the network statistic called degree centrality and can be said to dominate a research topic (White, 2003).

For example, in the PFNET above, HAFFAJEE_94 has the highest score on degree centrality in the network, making it the most ‘active’ node in the network. HAFFAJEE_94 is a ‘connector’ or ‘hub’ in this network, as is the other top-ranked cited references for degree centrality in Table 8.7.

GROSSI_95, PAPAPANOU_96, and GROSSI_94 (all belonging to cluster 5), have lower degrees of centrality than HAFFAJEE_94, yet the pattern of their direct and indirect connections allow them to access all nodes in the network more quickly than anyone else. They have the shortest paths to all other nodes, or they are closer to everyone else. This implies that these cited references, or in fact their cluster, are in the central position of the network.

A node with a high betweenness score has great influence over what ‘flows’ in the network; such nodes play a ‘broker’ role in the network. From Table 8.7 we can observe that GROSSI_95 has by far the highest betweenness score, which indicates that this reference has one of the most important positions in the network. From Figure 8.4 it appears clearly that GROSSI_95 is the central node of the entire network.

The strings that connect the network are the most salient relative co-citation relations. Notice that the network has a main ‘artery’ going from NYMAN_82a in the left corner to STASHENKO_91 in the opposite right corner. The central node in the network, i.e., GROSSI_95, contains two ‘veins’, one in each direction, which again are sub-divided into minor ‘veins’. ‘Artery’ and ‘veins’ are often used as an analogy when describing the structure of PFNETs (e.g., White, 2003). The ‘artery’ typically indicates a chain of main clusters, where the sequential ordering of nodes is non-arbitrary and possibly reveals how major topical areas in a specialty area are connected. The ‘veins’ indicate minor topical areas gemmating from the main ‘artery’. In Figure 8.5 on the following page, the betweenness score determines the size of the nodes. From Figure 8.5 it can be observed that the clusters 1, 2, 3, 4, 5, 6 and 8 have high betweenness scores and are placed on the main ‘artery’. Further, cluster 13 gemmates from cluster 5, and cluster 10 gemmates from cluster 13. Likewise, cluster 9 gemmates from cluster 5, clusters 7 and 11 gemmate from cluster 9, and finally cluster 12 gemmates from cluster 11. It is evident from Figure 8.5 that cluster 5 is central to the network.

FIGURE 8.5. 'Artery' and 'veins' of the 2001 PFNET; node sizes indicate betweenness scores.

From the point of view of concept group creation, the interesting question is what these structural relations indicate; this is investigated in section 8.2.

This concludes the methodical steps of the second component. The present component has created intellectual base clusters of cited references from the sample of 2001 citing documents within periodontology. In addition, the cluster solution has been substantiated by a PFNET. Consequently, we designate the intellectual base clusters as concept groups, which, for the moment, treat some unspecified subject matter. The following section 8.2 will investigate and analyze the cited references within these concept groups, in order to identify the subject matter, and subsequently select candidate thesaurus terms related to it. Nevertheless, first we briefly summarize and discuss the main results from the present component.

8.1.2 *Summary and discussion of results*

The second component of the exploratory methodology investigates the ability of document co-citation analysis to group semantically related cited references in corresponding concept groups. The component is comparable to vocabulary

organization in traditional automatic thesaurus construction approaches. The second component, in conjunction with the third component, investigates the first research question concerning identification of candidate thesaurus terms. Consequently, a conclusion regarding the first research question is not possible until after the exploration of the third component. Nevertheless, some conclusions can be made in relation to the initial analyses of ‘vocabulary organization’, performed in the present second component.

The document co-citation analysis produces 13 clusters from 45 cited references. The methodical steps in document co-citation analysis contain some critical points, which are indicated in Figure 8.1 as ‘diamonds’. These are the choices of threshold values and proximity measures. The threshold values in the present study are chosen for pragmatic reasons. On the one hand, the initial citation threshold must suffice that highly cited references are incorporated for the subsequent citation context analysis. On the other hand, we need to delimit the number of references in order to obtain a manageable set of references. As a result, a high citation threshold value is chosen, resulting in the selection of 64 cited references.

The co-citation threshold value is based upon iterative trial and error clustering. The basis value is set to 0.18 based on former research by Small and Greenlee (1980). Iterative clustering is applied to indicate the most appropriate value in the close range of the basis. The co-citation threshold value further reduces the set of references to 45.

Obviously, the choice of threshold values influences the subsequent analysis. The purpose of the present study is to create concept groups based on complete-link clustering. The target is therefore references with significant relative co-citation values. In studies with such a purpose, it seems appropriate to use the Jaccard measure of 0.18 as a basis (or two times the value if the cosine is applied), and then establish the actual threshold value from iterative clustering of the data set. If the study only focuses on the cited references as concept symbols, then single-link clustering seems more appropriate (Small, 1986). A citation threshold value is still needed as concept symbols are of interest. Albeit, there is no need to enforce a rigorous relative co-citation threshold value, as the aim of single-link clustering is the ‘chaining’ of objects, contrary to the narrowly focused complete-link clusters (cliques). This may very well be a suitable approach to thesaurus construction, though not investigated in the present dissertation.

The choice of proximity measure is also a critical point in document co-citation analysis. Overall, the choice is between three types: distances, similarities, or correlations. The choice in the present case is the Jaccard similarity measure, which again is a so-called association measure. The choice of the Jaccard measure is based

on exclusion. As discussed in section 8.1.1.2, we need a measure that relies on the inner product. This excludes distance measures. Correlation coefficients apply a form of the inner product. Yet, in co-citation analysis, the correlation coefficients are troublesome from a mathematical point of view due to their dependence of moments around the mean. This leaves us with similarity measures that rely on the inner product. Several such measures exist. In order to conclude that the cluster result is robust and not just a result that changes with the choice of similarity measures, monotonicity between two differently defined similarity measures is investigated, a set theoretic and a geometric defined measure. The result of the Mantel statistical test between the Jaccard and the cosine measures shows a joint monotonicity between their values, thus, confirming the use of the Jaccard measure for the subsequent clustering.

Finally, the cophenetic statistical evaluation of the match between the cluster result and the original proximity values confirm that the derived clusters are fine solutions given the clustering premises. This result is further supported by the PFNET extracted from the same Jaccard matrix. There is a perfect match between the clusters derived from the cluster analysis and the placement and linking of nodes in the PFNET. We therefore conclude that given the choices of threshold values, proximity measure, and clustering algorithm, the cluster results are satisfactory. Whether, the clusters are suitable as concept groups is determined by the citation context analysis in the following section 8.2.

8.2 Third component: Term selection (citation context analysis)

The purpose of the third component of the proposed methodology is term selection. The third component investigates the ability of citation context analysis and noun phrase parsing for selection of candidate thesaurus terms, based on the concept groups established for the 2001 sample in the previous component,. Citation context analysis and its potential for thesaurus construction are presented in the chapters 5 and 6. Noun phrase parsing and phrase indexing are presented in Chapter 4. The third component of term selection is a natural extension of the second component concerning vocabulary organization. Together they explore the first and most central research question, whether the selected bibliometric methods are able to detect candidate thesaurus terms in a specialty area within the life sciences, where knowledge claims are primarily mediated through journal papers.

It is assumed that focus on recent citation contexts within the structure of citing documents makes it possible to identify current, contextual, and agreed upon candidate

thesaurus terms. The terms are expected to reflect aspects of their concept symbol, and more generally, the common subject matter of the concept group. As a result, the proposed term selection methodology is very different from traditionally term distribution methodologies used for thesaurus construction (see Chapter 3 and 4).

In order to identify and validate candidate thesaurus terms, the citation context analysis includes the following methodical steps:

1. Selection of a sample of citation contexts for each cited reference.
2. Quantitative identification of ‘consensus passages’ in order to indicate the level of terminological agreement among citation contexts of a cited reference.
3. Shallow parsing and extraction of commonly agreed upon noun phrases from the citation contexts of individual cited references within the concept groups.
4. Concept symbol analysis, which consist in a combined investigation of ‘consensus passages’ and extracted noun phrases attached to cited references. The purpose is to establish whether the references act as concept symbols. The phrasal expression of a potential concept symbol is determined from terminology in the ‘consensus passage’, in combination with the extracted noun phrases. Frequently occurring noun phrases of a concept symbol, as well as specifically selected phrases from its ‘consensus passage’, are designated as its portfolio.
5. Evaluation of the semantic coherence and naming of concept groups in accordance with the common conceptual meaning expressed by their member concept symbols, and the portfolios attached to the latter.
6. Finally, selection and validation of candidate thesaurus from the portfolio of noun phrases attached to the individual concept groups.

Verification of concept symbols in step 4 and concept groups in step 5 ensures that the basic assumptions behind the methodology of term selection are met. The concept symbol analysis in step 4 investigates the degree of consensus and contextual usage of terminology in citation contexts surrounding a cited reference. If a consensus usage is present, then it is assumed that a cited reference acts as a concept symbol. While the concept symbol analysis ensured contextual and consensus usage of terminology in relation to individual cited references, evaluation of concept groups ensures contextualization and consensus usage of noun phrases on the group level, provided there exists a semantic coherence within the group. The evaluation of the semantic coherence of concept groups makes it possible to identify primary and secondary candidate thesaurus terms. Primary terms occur in the portfolio of several concept symbols, whereas secondary terms usually only occur in a single portfolio. The

assumption is that important candidate thesaurus terms are more frequently found among the designated primary terms.

Consequently, the concept symbol analysis and the evaluation of the concept groups are conceived of as 'filtering procedures', which ensures a more solid basis for selection of candidate thesaurus terms.

Eventually, the selected primary and secondary candidate thesaurus terms are subjected to validation procedures. The validation will establish whether the second and third components of the proposed methodology, are able to select relevant candidate thesaurus terms. The results of the validation lead to an answer of the first research question concerning the applicability of bibliometric methods for selection of candidate thesaurus terms.

The validation procedures consist of matching primary and secondary candidate thesaurus terms with corresponding MeSH® descriptors, as well as periodontal terms defined in the *Glossary of Periodontal Terms* (2001). As described in Chapter 2, MeSH® is the standard authority in relation to index terms concerning periodontology. Likewise, *Glossary of Periodontal Terms* (2001) is a standard definition produced by the American Academy of Periodontology.

The following sub-sections present, discuss, and validates, the methodical steps of the third component. Sub-section 8.2.1 concern step one citation context selection. Sub-section 8.2.2 treats step two, identification of 'consensus passages'. Sub-section 8.2.3 concerns step three of noun phrase parsing. Sub-section 8.2.4 reports on the concept symbol analysis of step four. Sub-section 8.2.5 concerns step five, the evaluation and naming of concept groups. Sub-section 8.2.6 reports on step six, selection and validation of candidate thesaurus terms; and finally sub-section 8.2.7 briefly summarises the findings of the presented in the previous and present sections concerning term selection.

8.2.1 Citation context selection

In a citation context analysis, the citing papers are the objects of study. The preliminary step in the citation context analysis is to identify the text windows in the document structure, which are the target for concept symbol analyses and noun phrase parsing.

A set of rules is developed to define a citing paper and a citation context for the purpose of the present study. Any citing papers or citation contexts that did not meet the rules were discarded and a substitution was identified if possible.

- The citing paper must be in the English language.

- A reference marker is required that is actually embedded in the text of the citing document.
- Exact quotations from the cited papers are not used because we are interested in the terminology used by later citing authors.
- Traditionally, Small (1978; 1986), O'Connor (1982; 1983), and Rees-Potter (1987; 1989) have designated a citation context according to a certain sentence limit, typically up to three sentences from, or surrounding, the reference marker. We apply a less rigid citation context limitation. Almost all citing documents from the 2001 sample used in the present analysis are available in electronic form. The electronic form makes the manual process of citation context selection less cumbersome and swifter. As a rule thumb, we aim at citation contexts with the least possible number of sentences, but still sufficient for construction of a meaningful citation context. Most of the time, one to three sentences suffice, but in some cases we surpass three sentences in order to construct a meaningful citation context.
- Contrary to Rees-Potter (1987; 1989) and in accordance with Small (1986), we do not discard so-called 'perfunctory' or redundant references. As discussed in Chapter 5, 'perfunctory' references are cited references mentioned or acknowledged but not explicitly described in a citation context (Moravcsik & Murugesan, 1975). We believe that 'perfunctory' or redundant references can serve a function in relation to the present study. Such references may, for example, signal simultaneous and independent discovery, or indicate the availability of more than one good source for the same concept (Small, 1982). For example, the 2001 sample of citing papers from the specialty area of periodontology have a mean of 39 references per paper, with a mean paper size of nine pages (median of eight). This indicates a high consumption of cited references within periodontology. There seems to be a logical pattern in the consumption, as especially the reporting of experimental studies is divided into several small-sized journal papers. For that reason, an active publication pattern can be observed. It is therefore apparent that some of these papers are so closely related in origin, that when later citing authors conceptualize their findings and express them in more general terms, he or she may very well need to refer to several of these cited references. This leads to citation contexts containing what would immediately be regarded as 'perfunctory' or redundant references. Consequently, it may very well be that two or more related cited references jointly denote the same concept symbol. For some examples see cluster 2 and 13 in Appendix 9. The purpose of the present concept symbol analysis is to identify terminology preferably attached to individual cited

references. By allowing 'perfunctory' references, it becomes difficult to distinguish which words and phrases belong to which references in a citation context. But, it is assumed that the sample of citation contexts eventually yield sufficient information concerning the individual cited references, in order to judge whether they are concept symbols or not. If a reference is 'perfunctory' in several citation contexts, but the terminology used in these contexts are agreed upon, then it is fair to assume that the cited reference is semantically related to this terminology usage, 'perfunctory' or not. Moreover, we do in fact compensate to some degree for 'perfunctory' references because we penalize citation contexts with multiple references in the concept symbol analysis explained below. Most importantly, by allowing 'perfunctory' references, we save time in an otherwise time consuming analysis.

- Multiple citation contexts within one document may be used as long as the above rules are followed.

With these rules in mind, the analysis commences by selecting a sample of citation contexts. The bibliographies of the 2001 set of 'enhanced document representations' are randomly scanned to provide at least five citation contexts for each of the 45 cited references. No upper threshold limit is chosen.

Please note that the procedure deliberately does not take into account multiple citing of a reference within a document. For example, initially five different citing papers for cited reference *X* are located. The citation contexts to reference *X* is subsequently identified in these five citing papers. However, some of the papers may have cited reference *X* several times, thus, the number of citation contexts to cited reference *X* extends beyond the initial five. The threshold value of *at least* five implies a larger probability of highly co-cited references to emerge in several 'extra' citation contexts, contrary to the low co-cited reference. Thus, the rationale behind the threshold of at least five citation context is twofold. First, we expect that the majority of cited references eventually will attain more than five citation contexts due to the sampling procedure. Secondly, we suppose that a minimum of five citation contexts may be a sufficient number for identification of concept symbols. Eventually, 88 citing documents were needed to obtain at least five citation contexts for each cited reference. The 88 citing papers from 2001 are listed in Appendix 10.

The 88 citing documents are manually scanned and the citation contexts surrounding the reference markers are identified. The citation contexts are copied into aggregated text files constructed for each individual cited reference. As stated above, the procedure is less cumbersome and swifter than for example that of Rees-Potter

(1987; 1989), since most of the citing documents are available in electronic form (i.e., pdf-format or mark-up language).

The 88 citing documents produce a sample of 580 citation contexts. That is an average of 12.9 citation contexts per cited reference with a median of 11. The highest number of contexts to one cited reference is 34 and the lowest is five. Further, only two cited references ended up having the minimum number of five citation contexts attached to it. A histogram of the distribution of citation contexts per cited reference is presented in Appendix 11.

The sample of 580 citation contexts represents the 45 cited references in the 13 created concept groups and is the basis for the present term selection methodology. The following sub-section presents the methodical step of ‘consensus passage’ identification. ‘Consensus passages’ are needed for the concept symbol analysis.

8.2.2 The notion of ‘consensus passages’

Two important assumptions pertaining to the present methodology are the consensus usage of terminology in the citation contexts of concept symbols, and the conceptual importance of this terminology. In principle, we follow the notion of Small (1978), who in relation to citation context analysis, regards concepts as residing in the mind and expressed via terms in language. Small (1978) does not restrict these terms to abstract or theoretical formulations, instead he proclaims that in connection with citations, any statement that may be taken as characterizing or describing the cited reference will do. These include, for example, experimental findings, methodologies, types of data, theoretical statements, equations etc. (Small, 1978, p. 329).

In order to apply noun phrases from the citation contexts, we need to verify whether the cited references act as concept symbols. A cited reference is a concept symbol if there is an agreement in the terminology used to describe the reference in citing papers (Small, 1978). Traditionally, concept symbol analysis implies manual investigation of all or some of the citation contexts and manual extraction of sentences or phrases. However, it is a very time consuming process to investigate all or even some of the citation contexts.

Instead, we apply a modified version of the ‘consensus passage’ procedure in conjunction with frequently occurring noun phrases in order to determine whether a cited reference acts as a concept symbol. The ‘consensus passage’ procedure originates with Small (1986), who wanted to identify citing ‘consensus passages’ for his ‘specialty narrative’ thus, the procedure is known as identification of ‘consensus passages’. The ‘consensus passage’ procedure identifies the citation context that best express the ‘consensus’ terminology in the sample of contexts to an individual cited

reference. The procedure assumes that most 'passages' for a highly cited reference express the same concept, but with slightly different terminology. It is therefore possible to select the 'passage' which is the most characteristic or typical of the group by virtue of using the most frequently encountered words to describe the idea of the cited work (Small, 1986, p. 101). As a result, the 'consensus passage' procedure is quantitative, which implies that all citation contexts are given a score that reflects their use of consensus terminology. Instead of manual analysis and interpretation of all citation contexts, a score is used to identify a 'consensus passage'. The concept expression in this passage, given a sufficiently high consensus score, is used to characterize the concept symbol of the cited reference. In the present application, the result of the 'consensus passage' procedure is compared to the lists of extracted commonly used noun phrases in order to determine the phrasal expression of the concept symbol. This is to ensure that the expression of the concept symbol will resemble the potential consensus usage of the noun phrases. This procedure alleviates the procedure of concept symbol identification.

In relation to thesaurus construction, Rees-Potter (1987; 1989) applied a different approach to citation context analysis. She identified the name of concept groups through an analysis of abstracts and reviews. No consensus or uniformity measures were applied. Consequently, all noun phrases in the citation contexts within the concept groups were initially selected. The noun phrases were subsequently evaluated by subject experts and matched with Library of Congress Subject Headings®. As a result, this approach selected several irrelevant phrases, as all phrases were initially considered candidate thesaurus terms. Further, the approach was very time consuming because the concept group analysis and phrase selection procedures were done manually.

8.2.2.1 Identification of 'consensus passages'

Noun phrases are used as the basic unit for candidate thesaurus terms, whereas single word usage in citation contexts is used for the 'consensus passage' procedure. The 'consensus passage' procedure is not interested in the exact ordering of words in phrases. What is interesting is the degree of overlap of content bearing words between the citation contexts chosen for a potential concept symbol. If it is assumed that most citation contexts express the same concept, but in slightly different words, then it is possible to select that citation context which is the most characteristic or typical of the group by virtue of using the most frequently encountered words to describe the concept

of the cited reference. Notice we are not interested in total word frequencies⁵²; conversely, we apply a unit-wise count, which implies that a word is counted only once for each citation context it occurs in⁵³. Unit-wise counting is more appropriate, as we are interested in the consensus usage of terminology *between* the citation contexts. Thus, we investigate whether a word is present or not. The application of unit-wise counting ensures normalization for the different lengths of citation contexts.

The text of the citation contexts is subjected to a lexical analysis similar to that of single term indexing presented in Chapter 3. A mild form of suffix stemming is applied to standardize plural endings. Function words are removed by use of the generic stop word list in the Bibexcel software (www.umu.se/inforsk/). In addition, a constructed dictionary of common periodontal abbreviations is applied for standardization purposes. The remaining potential content bearing words in the citation contexts are indexed.

The ‘consensus passage’ score is obtained by running a script that counts the unit-wise frequency of occurrence of all content words across all citation contexts associated with each cited reference. The maximum possible frequency of a word is equal to the number of investigated citation contexts for an individual cited reference. The content words obtain a weight equal to their unit-wise frequency of occurrence. A minimum weight threshold of two or three is imposed depending on the number of citation contexts attached to a cited reference⁵⁴. The script then scores each citation contexts for ‘representativeness’ by summing the present weights equal to or greater than the threshold value.

Two normalization functions are incorporated. As mentioned above, we penalize citation contexts with multiple references. This is done by dividing the summed weights by the number of references cited in the citation context. Finally, the adjusted score is further divided by the sum of all unit-wise word frequencies above the threshold value, i.e., ‘sum of weights’. The latter normalization produces a normalized overlap score. The normalized score equals one if a citation context contains all words above the threshold. A score of zero is obtained if a citation context contains only words that occur uniquely in that contexts. If all citation contexts have zero scores, then no common words exist and it can be said that there is no consensus on the meaning of the reference. If all scores are one, then each citation context contains the same words above the threshold, and there is complete consensus on meaning. Obviously, most cases lie in between these extremes, and contexts vary in the degree

⁵² Word frequencies correspond to term frequencies (TF) introduced in Chapter 3.

⁵³ Unit-wise count correspond to document frequency (DF) introduced in Chapter 3.

⁵⁴ Results showed that the difference in threshold weights between 1 and 2 did not have any significance at all.

to which they may be said to 'represent' or be 'typical of' all other contexts. Hence it is possible to select the most representative citation context or 'consensus passage' of the set. The example in Table 8.8 shows seven citation contexts for LOE_63. Also indicated are the number of references cited in the citation context; the adjusted score, which is normalized by the number of references cited; and the normalized score, which is the normalized overlap score based on the sum of all weights. The latter is shown further below in Table 8.9.

TABLE 8.8. Citation contexts for LOE_63 and the calculation of the normed 'consensus score'.

Citation context	Cited references	Adjusted score	Norm consensus score
"Plaque Index (28) and Gingival Index (18) were recorded."	2	19/2 = 9.5	0.23
"The measures evaluated were changes in <i>clinical inflammation</i> , ⁹ <i>plaque scores</i> , ¹⁰ <i>probing depths</i> , and <i>clinical attachment levels</i> ."	2	23/2 = 11.5	0.28
"Data recorded during each examination included age, self-reported smoking (current smoker or non-smoker), and betel nut chewing status (current user or non-user), bacterial <i>plaque</i> (<i>Plaque Index</i>), ⁵¹ and calculus accumulation (CI), ⁵² <i>gingival inflammation</i> , (GI), ⁵³ ..."	3	26/3 = 8.7	0.21
"All subjects underwent <i>clinical</i> periodontal examination including the measurement of <i>probing depth</i> (PD), <i>attachment level</i> (AL), <i>gingival index</i> (GI), ²⁰ <i>plaque index</i> (PI), ²¹ ..."	2	32/2 = 16	0.39
"Clinical parameters were measured by a single skilled examiner (A.A.) and included <i>probing depth</i> (PD), <i>bleeding on probing</i> (BOP), and <i>gingival index</i> ²¹ (GI)."	1	23/1 = 23	0.56
<i>Gingival index</i> (GI): GI was used to assess the severity of <i>gingival inflammation</i> . ²³ "	1	15/1 = 15	0.36
" <i>Plaque index</i> (PI) ²⁷ and <i>gingival index</i> (GI) ²⁸ scores ranged from 1 to 2 and from 2 to 3 for all teeth, respectively."	2	19/2 = 9.5	0.23

TABLE 8.9. List of content words that appear in two or more citation contexts to LOE_63.

Words	Frequency
gingival	6
index	6
plaque	5
clinical	3
depth	3
inflammation	3
probing	3
attachment	2
examination	2
included	2
level	2
recorded	2
scores	2
Sum of all weights	41

Table 8.9 shows the list of content words appearing in two or more citation contexts attached to LOE_63, as well as the sum of all weights for these content words.

In Table 8.8 the italicized content words are the ones that appear in more than one citation context to a cited reference. These content words and their frequencies are shown in Table 8.9. From Table 8.8, we can observe that the fifth citation context from the top has the highest norm 'consensus score' at 0.56. In order to obtain the score of 23, the sum of weights for the citation context (23) is divided with the number of reference cited in the citation context (1). The score of 23 is subsequently divided with the sum of all weights (43).

From the example above it becomes clear that citation contexts where the cited reference is mentioned exclusively, are most likely the ones that acquire the highest 'consensus scores'. This is intentional as such citation contexts most likely are the best for identification and analysis of concept symbols. Appendix 12 list 'consensus scores' for all citation contexts. The following sub-section presents the extraction of noun phrases by use of shallow parsing. Frequently occurring noun phrases are the basis for candidate thesaurus terms and are needed in combination with the 'consensus passages' for the concept symbol analysis.

8.2.3 *Extraction of noun phrases by use of shallow parsing*

A major methodical step in citation context analysis is the extraction of basic units of analysis, such as phrases or sentences, from the citation contexts. The present methodology focus on noun phrases as basic units of analysis. As described in the Chapters 2 and 3, noun phrases are the most important semantic units for indexing⁵⁵. The rationale for the use of noun phrases as index terms is that they represent more meaningful concepts than individual words. Noun phrases are believed to be content bearing units and thus good indicators of the content of a text. In Chapter 3, we list some of the motivations given by Anick and Vaithyanathan (1997) behind the use of noun phrases to describe concepts in natural language text. One of these motivations is that noun phrases are widely used across sublanguage domains to describe concepts succinctly. It is therefore evident to immediately identify and extract such phrase from the citation contexts of citing documents within periodontology, when the ultimate aim is identification of agreed upon candidate thesaurus terms.

⁵⁵ Arppe (1995) reports that between 80 to 95% of the terms listed in thesauri, or index lists of bibliographic databases, are noun phrases (including single nouns).

The exploratory methodology applies the shallow noun phrase parser *Connexor* (www.connexor.com)⁵⁶ to select phrases from the citation contexts of citing documents. Noun phrase parsing and phrase indexing is presented in Chapter 4. The application of parsing techniques for phrase extraction in connection with citation context analysis is very different from former approaches. Small (1978; 1986) and Rees-Potter (1987; 1989) manually selected noun phrases from citation contexts, whereas O'Connor (1982; 1983) used automatic statistical techniques to extract single words. The application of a noun phrase parser greatly reduces the workload of citation context analysis. Likewise, the application of a parser also ensures identification of terms that represent concepts in a more meaningful way than individual words alone.

Connexor (www.connexor.com) is a shallow syntactic parser based on a functional dependency grammar that produces part-of-speech tags, noun phrase markers (i.e., corenp), and relational dependencies between constituents. Connexor is a continuation of the older *NPtool* noun phrase parser, which was specifically developed for automatic indexing purposes (Voutilainen, 1993). Syntactical phrase identification is described in sub-section 4.2.2 in Chapter 4. In addition, Figure 4.1 illustrates the generic process of shallow parsing, which correspond to the present application. The application of the Connexor parser in the present methodology requires that the selected citation contexts, for each cited reference, are represented as text files. As stated above, the fact that most citation contexts are acquired in electronic form alleviates this process. Forty-five text files are created, one for each cited reference. The individual citation contexts are marked within the text files, in order to distinguish the parsed contexts from each other. The text files are then parsed in Connexor and the parsed results are subsequently subjected to a small script that lists the extracted noun phrases. The initial lists of shallow parsed noun phrases from the citation contexts are presented in Appendix 12, where the bibliographic description of the individual cited references (45 in all), their citation contexts, and their selected noun phrases are presented.

Table 8.10 below illustrates the process of noun phrase parsing by use of Connexor in the present methodology.

⁵⁶ Parsing was carried out via access to the Unix version of *Connexor* at the University of Tampere, Department of Information Studies in Finland.

TABLE 8.10. Example of noun phrase parsing of citation contexts by use of the Connexor parser.

Citation context to be parsed:

“Guided tissue regeneration utilizes barrier membranes to isolate the root surface from gingival epithelium and flap connective tissue. This method preferentially allows cells from the periodontal ligament and endosteum of bone to populate the defect and diseased root. This method has been applied to treatment of furcation and angular bony defects and the results from short-term evaluation of treatment appear promising. 4-8”

The parsed result from Connexor, where noun phrases are marked with brackets:

<corenp base="dc"> dc </corenp> - <corenp base="guided_tissue_regeneration"> Guided tissue regeneration </corenp> utilizes <corenp base="barrier_membrane"> barrier membranes </corenp> to isolate <corenp base="the_root_surface"> the root surface </corenp> from <corenp base="gingival_epithelium"> gingival epithelium </corenp> and <corenp base="flap_connective_tissue"> flap connective tissue </corenp> .
 <corenp base="this_method"> This method </corenp> preferentially allows <corenp base="cell"> cells </corenp> from <corenp base="the_periodontal_ligament"> the periodontal ligament </corenp> and <corenp base="endosteum"> endosteum </corenp> of <corenp base="bone"> bone </corenp> to populate <corenp base="the_defect"> the defect </corenp> and <corenp base="diseased_root"> diseased root </corenp> .
 <corenp base="this_method"> This method </corenp> has been applied to <corenp base="treatment"> treatment </corenp> of <corenp base="furcation"> furcation </corenp> and <corenp base="angular_bony_defect"> angular bony defects </corenp> and <corenp base="the_result"> the results </corenp> from <corenp base="short-term_evaluation"> short-term evaluation </corenp> of <corenp base="treatment"> treatment </corenp> appear <corenp base="promising"> promising </corenp> .

Cleaning and presentation of the parsed result for one citation context:

guided_tissue_regeneration	the_periodontal_ligament	furcation
barrier_membrane	endosteum	angular_bony_defect
the_root_surface	bone	the_result
gingival_epithelium	the_defect	short-term_evaluation
flap_connective_tissue	diseased_root	treatment
this_method	this_method	promising
cell	treatment	

Notice that the lexical analysis in Connexor extracts the singular form of noun phrases. As an example, the phrase ‘*barrier membranes*’ is transformed to its singular form by removing the *s*-ending. The lexical analysis is based on a dictionary, which enables more sophisticated suffix removals in connection with plural endings. From the parsed example presented in Table 8.10 we can observe that determiners, such as *the* and *this*, are attached to the parsed noun phrases. Likewise, simple pre-modifiers such as *greater* or *larger* also remain attached to the noun phrases after parsing. As described in Chapter 2, determiners and simple pre-modifiers are not suitable for identification of candidate thesaurus terms. As a result, a small dictionary is constructed with the most prevalent of these determiners and simple pre-modifiers. A script then matches the parsed text files with the dictionary, and removes matching determiners and pre-modifiers from the text files.

Generally, shallow parsing algorithms operate with low error rates and parsed syntactical phrases are very precise, at least in relation to adjectival phrases. We have experienced some minor problems with Connexor in relation to parsing of prepositional phrases. Constituents in a prepositional noun phrase are tied together through an *'of'* preposition as in *'association of oral disease'*. In Connexor, such a prepositional phrase is parsed into separate constituents. This implies that the phrase *'associations of oral disease'* is divided into two noun phrases, which is *'association'* and *'oral disease'*. As described in Chapter 4, problems in relation to parsing of prepositional phrases are common to natural language processing. However, most often the splitting of prepositional phrases is accepted in relation to indexing, since alternative normalization procedures are very difficult to implement and essential not worth the effort. Obviously, the division of prepositional noun phrases into less specific phrases, to some degree reduce the semantic advantages of using phrases for indexing. However, as described in Chapter 2, too specific phrases are not desirable for indexing purposes either. In addition, the majority of phrases in natural language text are adjectival noun phrases. The latter ensures that a sufficient degree of specific noun phrases eventually will be extracted.

A more serious problem in relation to prepositional phrases is illustrated with an example from Table 8.10. Consider the phrase *'treatment of furcation and angular bony defects'*. This complex prepositional phrase refers to treatment of both *'furcation defects'* and *'angular bony defects'*. The prepositional phrase is divided into three new noun phrases, *'treatment'*, *'furcation'*, and *'angular bony defects'*. Notice that the grammar in Connexor is not able to identify that *'defects'* is the head noun for both *'furcation'* and *'angular bony'*. The result is that the phrase *'angular bony defects'* retains its head noun while *'furcation'* is treated as a noun phrase of its own. As the problem is not widespread, we decided not to treat it further, since treatment is very time consuming, and in all probability not worth the effort.

Finally, as described in Chapter 4, normalization procedures are important in phrase indexing. Normalization procedures, for example, compensate for the fact that phrasal concepts can be expressed by use of different syntactic structures, possibly combined with lexical variations in word use and morphological variants (Moens, 2000). It is therefore necessary to map different phrase expressions to a single form in order to do frequency analysis. However, the main purpose of the present component is to identify agreed upon terminology in the citation contexts. This implies that the degree of normalization has to be low in order not to impose a deduced *'superficial'* consensus terminology upon the citation contexts. As a result, some small obvious adjustments are made by use of head-modifier analyses. Head-modifier analysis is

described in Chapter 4. For example, there is no doubt that the citation contexts of OFFENBACHER_96b refer to the same concept of risk for '*preterm low birth weight infants*' in relation to periodontal disease. Nevertheless, several phrasal variations of this concept are observed in the citation contexts to OFFENBACHER_96b; for example, '*preterm low birth weight baby*', '*preterm low weight birth*', '*low birth weight infant*', or '*premature low birth weight*'. Evidently, normalization of these variants does not violate the main purpose of the analysis, which is identification of consensus terminology. The problem is mainly one of different ordering of modifiers within the noun phrase. Consequently, a simple modifier analysis groups the variants together and the most prevalent phrase is chosen as the standard form.

In manual thesaurus construction, phrases are factored immediately. While factoring is essential to thesaurus construction, we do not wish to implement it at this early stage. As stated above, we are interested in identifying consensus usage of noun phrases. This literally means consensus usage, thus if a phrase appears in many different variations, we do not consider it consensus usage, except for some obvious cases such as *preterm low birth weight infants* illustrated above. Factoring can be applied at the instance when the candidate thesaurus terms are identified. Head-modifier analyses are capable of identifying some of the variant phrases that need factoring.

The purpose of noun phrase parsing in the present component of the methodology is to extract the most commonly agreed upon phrases for each cited reference. This is the most persistent expressions in the citation contexts. As reported in Chapter 4, most phrases, often 85% or more, occur only once in a corpus and can be safely eliminated. This also points to the fact that frequently occurring noun phrases generally have lower absolute frequencies compared to individual words. As discussed in Chapter 4, lower frequencies of automatically extracted phrases are partly due to the specificity of the phrases, and partly due to the deviations in the constituent words of the parsed phrases.

As mentioned above, frequently occurring noun phrases are used in conjunction with the 'consensus passage' to determine the conceptual status of a cited reference. If a cited reference acts as a concept symbol, a portfolio of noun phrases is created and attached to the concept symbol. The portfolio comprises of frequently occurring noun phrases and perhaps noun phrases from the 'consensus passage' used to denote the concept symbol. Eventually all portfolios of identified concept symbols within a concept group are compared in order to determine the candidate thesaurus terms appropriate for the particular concept group.

Based on the 'consensus passage' and the list of frequently occurring noun phrases identified for each cited reference, the following sub-section report on the concept symbol analysis.

8.2.4 Identification of concept symbols

To reiterate, the overall purpose of the concept symbol analysis is to determine the degree of consensus usage of terminology in citation contexts surrounding a cited reference. If a consensus usage is present, then it is assumed that a cited reference acts as a concept symbol. The verification of concept symbols ensures that the basic assumptions behind the term selection methodology are met. This makes it possible to identify contextual and agreed upon terminology in the citation contexts attached to a concept symbol.

The concept symbol analysis is performed manually by comparing the statistically derived 'consensus passage' for a cited reference, with the parsed list of commonly agreed upon noun phrases. The 'consensus passage' expresses the degree of terminological consensus usage of single content words in the citation contexts. No attention is paid to the actual meaning of these content words. If a cited reference acts as a concept symbol, then the chosen 'consensus passage' very likely contains phrases or sentences that express the concept. This makes the 'consensus passage' attractive for identification of concept symbols.

In order to alleviate the process of concept symbol identification in the 'consensus passages', a process which requires some domain knowledge, a list of the most commonly agreed upon noun phrases is used to substantiate the analysis. The basic notion is that a 'consensus passage' and a list of noun phrases should be sufficient for a valid conceptual analysis. However, this is not always the case. For example, small deviations in the ordering of words within a noun phrase across different contexts does not influence the 'consensus' score. It can however imply that the 'consensus passage' phrase has a lower ranking on the list of frequently occurring parsed noun phrases. This is the case, if the phrase deviates in the ordering of word constituents, compared to the frequently occurring parsed noun phrases, which have equal ordering of constituents. In such cases, phrases from the 'consensus passage' can be used to supplement the conceptual expression.

Notice, it is always terminology drawn from either the list and/or 'consensus passage', which is used to define the concept symbol. Nevertheless, we prefer the use of frequently occurring noun phrases for the description of the concept symbols, since the frequently occurring noun phrases eventually provide the basis for selection of candidate thesaurus terms.

The ‘consensus score’ is used as an immediate impression of the degree of consensus usage of terminology in the citation contexts. A grading system is used to indicate the degree of consensus. As mentioned above, the scores may be inflated due to the appearance of multiple references in the citation contexts. Therefore, the ‘consensus score’ is only used as an indication. The incorporation of the list of noun phrases into the analysis ensures that a cited reference with a low ‘consensus’ score, which at the same time has a sufficient number of agreed upon noun phrases attached to it, will be identified as a concept symbol. The same is true when the noun phrase list is clear and the ‘consensus passage’ is unclear in their expression of a concept symbol. Consequently, the determination of whether a cited reference is a concept symbol is based on a complimentary analysis between ‘consensus passage’ and the list of commonly agreed upon noun phrase.

Notice, what is defined as ‘commonly agreed upon’ in relation to parsed noun phrases usually means lower frequencies than observed for single words. As described in Chapter 4, this requires an adjustment of traditional weighting functions and threshold values. As a rule of thumb, we apply a ‘citation context frequency threshold value’ of around $\frac{1}{3}$ of the sample size. This implies that in order to be selected for the list of frequently occurring phrases, a phrase must occur at least in $\frac{1}{3}$ of the citation contexts. Please note, it is a rule of thumb. In some cases, there is a marked gap between a few frequently mentioned noun phrases and those with lesser frequencies. In such cases, it is likely that only the most frequently mentioned phrases are selected.

The concept symbol analysis of the 45 cited references is meticulously described in Appendix 9. The cited references are located in their respective concept groups. Each cited reference is analyzed to establish whether it is concept symbol. The expression of the concept symbol must be apparent from the ‘consensus passage’ or from the list of frequently occurring noun phrases according to above mentioned prescriptions. If this is not the case, then the cited reference is discarded as a concept symbol based on the present data. In addition, its portfolio of noun phrases is also excluded from further analysis.

As mentioned above, the portfolio contains the phrases used to express the concept symbol. In addition, noun phrases on the list with frequencies equal to or above the ones chosen for concept symbol description, is also incorporated into the portfolio. Finally, a low frequency noun phrase that is used in the expression of the concept symbol because it appears in the ‘consensus passage’ is likewise incorporated into the portfolio. However, in such case we do not incorporate ‘extra’ phrases equal to or

above in frequencies. All other noun phrases from the list are excluded from further analysis.

Notice, the concept symbol analysis presented here is much more elaborate than envisaged for semi-automatic thesaurus construction due to the elaborate evaluation procedures. The aim with the methodology is identification and selection of candidate thesaurus terms based on concept group creation, concept symbol identification, automatic extraction of noun phrases, and interpretation of the noun phrases by comparison with consensus citation contexts. It is assumed that this procedure essentially provides literary warrant for contextual and agreed upon candidate thesaurus terms. The value of this methodology is what we investigate in the present case study. The following sub-section presents and discusses the main results of the concept symbol analysis.

8.2.4.1 Summary of results from the concept symbol analysis

The 45 cited references are investigated in order to determine whether they act as concept symbols for the 2001 research fronts of citing papers. The main result is summarized in Table 8.11 below. From Table 8.11, we can observe that 42 concept symbols are identified, which means that three cited references are excluded from the remaining analysis. The grounds for their exclusion are given in Appendix 9. Notice that HIRSCHFELD_78 and MCFALL_82 are highly co-cited references and therefore located in the same concept group. From Table 8.7 above, we can observe that both papers also have relative high centrality scores on closeness and betweenness. The closeness score is a measure of how close a node is to everyone else, while the betweenness score is a measure of the 'broker' role of node in a network. Figure 8.5 illustrates that HIRSCHFELD_78 and MCFALL_82 are central nodes in the network. The obvious reason for the centrality of these nodes is that they have a number of co-citation links with other nodes in the proximity matrix. We would therefore expect such central nodes to act as concept symbols, although this is not the case. We can only speculate on the reason for this failure. As described in Appendix 9, HIRSCHFELD_78 and MCFALL_82 are two long-term retrospective studies that investigate a range of different aspects of periodontal therapy. It may well be that these papers are cited in several different contexts, for example, when citing authors compare their experimental results to the corresponding ones in these classic studies? Nevertheless, the most closely related cited references of HIRSCHFELD_78 and MCFALL_82 in the network all treat aspects of 'furcation involvement'. The focus of 'furcation involvement' is indicated to some degree in several citation contexts, but not

sufficiently to warrant a designation as a concept symbol attached to HIRSCHFELD_78 and MCFALL_82 respectively.

TABLE 8.11. *Result of the concept symbol analysis for the 45 highly cited references in the 2001 sample of overlapping documents in periodontology.*

Concept group no.	Cited reference	Concept symbol of
1	HAMMARSTROM_97a	The role of enamel matrix proteins in periodontal regeneration
1	HAMMARSTROM_97b	The use of enamel matrix proteins for regenerative therapy
1	HEIJL_97a	The use of enamel matrix proteins for periodontal regeneration
1	HEIJL_97b	Enamel matrix derivative treatment leads to gains in clinical attachment level
2	GOTTLow_86	Guided tissue regeneration
2	NYMAN_82a	Guided tissue regeneration
2	NYMAN_82b	Guided tissue regeneration
3	BECK_96	Association between periodontal disease and coronary heart disease
3	DESTEFANO_93	Association between periodontitis and coronary heart disease
3	JOSHUPURA_96	Association between periodontal disease and coronary heart disease
3	MATILTA_89	Association between oral health and coronary heart disease
3	OFFENBACHER_96a	Not identified
3	OFFENBACHER_96b	Periodontal disease as a risk factor for pre-term low birth weight
4	GARRETT_96	Periodontal regeneration techniques for class II furcations
4	HAMP_75	Degrees of furcation involvement
4	HIRSCHFELD_78	Not identified
4	MCFALL_82	Not identified
4	OLEARY_72	Modified O'Leary plaque record score for oral hygiene
5	GENCO_96	Risk factors for periodontal disease
5	GROSSI_94	The smoking risk factor for periodontal disease and attachment loss
5	GROSSI_95	Smoking as a risk factor for periodontal disease and alveolar bone loss
5	KORNMAN_97	Polymorphisms of the interleukin-1 genotype as markers for an increased risk of chronic periodontitis
5	PAPAPANOU_96	Smoking, age, and diabetes as risk factors for periodontal disease
6	HOLT_88	P. gingivalis initiates progression of periodontitis
6	PIKE_94	Gingipains of P. gingivalis
7	HOLT_99	P. gingivalis bacterium
7	SLOTS_99	Association between periodontal disease and A. actinomycetemcomitans and P. gingivalis
8	MASADA_90	Interleukin-1beta
8	STASHENKO_91	Tnf-alpha
9	HAFFAJEE_94	The bacteria P. gingivalis, A. actinomycetemcomitans, and B. forsythus are pathogens for periodontitis
9	LOESCHE_85	High proportions of P. gingivalis and spirochetes in patients with 'early-onset periodontitis
9	MOORE_94	Bacterial species of periodontal disease

10	SOCRANSKY_94	Checkerboard DNA-DNA hybridization technique
10	SOCRANSKY_98	The microbial complex of <i>B. forsythus</i> , <i>P. gingivalis</i> , and <i>T. denticola</i>
11	BRAGD_87	Increased level of <i>A. actinomycetemcomitans</i> in 'localized early-onset periodontitis
11	SLOTS_82	Identification of <i>A. actinomycetemcomitans</i> by use of TSBV
11	SYED_72	Reduced transport fluid
12	LOE_63	Gingival index
12	SILNESS_64	Plaque index
13	BERGSTROM_89	Cigarette smoking as a risk factor for periodontitis
13	BERGSTROM_94	Smoking as a risk factor for periodontal disease
13	HABER_93	Smoking as a risk factor for periodontal disease
13	MACHTEI_97	Smoking as a risk factor for periodontal disease
13	STOLTENBERG_93	No influence of smoking on the prevalence of bacterial species involved in periodontal disease
13	ZAMBON_96b	Difference in the prevalence of periodontal pathogens, such as <i>B. forsythus</i> , between smokers and non-smokers

The mean 'consensus score' for the remaining 42 citation contexts is 0.52 (median is 0.56). This is considered a good score, which corresponds to the exemplary result of 0.48 from the study of leukaemia viruses by Small (1986, p. 102). It has to be mentioned that a considerable number of citation contexts, including the 'consensus passages', contain multiple references. This can lower the 'consensus' scores for individual affected concept symbols.

FIGURE 8.6. *Proportion of consensus passages to citation contexts distributed among document sections.*

Figure 8.6 shows the proportion of 'consensus passages' in relation to citation contexts, distributed among the individual document sections.

The distribution pattern of the sample citation contexts among the individual document sections is expected, as most citation contexts are found in the introduction section. What is interesting is whether the distribution of the sample citation contexts

among the different document sections, resembles the distribution of the statistically derived ‘consensus passages’. Such a comparison can give an indication of where we can expect to recognize the most agreed upon concept symbols in the document structure. From Figure 8.6, we can observe that the method and review⁵⁷ sections produce relatively more ‘consensus passages’ than we would expect from the distribution of citation context. Conversely, the introduction and discussion sections produce fewer ‘consensus passages’ than we would expect. The reason for these findings is that higher ‘consensus’ scores are obtained from citation contexts extracted from method and review sections. Consequently, these sections produce the most recognizable concept symbols.

Figure 8.7, illustrates the mean ‘consensus passage’ scores distributed among the individual document sections. The total mean score is 0.52, whereas the mean ‘consensus scores’ for ‘method’ is 0.64 and ‘review’ 0.54. Figure 8.7 substantiates the findings presented in Figure 8.6, however, the mean score for ‘consensus passages’ in the review section is not markedly higher than the total score.

FIGURE 8.7. Comparison of mean ‘consensus scores’ for individual document sections. TOTAL corresponds to mean for all document sections combined.

Consequently, Figure 8.6 and 8.7 indicate that the most recognizable concept symbols are located in the method section of citing papers of the present sample. This corresponds to our experience with the concept symbol analysis. The method sections do produce very clear concept symbols that denote methods, techniques, or instruments applied in experimental studies. For example, OLEARY_72 is a concept

⁵⁷ In this context, ‘review section’ denotes the review section in review papers. Review papers in periodontology do not follow the traditional document form used in experimental and theoretical papers; instead they have a large review section.

symbol of the 'modified O'Leary plaque record score for oral hygiene', SOCRANSKY_94 is a concept symbol for the 'checkerboard DNA-DNA hybridization technique', and SYED_72 is a concept symbol for a 'reduced transport fluid'. These concept symbols are not only easy to recognize from the 'consensus passages', they also appear very clearly in the list of frequently occurring noun phrases.

Notice that the discussion sections produce fewer citation contexts than expected. Likewise, the mean 'consensus score' obtained from 'consensus passages' extracted from the discussion section is markedly lower than the total mean. In the present case study, the discussion sections typically consider very specific aspects of a cited reference, usually in relation to findings reported on in the citing paper. This implies that cited references often appear exclusively in the citation contexts appearing in the discussion sections. Such citation contexts do not always refer directly to the agreed upon concept symbol due to their specificity. But citation contexts extracted from the discussion sections are likely to acquire a relatively higher consensus scores when most other citation contexts contain multiple references and thus lower scores. This also implies that when these contexts are chosen as 'consensus passages' their 'winning' score is usually mediocre as indicated in Figure 8.7. Consequently, the 'consensus passage' from the discussion section may not be very useful for the analysis. However, as stated above, the concept symbol analysis is complimentary. Therefore, if a cited reference acts as a concept symbol and the 'consensus passage' is difficult to interpret, then the list of commonly agreed upon noun phrases can be used instead to identify the concept symbol.

An illustrative example is NYMAN_82b from Appendix 9. The 'consensus score' is low at 0.27; however, it is not difficult to determine that NYMAN_82b is in fact a concept symbol. The major reason for this determination is the complimentary use of the noun phrase list for the analysis. The list ensures that potential concept symbols are revealed if consensus terminology is present in the sample of citation contexts, but not clearly expressed in the 'consensus passage'.

In the end, what matters is whether the 'consensus passage', in combination with the noun phrase frequency lists, is able to reveal a potential concept symbol for a particular cited reference. This is clearly the case in the present case study as documented in Appendix 9. Given the citation context sampling restriction, we consider the result of 42 concept symbols out of 45 cited references as satisfactory. Moreover, as indicated in Appendix 9, while the three discarded references do not act as concept symbols, they still reflect upon the general subject matter of their parent concept group. A

larger sample of citation contexts may therefore give more indication whether these references indeed are concept symbols or not.

The citation context analysis applied in the present methodology is not very restrictive in relation to delimiting citation contexts. The aim is to produce ‘real’ citation contexts, and not some arbitrarily truncated ones. We suppose that such an approach will eventually produce the most representative noun phrases for a given concept symbol. An alternative approach would be to reduce the length of citation contexts in order to identify concept symbols more easily. This will no doubt speed up analysis, but fewer noun phrases will eventually be produced. The risk pertaining to arbitrarily defined small citation contexts is that no concept symbols can be identified. Thus, application of short-span citation contexts, probably requires a larger sample of contexts.

Finally, a portfolio of significant noun phrases is attached to each concept symbol. The following step concerns evaluation and naming of concept groups, which is based on the concept symbols and their attached portfolios of noun phrases, identified and selected by the present concept symbol analysis.

8.2.5 Evaluation and naming of concept groups

The purpose of concept groups in the present methodology is to strengthen the conceptualization of their attached concept symbols. This leads to a further contextualization of the noun phrases characterizing the group of individual concept symbols. As a result, concept symbols within a concept group, and their attached portfolio of noun phrases, are jointly investigated to ensure a more solid basis for selection of candidate thesaurus terms. The combined investigation of concept symbols within concept groups eventually makes it possible to identify primary and secondary candidate thesaurus terms. The present methodical step serves as an evaluation of the semantic coherence within the concept groups. Lack of semantic coherence means that the expected solid basis for selection of candidate thesaurus terms is disrupted. At the same time, the evaluation also serves as an extra validation of the clustering result produced by the second component.

In sub-section 8.1.1.3.1, the cophenetic correlation statistic is used to validate the document co-citation clustering result. The statistical validation focuses on the match between the derived clusters and the original proximity values. A strong correlation indicates solidly derived concept groups, which are expected to contain semantically related cited references. Consequently, in a semantically coherent concept group, all members must unequivocally refer to some common conceptual meaning. Obviously, we cannot determine the common conceptual meaning of a concept group until we

examine the conceptual meaning of its individual members. The previous step of concept symbol analysis investigates the conceptual meaning of individual members of the concept groups. As a result, the sum of meanings reflected by concept symbols in a concept group, eventually defines the common conceptual meaning for the parent concept group. Thus, concept symbols are used to name their parent concept groups.

The 2001 document co-citation clustering result created by the second component is statistically satisfactory. We therefore expect to conclude that the derived concept groups are semantically coherent, which means that their concept symbols reflect a common concept. Consequently, the present evaluation determines whether concept groups are semantically coherent. Semantic coherence is essential because it strengthens the conceptualization of concept symbols, which is needed in order to identify primary and secondary candidate thesaurus terms.

In sub-section 8.1.1.4.1, the cluster result is visualized in a PFNET. The network representation is an interpretation of the mutual structure between the derived concept groups. The network representation contextualizes the individual concept groups, thus, we expect semantically related concept groups to be located near each other or connected in some meaningful way.

Consequently, naming of concept groups entails an analysis of the different concept symbols and their attached portfolios of noun phrases, which identifies and determines the common concept reflected upon in a concept group. The network representation and its corresponding centrality measures are used as tools in the present analysis.

In order to substantiate the evaluation and naming of the concept groups, a quantitative indication of semantic coherence is also investigated. The degree of semantic coherence is indicated by use of a quantitative measure developed by Braam, Moed and van Raan (1991a, pp. 236-237) for a related purpose. The degree of semantic coherence can be measured by comparing the individual portfolios of noun phrases attached to the concept symbols in a group, with an 'aggregate portfolio' of noun phrases that represents the entire concept group. The 'aggregate portfolio' is represented as a vector that comprise all different noun phrases appearing in the individual portfolios of the group's concept symbols. The individual portfolios are likewise represented as vectors with a length that corresponds to their number of different noun phrases. A binary count is used to indicate whether a noun phrase is present or absent in the vector representations. Notice, that the 'aggregate portfolio' representing the concept group only contains presence counts, as it is a representation of all different noun phrases in the group.

The binary Ochiai similarity coefficient (18) is used to measure the similarity between the individual portfolios and the 'aggregate portfolio of the concept group. This implies that a similarity result is obtained for each portfolio attached to a concept symbol within a concept group. According to Braam, Moed and van Raan (1991a, p. 240), the average value of the individual similarities indicate the 'semantic coherence' within the concept group. In their study, Braam, Moed and van Raan (1991a, pp. 240-241) obtained average similarities in the range of 0.36 to 0.44, which they considered sufficient in order to conclude that their groups were coherent. Thus, we use their results as a baseline for evaluating the average similarities computed for the present concept groups. Finally, the frequency distribution of MeSH® descriptors, appearing in a concept group's research front, is used to evaluate the identified common concept.

The analysis that leads to evaluation and naming of the 13 concept groups are presented in Appendix 13. The appendix lists the individual concept groups, their attached concept symbols, and the most prevalent major MeSH® descriptors extracted from citing papers in the 2001 research fronts. The MeSH® descriptors are divided into primary and secondary descriptors in accordance with their frequency of occurrence. The frequency distributions are created offline with the Bibexcel software (www.umu.se/inforsk/). This is possible due to the 'enhanced document representations' created in the first component of the exploratory methodology. The 'enhanced document representations' include MeSH® descriptors as well as cited references. Appendix 14 illustrates an 'enhanced document representation'. The MeSH® descriptors are extracted from citing papers in the 2001 sample that cite members of a particular concept group. We consider citing papers that cite several members of a particular group as especially important. As a result, their descriptor frequencies are weighted corresponding to the number of references they cite in the group. For example, if a paper cites three references in a group, each of its MeSH® descriptors attains a weight of three.

Based on the 'core' phenomenon presented in Chapter 5, it is assumed that the distribution of current index terms in a research front somehow reflects upon the common subject matter of a concept group. Nevertheless, we can expect some discrepancy, especially when concept groups denote a specific concept. In such cases, it is questionable whether specific MeSH® descriptors are visible in the frequency distributions. Frequency distributions are more likely to depict broad descriptors as they generally accumulate more frequencies. Citing papers may appear in several research fronts, which is reflected in the assignment of different index terms to these papers. Consequently, index terms that reflect upon other research fronts is likely to

appear in a distribution for a specific concept group if sufficient numbers are encountered. However, in most cases we expect the 'core' phenomenon to prevail, thus, indicating the semantic meaning of a concept group.

Compared to their constituent members, the individual concept groups most likely denote broader concepts. The evaluation is performed as an interpretation of the derived conceptual meaning of the concept group, and the one indicated by the primary and secondary MeSH® descriptors. The result verifies the semantic coherence of the concept group. Finally, the concept group is given a conceptual name that reflects the common conceptual meaning shared by the members of the group. The following subsection presents and discusses the main findings.

8.2.5.1 Summary of results from the evaluation and naming of concept groups

The analysis that leads to the evaluation and naming of the 13 concept groups are thoroughly described in Appendix 13.

TABLE 8.12. *Name of concept groups and quantitative indication of semantic coherence. The semantic coherence value is the average similarity scores of individual portfolios of concept symbols within a concept group compared to the 'aggregate portfolio' of noun phrases for the entire group.*

Concept group no.	Name of concept groups	Semantic coherence
1	Enamel matrix proteins	0.620
2	Guided tissue regeneration	0.902
3	Complications of periodontal disease	0.650
4	Furcation involvement	0.548
5	Risk factors for periodontal disease	0.585
6	Periodontitis progression, p. gingivalis	0.762
7	Periodontal pathogen, p. gingivalis	0.787
8	Cytokines	0.707
9	Periodontal pathogens	0.623
10	Classification of bacteria	0.774
11	Periodontal pathogen, A. actinomycetemcomitans	0.638
12	Periodontal index	0.707
13	Risk factor of smoking	0.592

The main result is presented in Table 8.12 above, together with the quantitatively derived measure of semantic coherence. The result of the evaluation and naming of the concept groups is very clear. All groups are semantically coherent. This implies that the concept symbols within the concept groups unequivocally refer to some common concept. The unambiguous result of the evaluation and naming of the concept groups is further substantiated by the high average similarity scores for each group. The scores indicate quantitatively the degree of semantic coherence for the individual concept groups. The ‘coherence scores’ in the present analysis, ranging from 0.585 to 0.902, are far above the results obtained by Braam, Moed and van Raan (1991a, pp. 240-241). Thus, compared to their results and subsequent conclusions, the present quantitative coherence results are very convincing.

In general, the result substantiates the statistical validation of the clustering results. Hence, we have 13 concept groups that more or less specifically reflect upon important concepts of periodontology. Notice, it is important concepts of periodontology as expressed through the co-citation behaviour and consensus usage of terminology among citing authors in 2001; whereas in manual thesaurus construction subject experts often designate important concepts. Together the concept groups contain 42 concept symbols and a similar number of noun phrase portfolios. The portfolio of noun phrases is the central object for term selection reported on in the following subsection.

Consult Figure 8.8 on the following page. It is the PFNET presented in Figure 8.4, but, instead of cited references, the present representation visualizes the names of the concept groups and the potential semantic relations between them. From the analysis in Appendix 13, we can overall conclude that the network representation is semantically meaningful. The connectivity and placement of nodes indicate semantic relationships useful for thesaurus construction purposes. Generally, it seems that ‘arteries’ most often indicate what corresponds to associative relationships, and ‘veins’ often indicate what corresponds to hierarchical relationships. However, let us stress that this is only an indication based on the present data and applied techniques. Conceptual and terminological relationships and meaning result in certain statistical patterns, but as described by Soergel (1974, p. 449), statistics should not take precedence over human judgement, though it can provide the basis for some useful decisions. This is elaborated on in section 8.3 concerning term associations.

For example, from the analysis in Appendix 13, we can establish that concept groups 1, 2, and 4 have an associative relationship between them. Concept groups 1, 2, and 4 are located on the lower part of the vertical ‘artery’, Most strongly between

guided tissue regeneration and *enamel matrix proteins*. Both concepts are related to periodontal regeneration. Further, we can deduce that several 'veins' appending from the 'artery' indicate what corresponds to hierarchical relationships. This is the case with the concept group 5 and 13. It is also the case within concept group 3. Concept group 3 concerns *complications of periodontal disease*; this general conception is deduced from the concept symbols located on the main 'artery'.

FIGURE 8.8. PFNET representation with concept group names indicated. CG is an abbreviation for concept group. The dotted circle to the left indicates the semantic similarity of the concept groups in this part of the network. The dotted line to the right indicates that concept group 10 is semantically related to the concept groups in the left side of the network.

The small appendix to the main 'artery' within concept group 3 concerns the specific complication of *low-birth weight of infants* in connection with periodontal disease.

Notice, that the link between concept group 5 and concept group 9 most likely represent an associative relationship. However, a hierarchical relationship is indicated between concept groups 7, 9, and 11, with group 9 as the broader concept; this semantic relation is indicated in Figure 8.8 with a dotted circle.

One minor problem was detected in the process of concept group validation, which is related to the relationship between concept groups 7, 9, 10, and 11. The analysis revealed that concept groups 7, 9, 10, and 11 are semantically coherent and related to each other see Appendix 13. The analysis shows that concept group 9 focuses on general aspects of *periodontal pathogens*, whereas concept groups 7, 10, and 11 focuses on more specific aspects of *periodontal pathogens*. It therefore seems reasonable to suggest that a hierarchical relationship exist between these concept groups, at least groups 7, 9, and 11.

Indications from the dendrogram in Appendix 6 show that a slightly lower threshold value would result in the merger of concept groups 9 and 10 in a nested cluster. Due to the strong complete-link clustering criterion, we can expect members of the two concept groups to be semantically related. Thus, it would seem natural if concept group 10, in some way is connected to concept group 9. However, the PFNET representation gives another impression.

Concept group 10 concerns *isolation* and *classification of bacteria* (periodontal pathogens) and is indicated by the dotted circle on the right side of the network. The group appears as an appendix to concept group 13, which concerns *smoking as a risk factor for periodontal disease*. As described in the analysis of concept group 13 in Appendix 6, the relationship between concept group 10 and 13 is most likely an associative one, due to the relation between bacterial species and the risk factor of smoking. The analysis of concept group 10 in Appendix 6 thoroughly explains why the PFNET weights the relation to concept group 13 higher, instead of the more apparent one to concept group 9. Briefly summarized, the PFNET focuses on strong links between nodes, while the clustering criterion focuses on complete internal linkage between members in order to produce solid clique-like clusters. Remember this is a very strong cluster criterion. Concept group 10 is a solid and semantically coherent cluster. If concept group 10 was to merge with another cluster, then the members of both clusters should have ‘complete linkage’, otherwise the merger fails due to the cluster criterion. Concept group 10 has complete linkage with all members of concept group 9. This is the reason why they will cluster at a lower threshold value. However, the fact that members of concept group 10 has complete-linkage with members of concept group 9, does not necessarily imply that these are the strongest individual links to nodes (cluster members) external of concept group 10. In fact, the

strongest individual link to a node external to concept group 10 is the one to ZAMBON_96b is concept group 13.

The Pathfinder algorithm reduces the number of links and retains the strongest ones. Thus, in the present case the strongest link and not the average best 'fit' determines the location of concept group 10 in the network. Notice that if we alter the threshold value and merge concept group 9 and 10, the PFNET representation would not change as the link to concept group 13 is still the strongest.

The incident is neither the result of an arbitrarily clustering, as the cluster is semantically coherent, nor is it an arbitrarily network representation, as the algorithm is set to reveal the strongest links. In most cases, the links considered for the complete-linkage clustering criterion corresponds to the strongest links emphasized by the Pathfinder algorithm. The incident therefore illustrates the advantage of visualizing the cluster result in a network representation. Without the network representation, we would most likely not have detected the associative relationship between the *classification of bacteria* and *smoking as a risk factor* for periodontal disease. Bear in mind, that the more apparent semantic relationship between *classification of bacteria* and *periodontal pathogens* is revealed by the present third component of exploratory methodology. We can confirm the relationship by consulting the dendrogram, but the actual relation is revealed due to the citation context analysis.

The latter example shows the usefulness of visualization of cluster results. Such visualization serves to contextualize the identified concept groups, and eventually their constituent concept symbols and noun phrases.

The methodical steps of the third component are essentially a filtering procedure, which ensures that candidate thesaurus terms are contextual, agreed upon, and therefore assumed to be important. Initially, noun phrases in citation contexts attached to a specific concept symbol are filtered due to their degree of common usage. Consensus terminology is assumed contextual, because it is identified in short-spanned text windows that reflect upon a common concept.

Concept symbols are heterogeneous. Some reflect specific or broad aspects of a common concept, while several others are identical. It is therefore very convenient to group them, and then utilize their communality as the basis for selection of candidate thesaurus terms. The extra conceptualization enforced by the concept symbol grouping makes it possible to filter the noun phrases in the portfolios once more. This filtering procedure enables differentiation between primary and secondary candidate thesaurus terms. Primary terms most likely reflect upon the common concept of the

group, whereas secondary terms more likely reflect upon specific aspects of the common concept expressed by individual concept symbols. It is assumed that the highest proportion of important candidate thesaurus terms is located among the primary terms.

The following sub-sections illustrate the term selection methodology, and eventually validates the selected candidate thesaurus terms for the 2001 sample of citing papers in periodontology. The result of the validation makes it possible to answer the first and most central research question.

8.2.6 *Selection of candidate thesaurus terms*

As stated in the introduction to the present section, the purpose of the third component of the exploratory methodology is term selection. Based on the concept groups established for the 2001 sample in the second component, the third component investigates the ability of citation context analysis and noun phrase parsing for selection of candidate thesaurus terms. In the previous sub-sections, we have introduced a filtering procedure, which ensures that potential candidate thesaurus terms are contextual, agreed upon, and therefore assumed to be important. Besides concept group creation, the method implies identification of concept symbols, extraction of noun phrases, and evaluation and naming of concept groups. Consequently, the second and third components explore the first and most central research question, whether the applied bibliometric methods are able to detect candidate thesaurus terms in a specialty area within the life sciences, where knowledge claims are primarily mediated in journal papers.

The final step in the process of term selection is the actual selection of candidate thesaurus terms. Term selection is based on the results from the previous methodical steps presented in the second and third components. We illustrate the procedure of candidate thesaurus term selection by examining the case of concept group 1 concerning *enamel matrix proteins*. The selected candidate thesaurus terms for all concept groups are listed in Appendix 15.

Initially, 13 concept groups are created by use of document co-citation clustering in the second component of the methodology. One of these concept groups contains four cited references:

- a. HAMMARSTROM_97a;
- b. HAMMARSTROM_97b;
- c. HEIJL_97a;
- d. HEIJL_97b.

The concept symbol analysis identifies concept symbols for all four cited references. The concept symbols are (the alphabetic order corresponds to the above-given alphabetic order of references):

- a. The role of enamel matrix proteins in periodontal regeneration;
- b. The use of enamel matrix proteins for regenerative therapy;
- c. The use of enamel matrix proteins for periodontal regeneration;
- d. Enamel matrix derivative treatment leads to gains in clinical attachment level.

In connection with the concept symbol analysis a portfolio of noun phrases is created for each concept symbol. This corresponds to the first filtering procedure. The portfolio comprises of frequently occurring noun phrases, and perhaps phrases from the 'consensus passage', which are used to characterize the concept symbol. The frequency analysis is based on 'document frequencies', which means unit-wise counting of noun phrases in citation contexts. The procedure normalizes frequency counting by only considering the presence or absence of a phrase in a citation context. 'High frequency' corresponds to consensus usage of terminology.

The portfolios for the four concept symbols in concept group 1 are comprised of the following noun phrases (the alphabetic order corresponds to the above-given alphabetic order of concept symbols):

- a. enamel_matrix_protein; enamel_matrix_derivative; periodontal_regeneration;
- b. enamel_matrix_protein; regenerative_therapy;
- c. enamel_matrix_protein; periodontal_regeneration
- d. enamel_matrix_derivative; clinical_attachment_level; gain; treatment.

The concept group evaluation ensures that there is a semantic coherence between the concept symbols. A name for the concept group is chosen, which reflects the concept common to the members; in the present case *enamel matrix protein*. This is an extra conceptualization of the four concept symbols. Instead of treating the concept symbols and their noun phrases separately, we use the communality of the group as the basis for term selection. This ensures more contextualization, as term selection is carried out on the group level and not individually for each concept symbol.

Subsequently, a second filtering procedure is applied, which separates the noun phrases from the individual portfolios in two categories, primary and secondary candidate thesaurus terms. The filtering procedure is based on a simple ‘document frequency analysis’. However, this time ‘document’ corresponds to the individual portfolios. The number of citation contexts attached to the individual concept symbol influences the frequency count of noun phrases in the portfolios. Hence, the application of document frequency at the portfolio level normalizes for the variances in frequencies of noun phrases between the portfolios.

Consequently, primary candidate thesaurus terms have higher frequencies across the portfolios than secondary candidate thesaurus terms. This implies that primary terms most likely appear in several portfolios within a group. Primary terms, therefore, most likely reflect upon the common concept of the group. Conversely, secondary terms are likely to reflect upon specific aspects of the common concept within a group. The selected primary and secondary candidate thesaurus terms for concept group 1 are:

Primary:	3	enamel_matrix_protein
	2	periodontal_regeneration
	2	enamel_matrix_derivative
Secondary:	1	treatment
	1	regenerative_therapy
	1	clinical_attachment_level
	1	gain [discarded]

A manual inspection is applied to discard what appear to be non-significant index terms, in the present case: *gain*.

This is the selection procedure of candidate thesaurus terms proposed and explored in present methodology. As mentioned above, all selected candidate thesaurus terms are presented in Appendix 15. Notice, that concept group 4 has equal frequencies of 1 among the selected noun phrases. We have decided to choose three of the phrases as primary terms as they essentially reflect upon the same aspect of ‘furcation involvement’. Otherwise, no problems were encountered with the selection procedure.

The following sub-section presents the results of the validation of the selected candidate thesaurus terms.

8.2.6.1 Validation of selected candidate thesaurus terms

The selected candidate thesaurus terms are evaluated by use of a quantitative validation procedure. The procedure consists in matching the selected primary and

secondary terms with possible corresponding terms in two authority lists, which is the MeSH® vocabulary and *Glossary of Periodontal Terms* (2001). As described in Chapter 2, MeSH® is the standard authority in relation to index terms concerning periodontology. Equally, *Glossary of Periodontal Terms* (2001) is a standard definition of periodontal terms produced by the American Academy of Periodontology. Overall, the validation procedure resembles that of Rees-Potter (1987). She matched extracted concepts in economics and sociology with those found in Library of Congress Subject Headings (Rees-Potter, 1987, p. 39).

The assumption behind the validation procedure is that the second and third components of the exploratory methodology are able to identify important candidate thesaurus terms among the selected primary and secondary terms. More importantly, it is assumed that the selection procedures are able to put forward a significant proportion of important candidate thesaurus terms among the selected primary terms. If the assumption is true, primary candidate thesaurus terms should acquire high overlap scores⁵⁸. Further, if the assumption is true, then a significant difference in the proportion of overlap between primary and secondary candidate thesaurus terms must be identified.

The matching procedure is separated, so that the first analysis concerns the MeSH® descriptors, and the second analysis concerns the periodontal terms from the glossary. The two analyses differ. MeSH® descriptors are index terms, whereas the periodontal terms from the glossary are domain specific definitions. The matching of selected candidate thesaurus terms with MeSH® descriptors is considered the more important of the two analyses, because the result indicates the ability of the methodology to produce important index terms. The matching of selected terms with terms in the glossary is used as an indication of the ability of the methodology to select important domain terms. Eventually, the results from the two analyses are compared, in order to investigate whether the results support each other.

Consequently, the analyses require different coding of data in order to perform the matching procedures. The results of the two analyses are presented individually. First, we consider the matching of primary and secondary candidate thesaurus terms with corresponding MeSH® descriptors.

8.2.6.1.1 Representativeness of selected candidate thesaurus terms in the MeSH® vocabulary

The matching procedure is carried out online by use of the MeSH® browser (www.nlm.nih.gov/mesh/MBrowser.html). The aim of the first analysis is to

⁵⁸ The overlap score is the proportion of the total overlap.

investigate the ability of the second and third components of the methodology to identify a significant proportion of important index terms among the selected primary terms. This implies matching primary and secondary terms separately with descriptors from the MeSH® vocabulary. A coding is needed, which can determine the level of match acquired. Four different categories are decided upon; these are:

- *Exact match*, as the wording indicates, this corresponds to a perfect match between the selected term and a MeSH® descriptor.
- *Entry term match* refers to cases where the search results in an automatic match with a corresponding relevant MeSH® descriptor, as a result of an entry term match. While this is not a perfect match, it is considered a match for the present analysis. Specific indexing policies have decided which terms should be entry terms in the vocabulary. Nevertheless, an entry term is an important term. As the selected candidate thesaurus terms have literary warrant, they are assumed to be contextual and agreed upon, we consider the match with entry terms as an exact match of important terms.
- *Partial match* refers to cases where the selected term does not automatically generate an exact or entry term match. However, the selected term does obtain a match of some kind, for example, a broader or narrower term. While there is match of some kind, it is considered a ‘no match’ for the present analysis.
- *No match* refers to all other instances.

A total of 60 selected candidate thesaurus terms are investigated, 21 primary and 39 secondary terms. The complete results are presented in Appendix 16 and summarized in Figures 8.9 and 8.10 below.

FIGURE 8.9. *Overlap score between selected primary candidate thesaurus terms and MeSH® descriptors.*

Figure 8.9 shows the overlap scores for the primary candidate thesaurus terms. The scores indicate the representativeness of selected primary candidate thesaurus terms in the MeSH® vocabulary. The result presented in Figure 8.9 is very promising. Overall, the categories deemed relevant, *exact match* and *entry term match* obtain a total score of 0.76. If we consider the individual distribution among the relevant matches, we can observe that *exact matches* noticeably have the highest score of 0.43. Considering that *partial matches* are matches of a kind, this leaves the score of *no matches* at 0.10. In absolute numbers, 0.10 covers two primary candidate thesaurus terms, 'bacterial species' and 'periodontal regeneration'. 'Bacterial species' are not represented in the glossary either, whereas 'periodontal regeneration' has a partial match in the glossary.

The high relevant overlap score for the selected primary candidate thesaurus terms suggests that the methodology produce a considerable number of important index terms in this category.

Figure 8.10 below shows the overlap scores for the secondary candidate thesaurus terms. The scores indicate the representativeness of selected secondary candidate thesaurus terms in the MeSH® vocabulary.

FIGURE 8.10. *Overlap score between selected secondary candidate thesaurus terms and MeSH® descriptors.*

The result presented in Figure 8.10 supports the result for primary terms presented in Figure 8.9. The result in Figure 8.10 indicates a lower degree of relevant overlap in the MeSH® vocabulary for the selected secondary candidate thesaurus terms. We can observe that the total overlap score for relevant matches drops noticeably from 0.76 for primary terms to 0.38 for secondary terms. The distribution between relevant matches,

partial matches, and no matches is almost equal. For the present analysis, the results imply that the proportion of important index terms among the selected secondary candidate thesaurus terms are lower compared to the primary terms. As a result, the secondary candidate thesaurus terms have a lower degree of representativeness in the MeSH® vocabulary. Consequently, the results support the findings for primary terms presented in Figure 8.9 and the assumptions presented in sub-section 8.2.6.1.

As in the present case, where data consist of frequencies (matches) in discrete categories, the non-parametric chi-square statistic can be used to determine the significance of differences between the two groups (Siegel & Castellan, 1988, p. 111). Thus, the proportion of overlap attained for the primary and secondary terms are subjected to the chi-square statistic in order to test whether the difference in the proportions is significant. The matching categories are merged into two groups. Exact and entry term matches are placed in a group coded *relevant matches*, whereas partial and no matches are placed in a group coded *non-relevant matches*. This produces a 2×2 contingency table, where columns indicate primary and secondary candidate thesaurus terms, and rows indicate relevant or non-relevant matching groups. The result of the chi-square test is presented in Table 8.13.

TABLE 8.13. *Chi-square test of proportional difference in relevant index terms between primary and secondary candidate thesaurus terms – based on match with MeSH® descriptors.*

Chi-square test of proportional difference in relevant index terms between primary and secondary candidate thesaurus terms (MeSH®)	
Chi-square (observed value)	11.28
Chi-square (critical value)	6.35
Degrees of Freedom	1
One-tailed p-value	0.001
Alpha	0.01

At the level of significance $\alpha = 0.001$ the decision is to reject the null hypothesis of no difference in the proportion of relevant index terms between primary and secondary candidate thesaurus terms. In other words, the proportional difference in relevant index terms between the primary and secondary candidate thesaurus terms is significant. Consequently, a significantly higher proportion of selected primary candidate thesaurus terms are represented in the MeSH® vocabulary.

The validation results, the high relevant overlap score and the significance test, confirm the assumption. The special selection procedures of the second and third components of the exploratory methodology produce important index terms among the selected primary and secondary terms. More importantly, the selection procedures enable the selection of a significantly higher number of important index terms among the selected primary terms. As a result, at least for the specialty area of periodontology, we can expect the methodology to identify a significant number of important and highly relevant candidate thesaurus terms among the selected primary terms.

8.2.6.1.2 Representativeness of selected candidate thesaurus terms in the Glossary of Periodontal Terms (2001)

The aim of the second analysis is to investigate the proportion of important domain terms among the selected primary and secondary candidate thesaurus terms. Thus, the result of the analysis indicates the ability of the exploratory methodology to select important domain terms. The validation procedure investigates the overlap between selected candidate thesaurus terms and their potential entries in the *Glossary of Periodontal Terms* (2001). Entries in the glossary are considered important domain terms for the present analysis. It is assumed that the special selection procedures of the second and third components of the methodology produce a significantly higher proportion of important domain terms among the selected primary candidate thesaurus terms. If the assumption is true, primary candidate thesaurus terms should acquire an acceptable exact match score. But more importantly, if the assumption is true, a significant difference in the proportion of overlap between primary and secondary candidate thesaurus terms must be identified. In order to determine the level of match, three different categories are decided upon:

- *Exact match* refers to a perfect match between the selected candidate thesaurus terms and the entry in the glossary.
- *Partial match* refers to a match where part of the candidate thesaurus term is covered by entries in the glossary. Partial matches are considered neutral for the present analysis.
- *No match* refers to all other instances.

The 60 selected candidate thesaurus terms, 21 primary and 39 secondary terms, are matched manually against potential entries in the glossary. The results are summarized in Figures 8.11 and 8.12 below.

FIGURE 8.11. *Overlap score between selected primary candidate thesaurus terms and important periodontal terms from *Glossary of Periodontal Terms* (2001).*

We can observe from the result presented in Figure 8.11 that primary candidate thesaurus terms have an *exact match* score of 0.48. The scores indicate the representativeness of selected primary candidate thesaurus terms in the *Glossary of Periodontal Terms* (2001). The exact match score is almost identical to the exact match score obtained in relation to the MeSH® descriptors presented in Figure 8.9. The present result is promising and acceptable as it indicates that half of the selected primary candidate thesaurus terms are important domain terms. In the present analysis, partial matches are treated as neutrals. However, in order to determine whether the methodology produces more important domain terms among the selected primary terms, we need to compare the results to the corresponding result for the selected primary and secondary terms. The overlap scores for the selected secondary candidate thesaurus terms are presented in Figure 8.12.

FIGURE 8.12. *Overlap score between selected secondary candidate thesaurus terms and important periodontal terms from *Glossary of Periodontal Terms* (2001).*

The result presented in Figure 8.12 is supportive of the result for primary terms presented in Figure 8.11. The result in Figure 8.12 indicates a lower degree of overlap in the *Glossary of Periodontal Terms* (2001) for the selected secondary candidate thesaurus terms. We can observe that the overlap score for relevant matches drops noticeably from 0.48 for primary terms to 0.29 for secondary terms. The score of 0.29 is slightly higher than the corresponding match score between secondary terms and their potential MeSH® descriptors of 0.23 presented in Figure 8.10.

For the present analysis, the results imply that the proportion of important domain terms among the selected secondary terms are less compared to the primary terms, because overall the secondary terms have a lower degree of representativeness in the glossary. Thus, the result supports the findings for primary terms presented in Figure 8.11.

Similar to the first analysis, the proportion of overlap attained for the primary and secondary terms are subjected to the chi-square statistic in order to test whether the difference in the proportions is significant. Partial matches are discarded, which produces a 2×2 contingency table, where columns indicate primary and secondary candidate thesaurus terms, and rows indicate exact or no match groups. Thus, the chi-square test only considers the proportion of exact matches to no matches. The result of the chi-square test is presented in Table 8.14 below.

TABLE 8.14. *Chi-square test of proportional difference in important domain terms between primary and secondary candidate thesaurus terms – based on match with Glossary terms.*

Chi-square test of proportional difference in important domain terms between primary and secondary candidate thesaurus terms (Glossary)	
Chi-square (observed value)	4.656
Chi-square (critical value)	3.841
DF	1
One-tailed p-value	0.031
Alpha	0.05

At the level of significance $\alpha = 0.050$ the decision is to reject the null hypothesis of no difference in the proportion of important domain terms between selected primary and secondary candidate thesaurus terms. In other words, the difference of important domain terms between the selected primary and secondary candidate thesaurus terms is

significant. Consequently, a significantly higher proportion of primary candidate thesaurus terms are important domain terms as represented in the *Glossary of Periodontal Terms* (2001).

The validation results, the exact overlap score and the significance test, confirm the assumption that the significant difference in the proportion of important domain terms between primary and secondary candidate thesaurus terms is due to the special selection procedures of the second and third components of the exploratory methodology.

The results of both validation analyses are very clear. The proportion of important index terms is significantly higher among the primary candidate thesaurus terms compared to the secondary ones. For every four selected primary candidate thesaurus terms, three have a relevant representation in the MeSH® vocabulary. Likewise, the proportion of important domain terms is significantly higher among the primary candidate thesaurus terms compared to the secondary ones.

Notice that some of the selected secondary candidate thesaurus terms are also important. Over one-third of the secondary terms have corresponding relevant matches in the MeSH® vocabulary. Remember that secondary terms are also chosen due to their contextual and consensus usage in relation to a concept symbol. Therefore, some of the selected secondary candidate thesaurus terms are also likely to be relevant for indexing considerations. Accordingly, secondary terms should also enter the pool of *candidate* thesaurus terms considered for thesaurus construction. Consequently, the results from the present case study of periodontology indicate that we can expect a significant number of primary terms in the pool of potential index terms to be important candidate thesaurus terms. In addition, we can also expect some of the selected secondary terms to be relevant candidate thesaurus terms. This leads to the conclusion that the bibliometric based term selection method introduced in the present exploratory methodology is able to select a significant number of important primary candidate thesaurus. The terms are current, contextual, and agreed upon, thus the second and third components of methodology are suitable for semi-automatic thesaurus construction.

8.2.7 *Summary of results*

The second and third components of the exploratory methodology investigates the most central research question in the present dissertation, whether the applied bibliometric methods are able to identify candidate thesaurus terms within a specialty area in the life sciences, where knowledge claims are primarily mediated through

journal papers. The exploration of the components in the present case study of periodontology clearly demonstrates that the applied bibliometric methods are able to select important candidate thesaurus terms within this particular specialty area. We believe that the special selection procedures inherent in the methodical steps of the components ensure that a significant number of the selected primary candidate thesaurus terms turn out to be important index terms. The conclusion therefore is that the applied bibliometric methods are suitable for selection of candidate thesaurus terms.

The key to this result is the focus on cited references as the primary unit of analysis. The citation context of a cited reference, acting as a concept symbol, is very interesting in relation to semi-automatic thesaurus construction. The terminology in the text window is most likely highly contextual as it reflects special characteristics in relation to the concept symbol. Likewise, consensus terminology is most likely found in the text window, as the cited reference acts as a concept symbol. These features are very important in relation to selection of index terms. Consequently, the proposed methodology establishes literary warrant for a significant number of important contextual and agreed upon primary candidate thesaurus terms. This is an excellent basis for further manual thesaurus construction.

As stated above, we believe that the special selection procedures inherent in the components eventually warrant the promising result for the selected primary terms. One crucial selection procedure is the document co-citation clustering that creates the concept groups. The correlation analysis confirmed that the clustering result is satisfactory, but the real confirmation is the verification that all concept groups are semantically coherent. There is no doubt that the unequivocal conceptual meanings among cluster members across all concept groups in the present study, is due to the implementation of the rigid complete-link clustering criterion, which produces clique-like clusters very suitable for conceptual purposes.

The results produced by the citation context analysis extend the findings of Rees-Potter (1987; 1989). We demonstrate the usefulness of citation context analysis for thesaurus construction purposes in a specialty area within the life sciences, dominated by journal papers. We can conclude that it is possible to identify concept symbols in citing journal papers in periodontology, and to extract noun phrases from their citation contexts. These procedures are useful for thesaurus construction.

The investigation also demonstrates the usefulness of noun phrase parsing in citation context analysis. Noun phrase parsing alleviates the time consuming process of phrase identification. In the present approach to thesaurus construction, noun

phrase parsing is especially suitable, as it extracts agreed upon terminology from which candidate thesaurus terms are eventually selected.

There are some deficiencies too. Document co-citation clustering as applied in the present dissertation only focuses on a small number of the core intellectual base references for the 2001 citing papers. This produces few concept groups. Nevertheless, the groups reflect the most visible research areas within periodontology, as reflected in the bibliographies of the 2001 citing papers. Several studies have suggested possible recall enhancing procedures in co-citation studies (e.g., Braam, Moed & van Raan, 1991a). However, in relation to the present application they do not seem very profitable. It gives no meaning to enhance the set of references by use of co-word analysis as suggested by Braam, Moed and van Raan (1991a), since the present methodology is strictly centered on concept symbols, i.e., highly cited references; this fact narrows the available references for analysis from the outset. More importantly, in order to create concept groups by use of the 'indirect approach' co-citations are needed. What seems more profitable for future research, is the application of bibliographic coupling in combination with identification of concept symbols in order to broaden the basis for concept group creation and term selection.

While not deemed a serious problem in the present analysis, it is noticeable that noun phrase parsing creates a considerable number of phrase variants. We believe that the methodology may profit by more elaborate phrase normalization procedures. One such procedure is to combine the parsed phrases with the result of a collocation analysis. The collocation analysis statistically joins the most prevalent two-word phrases. As the phrases are constructed statistically, all sorts of phrase combinations are produced. However, if the phrase combinations are matched against the list of parsed noun phrases, then it is possible to identify the most prevalent two-word noun phrases. This procedure can be used as normalization and factoring of the noun phrases, which eventually reduces the number of phrase variants created through parsing. The question is whether it is worth the effort.

This leads to the last consideration, whether the second and third components of the methodology concerning term selection are 'less resource demanding'. As presented here, the methodology may seem prolonged and time consuming, but please note we are to evaluate and discuss the individual proposed steps. The aim with the methodology is that the two components should be easy to apply. The only real time consuming step is the concept symbol analysis. However, the purpose of automatically identification of a 'consensus passage' and the extraction of commonly agreed upon noun phrases are specifically applied to reduce the otherwise time consuming process of concept symbol analysis.

The present section has investigated the ability of the second and third components to select candidate thesaurus terms. The following section investigates aspects in relation to term associations. Notice, that the second component of vocabulary organization and the third component of term selection already have created the basis for this analysis in the form of conceptual networks centered on the concept groups.

8.3 Fourth component: Conceptual network (term association)

The fourth component of the exploratory methodology investigates the ability of co-word analysis to disclose equivalence, hierarchical, and associative relationships in a conceptual network, based on the results of the second and third components. Thus, the fourth component explores the second research question of the dissertation. Accordingly, the fourth component is comparable to term association in traditional automatic thesaurus construction approaches, discussed in the Chapters 4 and 6.

In the proposed methodology, a conceptual network is centered on a concept group. The basic units of the network are noun phrases in the portfolios attached to concept symbols in a group. The rationale for the establishment of conceptual networks in the present methodology is previously presented in sub-section 6.1.4. Nevertheless, we briefly restate the most important preconditions for the present approach; preconditions that separates it from most other approaches.

As introduced in Chapter 3 and 4, traditional automatic thesaurus construction approaches consist of three successive steps, 1) term selection, 2) term association, and 3) vocabulary organization. The most important outcome of the successive ordering of the construction steps, is that term association and vocabulary organisation most often depend on direct first order co-occurrence analysis, albeit a notable exception is Latent Semantic Indexing. The Chapters 4 and 6 discuss some of the limitations related to the use of direct first order co-occurrence analysis. The most perceptible is the difficulty of associating synonymous and near-synonymous terms, because such terms rarely co-occur frequently in the same documents.

The methodology alters the successive ordering of the construction steps. As an alternative, it commences thesaurus construction with 'vocabulary organization'. The methodology explores an indirect approach to thesaurus construction, one based on cited references instead of term frequency distributions. As illustrated in Figure 6.3 and demonstrated in section 8.1, the indirect approach implies creation of concept

groups by use of document co-citation clustering, and not first order term co-occurrence analysis based on term distributions.

As demonstrated in section 8.2, a semantically coherent concept group contains semantically related concept symbols and their attached portfolios of noun phrases. It is important to emphasize that the selected set of noun phrases assigned to a concept group are not clustered due to their frequent co-occurrence in text windows. Conversely, the noun phrases appear together because they frequently share the same textual context, that is, the context surrounding a specific concept symbol in the citing papers. The citation contexts of concept symbols are most likely highly contextual in relation to subject matter. It is therefore assumed that most of the selected noun phrases are important and in some way semantically related. The conjecture is that some of these noun phrases co-occur directly with each other in the short-span citation contexts. While other noun phrases do not co-occur at all with each other in these contexts. Instead, these phrases frequently co-occur with the reference marker of the concept symbol. Hence, such noun phrases share a common context, the context of the concept symbol's reference marker. Thereby, all noun phrases become related to each other, either directly by occurring in the same context, or indirectly through their common co-occurrence with the concept symbol.

Theoretically, this opens up the possibility of bringing semantically related terminology into the analysis, terminology that rarely co-occur together but often share common textual contexts. This important basis is the major difference between traditional term association analyses, and the term association analyses applied in the present methodology.

The second and third components of the exploratory methodology have already created a contextual setting of semantically related noun phrases, where some are identified as primary and secondary candidate thesaurus terms, ready to be investigated for thesaural relations by use of co-word analysis.

The special thesaurus construction procedures of the second and third components imply that we can pursue what corresponds to first and second order co-occurrence analysis from the same basic $n \times m$ matrix. As it appears from the description above, the matrix is not an ordinary co-occurrence matrix. Besides traditional direct co-occurrences, the matrix also depicts noun phrases that frequently share textual context with each other. This implies that the matrix may contain semantically related noun phrases with low or no direct co-occurrence counts. The noun phrases are semantically related because they have similar 'association profiles'. An 'association profile' consists of the phrases that a particular noun phrase co-occur with. Thus, similar 'association profiles' means that two noun phrases share a number of co-

occurrence 'partners'. This enables second order co-occurrence analysis, where 'association profiles' of noun phrases are investigated instead of direct co-occurrence of phrases.

As a result, the basic $n \times m$ matrix for the co-word analysis consists of noun phrases (n) and citation contexts (m). Note we apply citation contexts as units for the present co-word analysis. The co-word analysis is based on unit-wise counting of noun phrases in citation contexts. In the third component, citation contexts are used as units in the frequency analysis that determined the agreed upon noun phrases of a concept symbol. Subsequently, portfolios are used as units in the frequency analysis that determined whether selected noun phrases were primary or secondary candidate thesaurus term. The aggregate of portfolios are applied in the latter frequency analysis in order to normalize for cases where there is a marked difference in the number of citation contexts representing the individual concept symbols in a group. If no normalization procedure is invoked, the latter condition will favour some noun phrases, which thereby attain too high frequency counts.

In the present co-word analysis, we return to the use of citation contexts as units for frequency counting because it ensures the most reliable basis for the co-occurrence counts needed for the analysis. Instead, the co-word analysis applies proximity measures with more sophisticated normalization functions than envisaged for the determination of primary and secondary candidate thesaurus terms. As a result, marked differences in the magnitude of noun phrase frequencies are compensated for in the present analysis by use of proximity measures.

The purpose of the fourth component therefore is to create a conceptual network from the selected terminology of one concept group by use of co-word analyses and network visualization techniques. A conceptual network does not directly correspond to a thesaurus entry. Even so, it is assumed that the conceptual network can reveal relations between contextual noun phrases within a specialty area, relations that may be useful for manual thesaurus construction. It is important to emphasize that we do not expect to create a 'perfect' network with exact relations, nor do we claim that this is possible. The assumption investigated relates to the fact that existing relationships between noun phrases based on their meaning result in certain statistical patterns of occurrence and co-occurrence of the phrases in citation contexts (Soergel, 1974, p. 449). We should therefore be able to conclude from the observed statistical patterns the meaning of some of these conceptual relationships. Automatic statistical methods, such as co-word analysis can assist in, but not replace, the intellectual effort needed for thesaurus construction. The advantage of such methods is that they can depict salient relations in a matrix otherwise difficult to deduce.

An illustrative case of one concept group is analysed, in order to explore the ability of co-word analysis to create a conceptual network and disclose potential conceptual relationships within it. Concept group 1 is chosen for the task, as it appears to reflect some of the characteristics pertaining to the special composition of the matrix presented above. Obviously, the specific findings depend on the frequency patterns within the particular concept group. Nevertheless, there is no reason why general findings pertaining to the investigated concept group, should not be valid for other groups, given similar patterns of data in their matrices.

The basis of the conceptual network is the primary and secondary candidate thesaurus terms selected in the third component. Most likely, one or several of these are important candidate thesaurus terms, which make them the focal units in the network. Moreover, we incorporate some less frequently occurring noun phrases from the portfolios in the concept group in order to broaden the analysis.

The analysis consists in exploring different proximity measures by use of the co-word analysis and then visualizing the resulting conceptual networks as PFNETs. In parallel, we analyze the results in order to verify whether potential thesaural relations become apparent. Validation is carried out by investigating the meaning of the potential relationships in textbooks of periodontology, the glossary of periodontology, as well as the MeSH® vocabulary (Jenkins & Allan, 2001; Glossary of Periodontal Terms, 2001; Newman, Takei & Carranza, 2002; www.nlm.nih.gov/mesh/MBrowser.html).

The following sub-section presents the noun phrases of concept group 1 used in the co-word analysis.

8.3.1 Presentation of the noun phrases selected for the co-word analysis

Because of the second and third components, three primary and three secondary candidate thesaurus terms are selected from concept group 1 for the co-word analysis. The selected candidate thesaurus terms are:

Primary:	3	enamel_matrix_protein
	2	periodontal_regeneration
	2	enamel_matrix_derivative
Secondary:	1	treatment
	1	regenerative_therapy
	1	clinical_attachment_level

The three primary candidate thesaurus terms are characterized by a high unit frequency (portfolio) within the concept group. For example, '*enamel matrix protein*' appears as the most persistent noun phrase in three out of four portfolios within concept group 1 (see Appendix 15). Obviously, the 13 concept groups created for the 2001 sample are not exhaustive in their coverage of periodontology. However, if we compare the three primary terms from concept group 1 with the appearance of other selected candidate thesaurus terms across the 13 concept groups, it becomes clear that they discriminate the present concept group from the others. The three primary terms are the focal units in the following creation of a conceptual network for concept group 1. The most obvious choice as common concept for the group, and indeed the one chosen in section 8.2.6, is *enamel matrix protein*. This is the most agreed upon noun phrase in the citation contexts attached to concept symbols within the present concept group.

Besides the six selected candidate thesaurus terms, another five noun phrases from the portfolios attached to concept group 1 are incorporated to broaden the co-word analysis. The selection of the additional noun phrases is based on a frequency analysis. As in the case of term selection, some common phrases such as *human*, *animal*, and *study* are excluded because they have low indexing value and low discriminatory power across concept groups.

In addition, a small head noun analysis is applied to identify phrase variants among the selected candidate thesaurus terms and noun phrases. At this stage, we find it more appropriate to apply phrase normalization since consensus usage has been determined. Thus, variant phrases can be normalized to the 'consensus' form. The procedure is a form of subsumption, as complex noun phrase are factored to their most common form. Notice it is a head noun analysis, which is applied only in the most evident cases to frequently occurring phrases, or phrases that become frequent as a result of subsumption. The result of the subsumption procedure is presented in Table 8.15.

TABLE 8.15. *Subsumption of head noun variants. The parsed phrases are indicated to the left and their frequency counts in parenthesis. The result of the subsumption procedure is the factored phrases to the right.*

Parsed phrase (frequency in parenthesis)	Factored phrase
acellular_cementum (3) layer_cementum (2) cementum (4)	➔ cementum
porcine_derived_enamel_matrix_derivative (1)	➔ enamel_matrix_derivative
porcine_derived_enamel_matrix_protein (3)	➔ enamel_matrix_protein

As a result the frequency counts of noun phrases in the citation contexts need to be adjusted after implementation of the subsumption procedure. Hence, the basis for the conceptual network is the following noun phrases presented in Table 8.16:

TABLE 8.16. *The noun phrases selected for the conceptual network of concept group 1. Primary candidate thesaurus terms are indicated in grey.*

Candidate thesaurus terms	Frequency
enamel_matrix_protein	25
periodontal_regeneration	7
enamel_matrix_derivative	26
treatment	14
regenerative_therapy	5
clinical_attachment_level	16
Additional noun phrases	
cementum	9
connective_tissue_attachment	7
Formation	7
intrabony_defect	7
periodontal_defect	8

The following description of *enamel matrix protein* serves as background knowledge and is used to evaluate the results of the co-word analysis. The selected noun phrases or their ‘derivatives’ are indicated in *italic*.

The periodontium is the tissues that invest and support the teeth, which include the gingiva, periodontal ligament, *cementum* and alveolar bone (*Glossary of Periodontal Terms*, 2001, p. 40). Specifically, the *cementum* is the tissue that provides for the *attachment* of periodontal ligament to the alveolar bone. More broadly, the periodontal ligament is the *connective tissue* that surrounds and attaches roots of teeth to the alveolar bone. *Enamel matrix proteins* induce the *formation* of *cementum* and are believed to favour *periodontal regeneration* in *treatment* of *intrabony defects*⁵⁹.

⁵⁹ *Intrabony* has the following entry in *Glossary of Periodontal Terms* (2001, p. 28), “[w]ithin a bone; ... See: Periodontal Bony Defects”. The corresponding entry for *Periodontal Bony Defects* is “[a]lterations in the morphological features of the bone ... defects may be subcategorized as follows ...” (*Glossary of Periodontal Terms* (2001, p. 38). One of the sub-categorizations is *Furcation Invasion*, which is further classified as class I, II, and III furcation invasions (*Glossary of Periodontal Terms* (2001, p. 38). Another sub-categorization of Periodontal Bony Defect is *Intrabony Defect*: “A

One *enamel matrix protein derivative* obtained from the development of porcine teeth is marketed under the name Emdogain® (Newman, Takei & Carranza, 2002, p. 812). The *enamel matrix derivative* Emdogain® is introduced in HEIJL_97b.

From this description of *enamel matrix protein*, we can observe that all noun phrases selected for the co-word analysis, more or less specifically, reflect upon the use of *enamel matrix proteins* for periodontal regeneration. It is very important to notice that we expect all of the selected noun phrases to be in some kind of semantic relationship with each other due to the selection procedures of the second and third components.

Moreover, we can also notice that '*periodontal regeneration*' and '*periodontal therapy*' are most likely equivalent terms in the present context as they both denote the use of *enamel matrix proteins* in connection with periodontal regeneration. Depending on the purpose of thesaurus construction, '*periodontal regeneration*' and '*periodontal therapy*' can be perceived as 'narrower terms' to '*treatment*', as '*periodontal regeneration*' is a special form of periodontal treatment (Jenkins & Allan, 2001, pp. 95-100). In fact, as illustrated in Newman, Takei and Carranza (2002, p. 812), the use of '*enamel matrix proteins*' as an adjunct in '*regenerative therapy*' means that the concept of '*enamel matrix protein*' typically is classified alongside '*guided tissue regeneration*' as a narrower term to periodontal regeneration. Likewise, the phrase '*periodontal defects*' corresponds to the entry 'periodontal bony defects' in the *Glossary of Periodontal Terms* (2001, p. 38). Further, '*inrabony defect*' is a 'narrower term' of 'periodontal bony defects', according to the *Glossary of Periodontal Terms* (2001, p. 39). Thus, '*inrabony defects*' is most likely a narrower term to '*periodontal defects*'. Evidently, '*formation*' reflects aspects in relation to '*cementum*' and '*cementum*' clearly has relations to '*clinical attachment level*' and '*connective tissue attachment*'.

Finally, there is the most important question regarding the difference between '*enamel matrix proteins*' and '*enamel matrix derivatives*'. The name of the latter indicates that it is a derivative of '*enamel matrix proteins*'. This corresponds with the definition presented above. Consequently, '*enamel matrix protein*' and *enamel matrix derivative*' appear to be in a hierarchical class inclusion relationship, where the narrower concept of '*enamel matrix derivative*' has all the characteristics of the broader concept of '*enamel matrix protein*'.

periodontal defect surrounded by two or three bony walls or a combination of these" (*Glossary of Periodontal Terms* (2001, p. 39).

Before we present the results of the co-word analysis, we need to discuss the important aspect of how to interpret statistically derived co-occurrence relationships. This is discussed in the following sub-section.

8.3.2 *Interpreting co-occurrence relationships*

As introduced in Chapter 2, a thesaurus contains three types of relationships, equivalence (synonymy and near-synonymy), hierarchical, and associative relationships. Soergel (1974, p. 113) distinguishes between two main types of relationships used in psychology to study term-term associations, these are *definitional relationships* and relationships of ‘*contextual contiguity*’ (for further elaboration see Tague, 1969). The relationships cut across the traditional typology used in thesaurus construction as definitional relationships for example include synonymy, near-synonymy, class inclusion, and topic inclusion. Similar, relationships of ‘contextual contiguity’ comprise part-whole and other hierarchical relationships, as well as empirically defined associative relationships. This implies that equivalence relationships correspond to definitional relationships, whereas associative relationships correspond to relationships of ‘contextual contiguity’. Different types of hierarchical relationships are divided among the two.

According to Soergel (1974, p. 112), the distinction between definitional relationships and relationships of ‘contextual contiguity’ is essential for the interpretation of term co-occurrence statistics. As a result, Soergel (1974, p. 452) elaborates on the distinctions and presents several possibilities that can be used to interpret high statistical associations between two single terms. We have revised and simplified the different categories, so that they apply to linguistically derived noun phrases. Consequently, high association between two phrases *A* and *B* can mean any of the following.

1. *A* and *B* are in a *definitional relationship*:
 - a. *A* and *B* are synonymous;
 - b. *A* and *B* are near-synonymous⁶⁰;
 - c. *A* and *B* designate concepts that are in a class inclusion or topic inclusion relationship.
2. *A* and *B* designate concepts that are in a *relationship* of ‘*contextual contiguity*’:

⁶⁰ Soergel (1974, pp. 110-112) defines near-synonyms (quasi-synonyms) with the following example: “terms *A* and *B* are quasi-synonyms [equivalent] because their meanings overlap widely. *A* may fully cover *B*, *B* being only slightly narrower; in this case *A* will be used as the preferred term. Thus, near-synonyms designate equivalent concepts or concepts that are similar in meaning”.

- a. Other hierarchical relationships such as part-whole relationship;
- b. Empirically connected associative relationships.

Obviously, the main problem is how to fit in the statistically derived associations with the above-mentioned categories. A very important aspect in this relation, is the size of the text windows used as units of analysis (in the present case the units of analysis correspond to citation contexts of multiple sentences). Equivalence relations are very unlikely in short span text windows such as individual sentences and abstracts⁶¹ (Soergel, 1974). Equivalence relations are more likely in larger text windows such as paragraphs and full text documents because authors tend to use synonyms and near-synonyms in order to achieve variety in their writing (see Chapter 4 for a discussion of lexical cohesion). However, the use of equivalent terminology in larger text windows is often not reflected in the results of a direct first order co-occurrence analysis. Typically, authors use several synonymous terms to describe a concept. This means that none of these terms acquire sufficiently high co-occurrence counts to reveal their equivalence relation in a direct co-occurrence analysis. As a result, equivalence relations are more easily detected by use of second order co-occurrence analysis as described above. As far as possible, this is also the case for class and topic inclusion relationships (Soergel, 1974, p. 453). This leads to the conclusion that *definitional relationships* are best detected by studying their second order associations.

Conversely, relationships of 'contextual contiguity' between phrases are typically found by studying their direct first order associations in multiple sentence text windows or beyond. In first order co-occurrence analysis, most of these relations appear as associative. It should be mentioned that all relations in principle can be found in first order analyses, though, as described above, some are more likely than others.

An empirical indication of a hierarchical relationship is very difficult to establish. According to Soergel (1974, p. 454), a hierarchical relationship can be surmised especially if there is a 'one-sided overlap'. This implies that if phrase *B* is considerably less frequent than phrase *A* and if almost all units containing phrase *B* also contain phrase *A*, then we may suspect that *B* designates a concept that is narrower than the concept designated by *A*. However, *B* may just as well be a rarely used synonym of *A*. Notice that it is quite likely that a specific noun phrase is used more often than a more general noun phrase. For this to be valid, sufficient co-occurrence

⁶¹ According to Soergel (1974, p. 452), 'abstracters' are very unlikely to use synonymous terms within an abstract.

counts must be obtained and the resulting association score between two phrases must be sufficiently high.

To summarize, we have some indications of how to interpret strong empirically derived associations between noun phrases by first and second order co-occurrence analysis. First order co-occurrence analysis most likely produces relationships of ‘contextual contiguity’ between noun phrases. In addition, we can expect most of these relations to be associative. Second order co-occurrence analysis most likely produces definitional relationships between noun phrases. Further, a strong one-sided overlap may indicate a hierarchical relationship, though the latter is very difficult to determine statistically by use of co-occurrence analysis.

The following sub-section presents the creation of the conceptual network by use of co-word analysis, which make up the fourth component of the proposed methodology.

8.3.3 *Creation of a conceptual network by use of co-word analysis: Special focus on primary candidate thesaurus terms*

To reiterate, two special features of the basis $n \times m$ matrix used for the present co-word analysis are important:

- All noun phrases in the matrix are by definition a priori related to the common concept of *enamel matrix proteins*. As a result, the fourth component tries to structure these conceptually related noun phrases by use of co-word analysis. Obviously, the result depends on the obtained frequency counts. Hence, only the strongest associations are reliable in relation to revealing conceptual relations. Notice, the fact that all noun phrases are semantically related does not necessarily mean that they have direct associations to the focal primary candidate thesaurus terms. Some noun phrases are empirically related to the primary terms through one or several mediator phrases.
- As described above, first and second order associations can be derived from the specifically composed matrix simply by altering the proximity measures used for the co-word analysis. Calculation of second order associations requires a proximity measure that focuses on vector comparison instead of individual conditional probabilities. Vector comparison is needed because ‘association profiles’ for two objects are investigated.

The focal units in the co-word analysis are the three primary candidate thesaurus terms. We are therefore primarily interested in revealing their mutual relationships. Nevertheless, other strong associations are obviously also relevant and interesting for the conceptual network. Thus, the remaining relationships are treated in sub-section 8.3.4.

Initially we want to determine quantitatively the strongest direct first order co-occurrences between the noun phrases. For this purpose, we must define a measure of association. Chapter 4 thoroughly discusses different types of proximity measures including association measures. As we have discussed in Chapter 4 and demonstrated in section 8.1, several of the different association measures are mutually monotonic. A commonly used association measure in co-word analysis is the so-called *equivalence index*, which resembles the Jaccard association measure. Both measures are conditional probabilities and they are monotonic to each other (Turner et al, 1988). The *equivalence index* is defined as:

$$E_{i,j} = \frac{(C_{i,j})^2}{(C_i \cdot C_j)} \quad (22)$$

Where, $C_{i,j}$ is the number of co-occurrences of noun phrase i and j in the citation contexts; C_i and C_j is the individual number of occurrences of noun phrase i and j in the citation contexts. $E_{i,j}$ measures the probability of noun phrase i appearing simultaneously in a set of citation contexts indexed by noun phrase j and, inversely, the probability of noun phrase j occurring if noun phrase i appears, given the respective collection frequencies (i.e., citation context frequency) of the two noun phrases. For this reason, $E_{i,j}$ is called 'a coefficient of mutual inclusion' (Turner et al, 1988). Hence, we apply the equivalence index in order to derive a proximity matrix of first order associations between the noun phrases. The equivalence proximity matrix for concept group 1 is presented Table 8.17.

TABLE 8.17. *Equivalence proximity matrix of first order association between the selected noun phrases in concept group 1. Only the lower-left half of the equivalence proximity matrix is shown since it is symmetric.*

		cementum	clinical_attachment_level	connective_tissue_attachment	enamel_matrix_derivative	enamel_matrix_protein	formation	intrabony_defect	periodontal_defect	periodontal_regeneration	regenerative_therapy	treatment
9	cementum											
16	clinical_attachment_level	0.00										
7	connective_tissue_attachment	0.40	0.00									
26	enamel_matrix_derivative	0.00	0.24	0.01								
25	enamel_matrix_protein	0.36	0.02	0.21	0.01							
7	formation	0.40	0.00	1.00	0.01	0.21						
7	intrabony_defect	0.00	0.32	0.00	0.27	0.00	0.00					
8	periodontal_defect	0.06	0.00	0.07	0.02	0.18	0.07	0.00				
7	periodontal_regeneration	0.00	0.00	0.00	0.05	0.14	0.00	0.00	0.00			
5	regenerative_therapy	0.02	0.00	0.00	0.01	0.20	0.00	0.00	0.40	0.00		
14	treatment	0.07	0.16	0.04	0.27	0.03	0.04	0.50	0.04	0.00	0.00	

The three primary candidate thesaurus terms and their mutual association scores are indicated in grey. The strongest associations are marked in **bold**. The numbers to the left of the noun phrase expressions are the occurrence counts of the phrases from the citation contexts (collection frequency). In traditional automatic thesaurus construction, a relationship between two terms is defined whenever the association is above a certain threshold. These relationships are then used indiscriminately in retrieval. However, as stated above, an association measure can mean different things and it is therefore advisable first to make sure, how the relation between two phrases should be interpreted.

The result from the third component concerning concept group 1, shows that the selected noun phrases for the present co-word analysis are semantically related as they all reflect upon the concept of *enamel matrix proteins*. This is confirmed by the description of *enamel matrix proteins* presented in sub-section 8.3.1.

In order to construe the conceptual network, we focus on the primary candidate thesaurus terms and the strongest associations between noun phrases in the matrix. This procedure tries to verify the most salient empirical relations in the network.

We commence by verifying the relation between ‘*enamel matrix protein*’ and ‘*enamel matrix derivative*’. From the matrix in Table 8.17, we can observe that the association score between ‘*enamel matrix protein*’ and ‘*enamel matrix derivative*’ is very low (0.01). For the moment, we do not consider the obvious relation between the two

phrases indicated in the phrasal expression of 'enamel matrix derivative'. We solely concentrate the analysis on the empirical patterns. We can observe that the two noun phrases have considerably higher occurrence counts than the other selected phrases. In addition, the magnitude of the counts is almost equal. It is therefore very interesting that their mutual first order association score is very low at 0.01. The score indicates that the two noun phrases rarely co-occur in the same citation context. From the co-occurrence counts, we can confirm that the phrases only appear together on two occasions. From a distributional point of view, we would expect the two phrases to co-occur more regularly given their high occurrence counts. But this is clearly not the case, we can therefore conclude that 'enamel matrix protein' and 'enamel matrix derivative' are not in a relationship of 'contextual contiguity'. Conversely, the two noun phrases must instead be in a definitional relationship.

This is an example of the 'indirect approach' introduced in the Chapters 5 and 6. The proposed exploratory methodology is able to identify and represent semantically related terminology in a matrix, despite the fact that the terminology hardly ever co-occurs directly in citation contexts. In contrast, the terminology is related indirectly because it shares a common mediator in textual contexts. In the present case, the noun phrases appear in citation contexts where the concept symbol (mediator) signifies 'enamel matrix proteins'.

A second order co-occurrence analysis based on the 'association profiles' of 'enamel matrix protein' and 'enamel matrix derivative' confirms that the two phrases are in a definitional relationship. Consider Table 8.18, which illustrate the 'association profiles' of the two noun phrases, as well as the computed second order association score.

TABLE 8.18. Second order co-occurrence analysis between binary 'association profiles' of enamel matrix protein and enamel matrix derivative. The two noun phrases are represented as binary vectors; their similarity score is indicated to the right.

	cementum	clinical_attachment_level	connective_tissue_attachment	enamel_matrix_derivative	enamel_matrix_protein	formation	intraabony_defect	periodontal_defect	periodontal_regeneration	regenerative_therapy	treatment	Score
enamel_matrix_derivative	1	1	1	0	1	1	1	1	1	1	1	0.84
enamel_matrix_protein	1	1	1	1	0	1	0	1	1	1	1	

An ‘association profile’ corresponds to a vector representation of noun phrases. Thus, the list of noun phrases associated in the first order with a particular noun phrase comprises the ‘association profile’ of that particular noun phrase. The degree of similarity between two ‘association profiles’ corresponds to the second order association between two noun phrases. In a second order co-occurrence analysis two decisions are important: 1) to determine whether binary or non-binary counts should represent the components in a vector; and 2) how to treat positive matches, mismatches, and negative matches⁶² in relation to vector length. Chapter 4 discusses the problems concerning these choices. In the present case, the Ochiai index (18) is applied. The Ochiai index is the binary application of the cosine angular association measure. As presented in Chapter 3 and 4, the cosine coefficient downgrades zero counts when it measures the degree of similarity between two vectors. The purpose of the present second order co-occurrence analysis is to establish to which extend two noun phrases share a ‘common context’. In this respect, a ‘common context’ means the common share of noun phrases. For this purpose, the binary presence-absence information is sufficient.

The second order association score of 0.84 presented in Table 8.18 is very high, thus, as concluded above, the two noun phrases are in a definitional relationship to each other. In sub-section 8.3.1, we claimed that the relation appear to be a class-inclusion relationship, where the narrower concept of *enamel matrix derivatives* has all the characteristics of the broader concept of *enamel matrix proteins*. Their mutual co-occurrence relations in the citation contexts, revealed by the present co-word analysis, could indicate that the noun phrases are applied as near-synonyms by citing authors. Their meaning overlap widely as ‘*enamel matrix proteins*’ fully cover ‘*enamel matrix derivatives*’, the latter being only slightly narrower. Nevertheless, the empirical evidence restricts our definition of the relation to one that belongs to the category of definitional relationships. Further, conclusions must be carried out by manual interpretation.

The relationship between the above-mentioned two candidate thesaurus terms and ‘*periodontal regeneration*’ is more blurred. From the equivalence matrix presented in Table 8.17, we can observe that ‘*periodontal regeneration*’ has a low association score with ‘*enamel matrix derivative*’ and a low to medium score with ‘*enamel matrix protein*’. As a result, the empirical observations immediately point in the direction of definitional relationships with ‘*enamel matrix derivative*’ and most likely a

⁶² *Positive matches* are the components whose value is 1 in each of two vectors; *mismatches* are components whose value is 1 in only one of the two vectors; *negative matches* are components whose value is 0 in each of the two vectors.

relationship of 'contextual contiguity' with '*enamel matrix protein*'. But, the second order co-occurrence analysis for '*periodontal regeneration*' and '*enamel matrix derivative*' results in a low score of 0.22. As a result, the first and second order association scores between '*periodontal regeneration*' and '*enamel matrix derivative*' indicate that the two noun phrases infrequently co-occur in citation contexts, and that they do not share many phrases between them in the contexts they appear in.

The first order association score between '*periodontal regeneration*' and '*enamel matrix protein*' could indicate a relationship of 'contextual contiguity'. In the case of first order associations, it is possible to investigate empirically whether the relationship might be hierarchical or associative. As stated above, it is very difficult to indicate a hierarchical relationship empirically, but Soergel (1974) suggests that a strong one-sided overlap may indicate such a relationship. In co-word analysis, the *inclusion index* is traditionally used for detection of hierarchies within subject areas (He, 1999). The inclusion index reads as follows:

$$I_{i,j} = \frac{C_{i,j}}{\min(C_i, C_j)} \quad (23)$$

Where C_{ij} is the number of citation contexts in which the noun phrases appear together; C_i and C_j are the individual occurrence frequencies of noun phrase i and j in the set of citation contexts; and $\min(C_i, C_j)$ is the minimum of the two frequencies C_i and C_j .

The inclusion index can be interpreted as a conditional probability. The notion is that when $C_i > C_j$, then noun phrase i is more general than noun phrase j and includes j , as a result the inclusion index measures the probability of finding i in a citation context given that j appears in it. Obviously, the hierarchic interpretation is problematic. As discussed above, in cases of empirically detected hierarchical relationships, it might as well be the narrower term that obtain the highest proximity score. Hence, it is more appropriate to indicate whether a hierarchic relationship is present or not, and then let the manual analysis decide the broader-narrow relation between the two terms.

The inclusion index score between '*periodontal regeneration*' and '*enamel matrix protein*' is 0.71. The result indicates that the two phrases could be in a hierarchical relationship to each other. Nevertheless, we have to be cautious with the empirically indicated relations of '*periodontal regeneration*' as the phrase only appears in seven citation contexts. The scattering of the seven citation contexts among different concept symbols, eventually secures that '*periodontal regeneration*' becomes a primary candidate thesaurus term. But, when constructing a conceptual network based on

frequency counts then the low number of units is problematic. Nevertheless, while the relation to ‘*enamel matrix derivative*’ is uncertain from the present data, we can conclude that ‘*periodontal regeneration*’ and ‘*enamel matrix protein*’ are in a relationship of ‘contextual contiguity’, perhaps a hierarchical relationships.

Based on the empirical results, we can conclude that ‘*enamel matrix derivative*’ and ‘*enamel matrix protein*’ are in definitional relationships to each other; ‘*enamel matrix protein*’ and ‘*periodontal regeneration*’ is in a relationship of ‘contextual contiguity’, probably a hierarchical one; whereas it is not possible to empirically detect a relationship between ‘*enamel matrix derivative*’ and ‘*periodontal regeneration*’.

8.3.4 *Detection of the strongest empirically thesaural relationships for the remaining noun phrases in the conceptual network*

The following co-word analysis procedure is pursued in order to verify the empirically derived important first and second order associations. Notice we apply some very strict thresholds due to rather low frequency counts.

1. All first order association scores in the equivalence matrix are investigated in order to verify relationships of ‘contextual contiguity’. A threshold of 0.14 is set to define the strongest associations among the noun phrases. The threshold value is chosen so that all phrases have at least one ‘strong’ relation with another phrase. The strongest association scores are the most reliable.
2. The verified relationships of ‘contextual contiguity’ are further examined to find out if they participate in a hierarchical relationship. The *inclusion index* (Appendix 17) is used to identify one-sided overlaps as prescribed by Soergel (1974). A threshold of 0.80 is set to define a one-sided overlap.
3. Weak first order relationships, the ones close to zero, are examined to verify whether they indicate definitional relationships. The *Ochiai index* (Appendix 17) is used to measure the similarity between the ‘association profiles’ in order to derive their second order associations. A threshold value of 0.80 is set to define a definitional relationship.
4. Association scores between 0.14 and close to zero are not considered because they are not empirically reliable. If anything, these relations probably indicate ‘contextual contiguity’.

The results of the co-word analyses are presented in Table 8.19. The lower-left half of the matrix in Table 8.19 is the equivalence matrix, which is the basis for the analyses,

and the upper-right half of the matrix indicates the empirically identified relationships between the noun phrases.

TABLE 8.19. *Empirically derived first and second order relationships by use of co-word analyses. The matrix is similar to the equivalence matrix in TABLE 8.17., however the upper-right half in the present symmetric matrix indicates the strongest empirically derived thesaural relationships by a code.*

	cementum	clinical_attachment_level	connective_tissue_attachment	enamel_matrix_derivative	enamel_matrix_protein	formation	intrabony_defect	periodontal_defect	periodontal_regeneration	regenerative_therapy	Treatment
cementum	-		CC		CC(H)	CC					
clinical_attachment_level	0.00	-		CC			CC				CC
connective_tissue_attachment	0.40	0.00	-		CC	CC				DR	
enamel_matrix_derivative	0.00	0.24	0.01	-	DR		CC				CC
enamel_matrix_protein	0.36	0.02	0.21	0.01	-	CC		CC	CC	CC(H)	
formation	0.40	0.00	1.00	0.01	0.21	-				DR	
intrabony_defect	0.00	0.32	0.00	0.27	0.00	0.00	-				
periodontal_defect	0.06	0.00	0.07	0.02	0.18	0.07	0.00	-		CC	
periodontal_regeneration	0.00	0.00	0.00	0.05	0.14	0.00	0.00	0.00	-		
regenerative_therapy	0.02	0.00	0.00	0.01	0.20	0.00	0.00	0.40	0.00	-	
treatment	0.07	0.16	0.04	0.27	0.03	0.04	0.50	0.04	0.00	0.00	-

CC = Relationship of 'contextual relationship'
CC(H) = One-sided overlap with a threshold value over 0.80, indicating a hierarchical relation
DR = Definitional relationship with a threshold value over 0.80

As stated above, in principle all the noun phrases investigated in the present co-word analyses are related to each other. The co-word analyses try to verify empirically the meaning of the strongest of these relationships. Thus, given the described co-word analysis procedure above, the results presented in Table 8.19 are the empirically strongest relationship obtainable with the present method.

Not surprisingly, the detected strong relationships of 'contextual contiguity' seem to be valid. A validation of the strong relations based on the description of enamel matrix protein presented in sub-section 8.3.1, verifies that most of the relations appear to be associative relationships. In order to investigate whether hierarchical relations can be detected among the first order associations, one-sided overlaps are searched for. Two one-side overlaps are identified with threshold values over 0.80. The relation between 'enamel matrix protein' and 'regenerative therapy' seems a valid case of a hierarchical relationship, whereas, 'cementum' and 'enamel matrix protein' is more

questionable. Several of the strong relationships of ‘contextual contiguity’ have one-sided overlaps with threshold values just below the 0.80 threshold. For example, as discussed in sub-section 8.3.3, the relation between ‘*periodontal regeneration*’ and ‘*enamel matrix protein*’ indicates a hierarchical relationship. This is indeed a valid interpretation of the relationship, yet the small number of frequencies entail that a strict threshold needs to be enforced in order to minimize arbitrarily defined relations.

The question concerning threshold values is also decisive in relation to detection of definitional relationships. Three definitional relationships with threshold values above 0.80 are identified. The relationships are however of mixed validity. The already analyzed relationship between ‘*enamel matrix derivative*’ and ‘*enamel matrix protein*’ is convincing, and is also the strongest second order association among the set of noun phrases. Conversely, it is questionable whether the relation between ‘*connective tissue attachment*’ and ‘*regenerative therapy*’, as well as the relation between ‘*formation*’ and ‘*regenerative therapy*’, are definitional relationships.

Similar to the case of relationships of ‘contextual contiguity’, several valid definitional relationships can be detected just below the 0.80 threshold. For example, both ‘*periodontal regeneration*’ and ‘*regenerative therapy*’, as well as ‘*regenerative therapy*’ and ‘*treatment*’, seem to be convincing definitional relationships. The former points to an equivalence relationship between ‘*periodontal regeneration*’ and ‘*regenerative therapy*’, while the latter points to a hierarchical relationship between ‘*regenerative therapy*’ and ‘*treatment*’. Nevertheless, several unconvincing definitional relationships are also detected just below the 0.80 threshold value, for example, ‘*clinical attachment level*’ and ‘*periodontal regeneration*’, as well as ‘*cementum*’ and ‘*enamel matrix derivative*’.

A major reason for the mixed results in relation to detecting definitional relationships pertains to the calculation of second order associations. It is a question of whether to apply binary or non-binary counts, and how to handle positive matches, mismatches, and negative matches, in relation to the number of components matched. As stated above, we think that binary counts are the most appropriate for the information we seek in the present co-word analysis. As the information we seek is the degree of shared noun phrases between two particular noun phrases. To obtain this information, the presence or absence of a noun phrase in any citation context with the noun phrases investigated is sufficient. Nevertheless, we investigate 11 noun phrases, thus, the ‘association profile’ for the noun phrase investigated is limited to 10 different phrases. This corresponds to a vector with a length of 10 components. The small number of units influences the succeeding proximity results. Two noun phrases are compared by aligning their vectors and comparing the resemblance of components

between them. The comparison method varies between proximity measures. As stated previously, it is a question of how to weight positive matches, mismatches, and negative matches of components, in relation to the length of the vector. No proximity measure favours all these aspects at the same time. Therefore, proximity results always favour some aspects at the expense of others. However, monotonic proximity measures always produce similar results, not similar association scores, but similar rankings of objects.

As discussed in Chapter 4, the Ochiai association measure applied in the present co-word analysis favours positive matches. This implies that mismatches and negative matches play a minor role when normalizing for the length of a vector. Consider the questionable definitional relationship between '*connective tissue attachment*' and '*regenerative therapy*', as well as the relationship between '*formation*' and '*regenerative therapy*'. The fact is that '*regenerative therapy*' never co-occurs with '*connective tissue attachment*' or '*formation*'. The association profile for '*regenerative therapy*' constitutes four different noun phrases, whereas the association profiles for '*connective tissue attachment*' and '*formation*' constitute six identical noun phrases. The second order association scores between '*regenerative therapy*' and the latter two noun phrases are high because all components (noun phrases) in the association profile of '*regenerative therapy*' have positive matches with the components in the profiles of the latter two phrases. This implies that the 'association profile' of '*regenerative therapy*' is very similar to that of '*connective tissue attachment*' and '*formation*'. Nonetheless, based on the description of enamel matrix proteins in sub-section 8.3.1, the relationships between '*regenerative therapy*' and '*connective tissue attachment*' or '*formation*' is more likely to be one of 'contextual contiguity' than a definitional relationship.

An alternative to the present procedure is to apply non-binary frequency counts instead. Besides the binary presence-absence information, non-binary proximity measures also consider the resemblance in magnitudes of component frequencies between the vectors. As described in Chapter 4, such measures also have downsides, for example in relation to radial monotonicity and component-wise monotonicity. To give an example, if the non-binary cosine measure is applied to calculate the second order association between '*enamel matrix protein*' and '*enamel matrix derivative*', then the association score drops to the mediocre 0.43. Yet, the association scores for '*regenerative therapy*' and '*connective tissue attachment*', as well as '*regenerative therapy*' and '*formation*' only make a small drop to 0.62 in both cases. The association score drops markedly in the former case because there is considerable difference in the distribution of magnitudes between the individual matching

components. The association scores in the latter case are stable because the magnitudes in the matching components resemble each other. As a result, it is our opinion that non-binary counts distort second order co-word analyses of the present kind. Non-binary counts may skew the statistical detection of ‘shared textual contexts’ between two noun phrases.

We therefore conclude that the co-word analysis is able to detect valid definitional relationships given binary frequency counts. However, faulty detections are unavoidable due to the different focus of the proximity measures. When the association profiles are relative short, as in the present case, it is therefore advisable to enforce a high threshold value knowing that some valid relations may be missed.

Consequently, we must conclude that the results based on the present data set are of mixed validity in relation to detection of definitional relationships. Yet, one result is very important. The definitional relationship between ‘*enamel matrix protein*’ and ‘*enamel matrix derivative*’ is the strongest second order association and the most important relation to be verified in the present conceptual network. The two phrases are clearly the most important in concept group 1; this is reflected in their high frequency of occurrence and the fact that they are selected as primary candidate thesaurus terms. Generally, we can therefore expect that definitional relationships are more easily detected given a sufficient number of occurrence counts. A sufficient number of occurrence counts can be expected at least among primary candidate thesaurus terms.

8.3.5 *Visualizing the conceptual network as a PFNET*

Figure 8.13 below illustrates the PFNET representation of the first order associations indicated by the equivalence matrix in Table 8.17. The three primary candidate thesaurus terms are indicated in upper case. Relationships are indicated by the abbreviations used in Table 8.19, and the corresponding association scores are indicated in brackets. The arrows indicate definitional relationships, where the dotted arrows signify that the relationships are questionable. Finally, the dotted circle around the noun phrases in the bottom half of the PFNET signifies that the terminology of the network can be divided into two sub-graphs.

The network representation indicates several interesting relationships between the noun phrases. Strongest links and ‘connectivity’ are the central issues in a PFNET. The present PFNET represents first order associations, thus, we can expect to find the strongest first order associations for each of the 11 noun phrases.

FIGURE 8.13. PFNET representation of 'first order association conceptual network' for concept group 1. (CC = Relationship of 'contextual relationship'; CC(H) = One-sided overlap, indicating a hierarchical relation; and DR = Definitional relationship with a threshold value over 0.80).

Further, several 'extra' relations are incorporated to ensure the connectivity of the network. Finally, we can also expect that the longer the geodesic distance between noun phrases the weaker their first order association. In other words, noun phrases that have definitional relationships between them are most likely located at some distance away from each other in the network.

The straight line from 'enamel matrix protein' via 'cementum' to 'treatment' is the center of the network. The three nodes have the highest betweenness scores (51.111, 71.111, and 46.667 respectively), which indicate that they are the focal nodes in the network (see Appendix 17). Consequently, this line is the main 'artery' of the network. The network can be separated into two sub-graphs. The larger sub-graph of

the network constitute the two focal nodes, '*enamel matrix protein*' and '*cementum*', together with the noun phrases appearing on the three 'veins' appended to these focal nodes. The noun phrases in this sub-graph are characterized by significant first order co-occurrences with '*enamel matrix protein*' (see Table 8.17).

The smaller sub-graph of the network constitutes the focal node of '*treatment*', together with the three noun phrases appended to it. The noun phrases in this sub-graph are characterized by significant first order co-occurrences with '*enamel matrix derivative*' (see Table 8.17). We can observe that the link between '*cementum*' and '*treatment*' is weak at 0.07, thus, we can deduce that this is the separation point in the network that partitions it into two sub-graphs. As a result, the noun phrase '*treatment*' has a 'broker role' in the network, as it ties the sub-graphs together.

As already discussed in sub-section 8.2.5.1 in relation to Figure 8.8, the topology of a PFNET can be very useful for manual thesaurus construction. Appendices from the main 'artery', focal nodes such as candidate thesaurus terms, as well as connectivity between the investigated nodes are interesting. In the present Figure 8.13, the 'veins' going out from '*enamel matrix protein*' indicates two relevant relationships of 'contextual contiguity', one with '*periodontal regeneration*', and one with '*regenerative therapy*'. Likewise, we can expect that '*formation*', '*connective tissue attachment*', and '*cementum*' have strong relationships of 'contextual contiguity' between them. Notice, that '*cementum*' also has a strong relationship of 'contextual contiguity' with '*enamel matrix protein*', and that '*formation*' and '*connective tissue attachment*' are also directly related to '*enamel matrix protein*'. Consequently, in a PFNET, relationships of 'contextual contiguity', either hierarchical or associative are directly visible. In addition, we can expect that most relationships of 'contextual contiguity' are associative relationships.

Most interestingly, the present methodology can create a proximity matrix that enables the PFNET to draw attention to second order associations as well. Noun phrases far apart in the network, and in the present case, separated into sub-graphs, may be in a definitional relationship with each other. Obviously, a first order PFNET cannot explicitly reveal a definitional relationship; it can however draw attention to the relationship. Besides the already discussed definitional relationship between '*enamel matrix protein*' and '*enamel matrix derivative*' illustrated by an arrow in Figure 8.13, there are also potential definitional relationships between '*intrabony defect*' and '*periodontal defect*' or between '*treatment*' and '*regenerative therapy*' and '*periodontal regeneration*'. Unfortunately, these potential definitional relationships do not acquire sufficient second order association scores in order to be detected by the co-word analysis.

We therefore conclude that a conceptual network created by use of co-word analysis should be visualized in a PFNET. The PFNET provides a sound basis for manual interpretation of possible equivalence, hierarchical and associative thesaural relationships. Remember that all noun phrases in the PFNET are semantically related, hence the PFNET might draw attention to second order associations otherwise not detected in the preceding co-word analysis due to the strict threshold values.

8.3.6 Summary of results

The fourth component of the exploratory methodology investigates the ability of co-word analysis to disclose equivalence, hierarchical, and associative relationships in a conceptual network. Thus, the fourth component investigates the second research question of the dissertation. The exploration of the fourth component in the special case study of concept group one from the 2001 periodontology sample, verifies that co-word analysis is able to identify thesaural relationships, albeit with a certain error rate.

As demonstrated in the present section, aggregate categories are needed in order to reveal empirically based thesaural relations by use of co-word analysis. Two categories proposed by Soergel (1974) are updated and applied in the present study of noun phrases. The two categories are relationships of 'contextual contiguity' and definitional relationships. Relationships of 'contextual contiguity' are best detected by studying first order associations between noun phrases, while definitional relationships are best detected by studying second order associations of noun phrases. First order associations include associative relationships and some types of hierarchical relationships, whereas second order associations include equivalence relationships and some other types of hierarchical relationships.

From the conceptual network result of concept group one, we can conclude that the co-word analysis is able to identify relationships of 'contextual contiguity' and definitional relationships. However, the results are of mixed value. First order relationships of 'contextual contiguity' are easy to detect empirically and generally produce valid results. It is even possible in some cases to indicate statistically that a relationship is hierarchical, although in general it is a very difficult procedure, which probably often will result in a number of defective relations.

The exploration of the fourth component demonstrates that the construction procedures of the second and third components of the exploratory methodology, makes it possible to identify first and second order term associations from the same matrix by use of co-word analysis. This is the 'indirect approach' introduced in Chapters 5 and 6. The findings verify that the proposed methodology is able to identify and represent

semantically related terminology in a matrix, despite the fact that the terminology hardly ever co-occurs directly in citation contexts. Instead, the terminology is related indirectly because it shares common textual context with the same concept symbol. This important basis is the major difference between traditional term association analyses, and the term association analyses applied in the present methodology. As a result, the co-word analysis is also able to detect significant second order definitional relationships, albeit with some margin of error.

As discussed in the present section, the problem of defective relationships relates to the frequency counts in the data set, as well as the proximity measures applied for the second order association analysis. Nevertheless, larger frequency counts are no guarantee for a valid result. Likewise, it is very difficult for a proximity measure to consider all vector aspects pertaining to 'association profiles', which are the objects for comparison in a second order association analysis. Thus, in all inevitability the co-word analysis applied in the present methodology will detect valid as well as defective definitional relationships.

As Bookstein and colleagues (2003) admit, to detect the precise type of relationship between two terms by purely statistical means seems not to be feasible. However, such methods, and the results they produce, are very interesting if they are used in combination with manual intellectual analysis (Soergel, 1974; Bookstein et al., 2003). Bookstein and his team (2003) suggest that statistical detection can be used as a front end for an automatic thesaurus construction system. Once statistically interesting term associations are found, another procedure involving a human indexer can classify each association as to the exact semantic relationship between the terms.

This view is in accordance with our objective of the fourth component, the creation of a conceptual network of statistically derived term associations. We do not expect to create a 'perfect' network with exact relations, nor do we claim that this is possible. The assumption investigated relates to the fact that existing relationships between noun phrases based on their meaning result in certain statistical patterns of the phrases in citation contexts. Therefore, the aim is to detect and present such patterns, so that their corresponding thesaural relationships can be considered for manual thesaurus construction. This is clearly possible by the proposed fourth component of the exploratory methodology.

We can conclude in relation to the second research question that the co-word method of the fourth component is able to detect the aggregate thesaural relationship categories introduced by Soergel (1974). The definition of the exact relationships, whether they are equivalence, hierarchical, or associative, is most reliably established by human intellectual interpretation. Further, as demonstrated in sub-section 8.3.5, a

PFNET is a solid visual presentation of a conceptual network created by the present co-word analysis procedure. The PFNET reveals strong first order associations, but a manual intellectual inspection may also reveal second order associations.

Notice that all noun phrases incorporated into the co-word analysis are semantically related. Consequently, the co-word analysis detects strong empirical first and second order relations among the semantically related phrases. The PFNET visualizes and thereby contextualizes the co-word results. Thus, we conclude that it is possible by use of co-word analysis, based on a matrix generated by the second and third components of the exploratory methodology, to produce a conceptual network useful for manual thesaurus construction.

This concludes the thesaurus construction components of the exploratory methodology. The following fifth and last component explores the monitoring of potential terminological and conceptual changes by use of bibliometric methods. Monitoring of such changes is useful for maintenance of thesauri.

8.4 Fifth component: Monitoring of terminological and conceptual changes by use of bibliometric methods

The fifth and final component of the proposed methodology investigates the applicability of bibliometric methods as a tool for monitoring terminological and conceptual changes within a specialty area for the purpose of thesaurus maintenance. Thus, the fifth component explores the third research question in the present dissertation.

In Chapter 6, we discussed that terminology used in documents to describe concepts may change over time. For example, new concepts for which there is no accepted terminology may appear; likewise, terms fall into disuse and disappear, as the concepts they represent change meaning or become obsolete. In other words, language usage is dynamic. The dynamism of language entails that thesaurus terms need to be maintained so that they most correctly reflect contemporary concepts.

As demonstrated by the second and third components of the methodology, a semantically coherent concept group contains concept symbols and portfolios of noun phrases, which reflect upon the same overall concept of the group. The assumption behind the present fifth component is that terminological change over time in the citation contexts attached to a group of semantically related concept symbols, to some degree reflects conceptual change within the specialty area.

The fact that concept symbols are cited references facilitates a retrospective bibliometric analysis of the cited references and their parent concept groups. Hence, to investigate potential terminological changes in relation to a concept, a retrospective bibliometric analysis of a concept group, its member concept symbols, and their attached portfolios of noun phrases, is carried out. It is thereby possible to trace the origin and evolution of a concept group. Furthermore, by investigating the terminology used in citation contexts over time to describe the concept symbols of the group, it may be possible to trace the origin and evolution of the common concept of the group. This implies that we go back in time and examine possible changes in the terminology used in citing papers to convey the concept symbols of a concept group. Hence, we use the cited references and their citation history as a guide to the monitoring of changes in terminology. Changes over time in terminology used to convey concepts may indicate conceptual change.

It is assumed, that potential terminological changes manifest themselves in changing compositions of the noun phrase portfolios attached to a concept group. Conversely, continuity in the usage of a concept manifests itself in a stable composition of the portfolios over time. Notice that the fifth component investigates terminological changes by monitoring concept groups over a period of time. Obviously, an investigation of appearing and disappearing concept groups in annual samples may also indicate conceptual changes within a specialty area. Although, disappearance of concept groups might as well be the result of the co-citation analysis, which tend to emphasize current ‘hot topics’ in a specialty area. Nevertheless, this is not investigated in the present analysis.

Two supportive types of bibliometric methods are applied in the fifth component in order to investigate potential terminological and conceptual changes over time in the specialty area of periodontology. The first method is based on quantitative research methods used to map dynamical aspects of science and technology (Braam, Moed & Van Raan, 1991a; 1991b). The second method is a descriptive bibliometric ageing study of concept symbols (Aversa, 1985; McCain & Turner, 1989).

The first method is based on the methodological steps of the second and third components of the present methodology. From the case study of the 2001 sample, a concept group and its member concept symbols, including their portfolios of noun phrases, are chosen for analysis. The co-citation history of the chosen concept group and its member concept symbols is retrospectively investigated in the three remaining annual samples of citing papers introduced in Chapter 7. The four annual samples of citing papers have an interval of four years between them going back to 1989 (i.e.,

2001, 1997, 1993, and 1989). As a result, the analysis traces the origin and evolution of the 2001 concept group.

The retrospectively identified concept groups are subjected to the methodological steps of the third component, in order to select portfolios of noun phrases attached to the concept symbols of the group. The portfolios of noun phrases are used to create a *concept profile* for the concept group. Remember that members of a concept group most likely refer to a common concept. Thus, the concept profile characterizes the common concept of the group. The concept profile is the basis used for comparison of groups over time. The retrospective analysis therefore compares concept profiles of the different annual samples with the purpose of investigating continuity or change in these profiles. The concept profile analysis is a second order co-occurrence analysis.

The second method is a bibliometric ageing study, which is applied in parallel with the concept profile analysis to support the latter. In Chapter 5, we presented two related bibliometric ageing studies that investigate the citation histories of individual cited references in order to typify general diachronous aging patterns of scientific papers (Aversa, 1985; McCain & Turner, 1989). The results from the studies indicate two different ageing patterns. One group of papers has an ageing pattern in which the received number of citations peak 2-3 years after publication and decline rapidly thereafter. Citations for the other more highly cited group of papers, peak 5-6 years after publication with a slow subsequent decline.

By use of citation context analysis, McCain and Turner (1989) confirmed previous speculation by Garfield (1975; 1984) that the two aging patterns are related to the role cited references play in subsequent citing papers. References cited for their theoretical or technical contributions peak later and age less rapidly than those cited for specific research findings and experimental results. McCain and Turner (1989) suggest that the former subset of papers inspire to new work or offer superior analytical methods of longer duration. The latter subset of papers apparently has immediate and limited application to the field.

As a result, the diachronous ageing pattern of concept symbols from the chosen 2001 concept group is studied in order to verify the role played by these concept symbols in later citing documents (McCain & Turner, 1989). It is assumed that the role played by core concept symbols of a concept group may give some indication of the role, usage, or 'staying power' of the common concept they represent. Core concept symbols imply membership of a concept group for a succession of years.

The fifth component is explored by a retrospective study of a single concept group from the 2001 sample. The co-citation history of the concept group is traced in the three remaining annual samples. The second concept group concerning *guided tissue*

regeneration is chosen for the analysis. The concept group is chosen mainly due to the publication years of its cited references. These references are published before the time period of the present study. Hence, it increases the likelihood of studying the concept group and its members for a longer period of time. In addition, the diachronic bibliometric ageing patterns of the members of the 2001 concept group is revealed in order to support the concept profile analysis.

The following sub-section 8.4.1 presents the retrospective concept profile analysis for the 2001 concept group concerning *guided tissue regeneration*.

8.4.1 Concept profile analysis

As described in Appendix 9, *guided tissue regeneration* is a regenerative technique, which is based on the assumption that periodontal ligament cells have the potential for regeneration of the attachment apparatus of the tooth. Central to the technique of guided tissue regeneration is the application of a barrier membrane. The technique derives from classic studies by NYMAN_82a, NYMAN_82b, GOTTLOW_84, and GOTTLOW_86.

Concept group two in the 2001 sample concerns the concept of *guided tissue regeneration*. The group contains three concept symbols that clearly reflect upon *guided tissue regeneration*, NYMAN_82a, NYMAN_82b, and GOTTLOW_86. The references are among the classic studies that introduced the technique in the early and mid 1980s. The cited references are the basis for the present retrospective analysis. Thus, the cited references are the intellectual base for citing authors in 2001 referring to the concept of *guided tissue regeneration*.

The first step in the retrospective concept profile analysis is to create concept groups for the three remaining annual samples. Document co-citation clustering, similar to the procedure presented in section 8.1, is carried out for the 1989, 1993, and 1997 samples. Note, the three concept symbols from the 2001 group are used to identify the potential concept group(s) denoting *guided tissue regeneration* in the three remaining annual samples. Consequently, we do not consider other concept groups in the present study.

As presented in section 8.1, citation analysis is applied to identify highly cited references for the document co-citation analysis. Unfortunately, none of the three concept symbols from the 2001 sample has sufficient citation counts in 1989 to warrant their inclusion in a co-citation analysis. Table 8.20 presents the citation counts and index scores for the three concept symbols for each of the four annual samples applied in the present analysis. The index score indicates the share of all citations in an annual sample received by the individual cited references. The index score makes it

possible to observe the relative rise and decline of citation counts for a cited reference during a period under study.

TABLE 8.20. *Citation counts and index scores for the three cited references from the 2001 concept group concerning guided tissue regeneration. Citation counts are the actual number of citations in the samples of overlapping documents. The index score is the relative citation count for annual samples. The index score makes it possible to compare citation counts across the annual samples.*

Cited reference	1989		1993		1997		2001	
	Citation count	Index score	Citation count	Index score	Citation count	Index score	Citation count	Index score
NYMAN_82a	3	5.11	23	12.12	23	7.84	13	4.10
NYMAN_82b	3	5.11	24	12.64	30	10.22	16	5.03
GOTTLOW_86	2	3.41	33	17.38	31	10.57	14	4.41

It is clear from Table 8.20 that the low citation counts in the 1989 sample do not warrant an inclusion of the three cited references in a document co-citation analysis. In addition, the index score for the 2001 sample is almost equal to that of 1989. However, in the 2001 sample, the total number of citations is much higher than the 1989 sample. Thus, citation counts of 13, 16, and 14 respectively, warrant inclusion in a co-citation analysis. We therefore proceed to the 1993 and 1997 samples. Here the actual citation counts and the annual shares of citation counts, for all three cited references increase remarkably. In fact, the three cited references are among the top-10 ranked on the citation frequency lists for both annual samples. This means that the three references and the concept they represent play a very important role in defining the 1993 and 1997 research fronts within periodontology.

It is a remarkable increase as it comes 7 and 11 years after the publication of these cited references. Figure 8.14 shows the bibliometric ageing pattern of the three cited references. The graphs illustrate the percentage of the received citations distributed annually from the publication year plus 1 to 2001 for the three cited references. The annual samples investigated are indicated with circles. The year 1992 is interesting, as *guided tissue regeneration* is introduced as a descriptor in MeSH®, six years after the coinage of the term in GOTTLOW_86.

FIGURE 8.14. Citation history and ageing pattern for NYMAN_82a, NYMAN_82b, and GOTTLOW_86. 'guided tissue regeneration' is introduced as a MeSH® descriptor in 1992. The circles indicate the annual samples investigated in the present case study. PY+ specifies the number of years from publication of the reference to its 'citation peak'.

The three graphs in Figure 8.14 show similar ageing patterns. In the immediate years after their publication, these cited references 'live a quiet life'. Then around 1992 a sudden increase in received citations occur, and at the same time *guided tissue regeneration* becomes a descriptor in MeSH®. The latter is an indication of the growing importance of the concept. In the immediate years following 1992, the three cited references have their 'years of glory', with citation peaks 16 (NYMAN_82a), 13 (NYMAN_82b), and 9 (GOTTLOW_86) years after their publication. Finally, a subsequent slow decline in received citations is then observed for all three cited references.

The citation peaks are beyond the category of 5-6 years determined by McCain and Turner (1989), yet the aging patterns clearly belong to this category. This implies that the three cited references, NYMAN_82a, NYMAN_82b, and GOTTLOW_86, most probably are cited for their theoretical or technical contributions to the specialty area of periodontology, and that these papers have inspired to new work or offered superior analytical methods of longer duration (McCain & Turner, 1989). Knowing that the papers are classics and foundational to the regenerative technique of *guided tissue regeneration*, the role indications suggested by McCain and Turner (1989) seems valid. The fact that these cited references are the only ones left in the 2001 concept group representing the concept of *guided tissue regeneration*, evidently shows the inspiration and influence these papers have had on later citing authors. Eventually, the

inspiration and influence of these concept symbols indicate the 'staying power' of the concept they represent, i.e., guided tissue regeneration.

Returning to the concept profile analysis, concept groups are created for the 1993 and 1997 samples. These annual samples cover the 'years of glory' with high citation peaks of the three cited references from the 2001 concept group.

Appendix 19⁶³ shows the resulting dendograms of the document co-citation cluster analyses. From the two clustering results in Appendix 19, we can observe that the original three cited references from the 2001 sample are solidly located together in the largest concept group in both samples. The two concept groups and their cited references are presented in Table 8.21.

TABLE 8.21. *The 1993 and 1997 concept groups denoting 'guided tissue regeneration' and their cited references. The three cited references constituting the 2001 concept group is marked in bold.*

1993: Concept group 4		1997: Concept group 3	
BECKER_87	PONTORIERO_88	BECKER_88	SCHALLHORN_88
GOTTLOW_86	CAFFESSE_90	CAFFESSE_90	GOTTLOW_84
NYMAN_82b	LEKOVIC_89	PONTORIERO_88	NYMAN_82a
GOTTLOW_84	SCHALLHORN_88	GOTTLOW_86	NYMAN_82b
NYMAN_82a	PONTORIERO_89	PONTORIERO_89	MELCHER_76
BECKER_88	MELCHER_76		

The subsequent step is to create a noun phrase portfolio for each individual cited reference of the two concept groups. According to the third component, we ought to identify concept symbols in connection with creation of the portfolios. But, in the present conception of the retrospective analysis, we are not interested in selecting primary and secondary candidate thesaurus terms. We are, however, interested in identifying commonly agreed upon noun phrases within the concept group. The most persistent of these noun phrases are then incorporated into the concept profile denoting the concept group. It is assumed, that the persistent noun phrases are a sufficient basis for the present analysis to infer the common concept referred to by the members of the concept group. Nevertheless, the term selection procedure may be a more solid procedure for the retrospective analysis.

Text files comprising the annual citation contexts for the individual cited references are generated and subsequently parsed with the Connexor shallow parsing software

⁶³ Appendix 19 also include an electronic adjunct on the CD-rom, where the remaining cited references from the 1993 and 1997 concept groups are presented similar to Appendix 12.

(www.connexor.com). This procedure is similar to the one described in sub-section 8.2.3. Likewise, the selection of noun phrases are similar to the procedure introduced in section 8.2.4, where as a rule of thumb we apply a 'citation context frequency threshold value' of around $\frac{1}{3}$ of the sample size. This implies that in order to be selected for the list of frequently occurring phrases for a cited reference, a phrase must occur at least in $\frac{1}{3}$ of the citation contexts. Similar to the third component, the counts are binary. Appendix 20 presents the results of the frequency analyses for the cited references in the three concept groups (i.e., 1993, 1997, and 2001). Appendix 20 shows two matrices for each concept group, six in all. The first matrix indicates the binary counts of noun phrases in the citation contexts of individual cited references, the second matrix is an aggregate binary count of whether a noun phrase is present or absent in the citation contexts for a particular cited reference.

From Appendix 20, we can observe that *guided tissue regeneration* clearly is the most common noun phrase used in the citation contexts of all cited references in the two concept groups. The data are so unequivocal that it is fair to assume that the common concept referred to in the 1993 and the 1997 concept groups is *guided tissue regeneration*, precisely as in the case of the 2001 concept group.

The two concept groups created from the 1993 and 1997 samples, and investigated in the present analysis, are the largest groups in their respective samples. Hence, from a co-citation perspective, at least in the years 1993 and 1997, the 'hottest' concept in periodontology is *guided tissue regeneration*.

The purpose of the present analysis is to monitor potential terminological changes over a period, as it is assumed that terminological changes to some degree reflect conceptual changes. The concept of *guided tissue regeneration* seems to be stable for the period studied. A quantitative comparison of concept profiles attached to the annual concept groups may verify whether there is continuity in the terminology used to characterize the concept for the period studied, or if terminological change is indicated.

A concept profile consists of the most persistent noun phrases in a concept group. Thus, only noun phrases attached to more than one cited reference in a group are included. This procedure resembles the selection of primary candidate thesaurus terms introduced in sub-section 8.2.6.

The decision to only use noun phrases attached to more than one cited reference also serves as a normalization procedure. For example, it appears from Appendix 20 that several variations of '*barrier membrane*' are represented among the noun phrases. Several of these variations only appear once; hence, they are removed due to the normalization procedure. Notice, that the concept profile analysis is carried out by use

of binary counts. It is therefore a question of whether '*barrier membrane*' is present or absent in the profile, and not how many times, or in how many different variations. Obviously, the normalization procedure needs manual verification as not all variations are dealt with. For example, in the 1993 concept group '*barrier membrane*' and '*eptfe*' both appear more than once. They are equivalent phrases and should be merged. Further, normalization across the concept profiles is also needed in order to perform the second order co-occurrence analysis between the profiles. The noun phrase '*eptfe membrane*' is used in the 1997 profile, consequently, we normalize '*barrier membrane*' and '*eptfe*' in the 1993 profile to '*eptfe membrane*' by merging the phrases.

Based on the aggregate binary matrices presented in Appendix 20, in combination with the selection and normalization procedure described above, Figure 8.15 presents the three concept profiles investigated in the present analysis.

FIGURE 8.15. Similarity between the three concept profiles calculated by use of the Ochiai measure. The circles illustrate the concept groups and their attached noun phrases, which constitutes the concept profile. The numbers on the lines indicate the similarity score.

In addition, Figure 8.15 also reveals the similarity between the concept profiles. The similarity is calculated as a second order co-occurrence analysis by use of the binary Ochiai association measure (18). The Ochiai measure expresses the relative number of noun phrases in the intersection of concept profile *X* and concept profile *Y*. The Ochiai measure normalizes for the vector length of the concept profiles. Thus, a long profile can be 'penalized' for its 'representational richness' if it does not correspond to the 'richness' in the short concept profile it is matched with (Jones & Furnas, 1987).

8.4.2 Results of the retrospective analysis

From the perspective of co-citation analysis, there is no doubt that the concept of *guided tissue regeneration* plays a central role in periodontology in 1993 and 1997. In 2001, the centrality of the concept is reduced, as the concept group reflecting it is limited to three cited references. Nonetheless, it is three classic references, which 15 to 20 years after their publication, still receive a sufficient number of citations and co-citations to be included among the core intellectual base of periodontology in 2001. It is also clear that, still in 2001, these three concept symbols commonly denote the concept of *guided tissue regeneration*. Consequently, it seems that the concept has been stable in the period under study.

Not surprisingly, the number of cited references in the concept groups seems to influence the number of noun phrases incorporated into the concept profiles. We can observe that the number of cited references in the concept groups representing *guided tissue regeneration* is reduced over the years to the three core references in 2001. A similar pattern can be observed for the number of noun phrases in the concept profiles. The number is reduced over the years concurrently with the reduction of cited references in the concept groups. Thus, in 2001 only three noun phrases constitute the concept profile.

The concept profile analysis indicates a declining similarity between the annual profiles as the time span between them increases. According to the assumption of the present analysis, this should be interpreted as an indication of terminological change, but this seems questionable in the present case. As stated above, the concept of *guided tissue regeneration* seems to be stable for the period under study. A closer inspection across the concept profiles in fact indicates terminological continuity instead of terminological change. What really changes is the number of noun phrases in the individual profiles. Obviously, the larger 1993 and 1997 concept profiles are more elaborate in their characterization of *guided tissue regeneration* than the 2001 profile. But, it is questionable whether this is an expression of ‘terminological change’ reflected among the phrases in the 2001 profile. It is more likely a simple result of the declining number of cited references in the 2001 concept group, which leads to fewer noun phrases in the concept profile. In fact, it is very likely that there is terminological continuity in the use of noun phrases across the three annual concept profiles. Notice, that the 1993 and 1997 profiles share four noun phrases, while the 2001 profile shares two out of its three noun phrases with 1993 and 1997 profiles. Yet, the similarity between the 2001 and 1993 profiles are lower, than the similarity score between the 2001 and 1997 profiles. The different similarity scores is the result of the diverse lengths of the concept profiles.

The fact is that two noun phrases, *guided tissue regeneration* and *periodontal ligament*, persistently continue to appear in all three concept profiles.

We therefore conclude that the present quantitative concept profile analysis is biased due to the different lengths of the concept profiles. The similarity scores indicate a terminological change in the 2001 citation contexts reflecting the concept of *guided tissue regeneration*. However, all other indications point to conceptual stability and terminological continuity for the period under study. The discrepancy must be ascribed to the different lengths of the concept profiles, especially the very short 2001 profile. A more solid analysis should ensure that profiles do not differ widely in length. For example, the similarity score between the 1993 and 1997 profiles are more reliable, as 0.68 indicates a strong similarity, or rather continuity in the terminology. Besides the overlap of noun phrases, the reliability of the similarity score is most probably a result of longer concept profiles, and almost equal length of their vectors.

8.4.3 Summary and discussion of results

Based on the exploration of the fifth component in the special case study of concept group 2 from the 2001 periodontology sample, we have to conclude that in its present application the concept profile analysis is not able to monitor terminological changes over a period of time in periodontology.

As discussed above, varying lengths of compared concept profiles, especially small versus long profiles, can lead to faulty interpretations of terminological changes, when this in fact is not the case. To some degree, this error can be corrected for, by ensuring that the concept profiles do not differ widely in their vector lengths, as well as enforcing a minimum number of phrases in the profiles if this is possible.

It might be that the fifth component is more suitable for depicting conceptual stability rather than conceptual changes. As illustrated in the present case study of concept group 2, and supported by Braam, Moed and van Raan (1991a; 1991b), core cited references appearing together in a successive number of annual concept groups indicate stability in the intellectual base of the specialty area of periodontology. When the core cited references have consensus terminology attached to them because they act as concept symbols, then it is a sign of conceptual stability or 'staying power' of the concept. Conversely, terminological changes should instead be searched for in concept groups where the publication year of the cited references is so recent that no core references have yet been established.

In relation to the present application of the concept profile analysis, noun phrases from the portfolios of core cited references should perhaps therefore attain higher weights in the concept profiles to indicate the importance of these phrases in the

formation and evolution of a concept. A change of weights implies the use of non-binary proximity measures.

It is well known, and clearly illustrated in the present case study of periodontology, that co-citation analysis reflects the current ‘hot topics’ in a specialty area. The present study has demonstrated what this may imply for the composition of concept profiles over a period of time when the concept group is reduced to few cited references. As discussed in Chapter 5, citation analysis is closely related with ‘obsolescence-like’ conditions of references over time. The fact, that a cited reference becomes ‘obsolete’ or a concept group is gradually reduced and eventually disappears from the result of a cluster analysis, does not necessarily mean that the concept it may reflect has also become obsolete. Obviously, this indicates that the proposed methodology explored in the present doctoral dissertation, is best suitable for detecting ‘hot’ concepts reflected upon in research fronts. It also indicates that monitoring terminological changes quantitatively by use of co-citation analysis and concept profile analysis is very difficult and perhaps not appropriate. Thus, the present application of bibliometric methods is unsuitable for its purpose, which is the monitoring of terminological and conceptual changes within a specialty area

9. Summary and conclusions

The preceding chapters contribute to the entirety of the doctoral dissertation with reference to the formal presentation, the theoretical justification, and the empirically exploration of the main objective investigated in the dissertation.

The main objective of the dissertation is to reintroduce, and further extend, the theoretical and methodological aspects of bibliometric methods to the research area of knowledge organization for the purpose of semi-automatic thesaurus construction. In order to pursue the main objective of semi-automatic thesaurus construction, a proposed bibliometric-based methodology is explored as a supplement to manual intellectual thesaurus construction. The methodology is used as a framework for the investigation of the ability of bibliometric methods to identify candidate thesaurus terms and thesaural relationships, as well as to monitor potential terminological and conceptual changes within a specialty area. Further, due to the success of the bibliometric methods the proposed methodology might also function as a guideline for others of how to apply bibliometric methods to semi-automatic thesaurus construction and maintenance. That is with the exception of the fifth component of the methodology. The intention to explore and further extend the theoretical and methodological aspects for a novel bibliometric based approach to thesaurus construction has been fulfilled.

The present chapter serves several purposes. First, it provides a summary of the main motivations and findings that have inspired and supported the development of the bibliometric based methodology, and eventually defined the objective of the dissertation. This is the focus of treatment in the Chapters 2, 3, 4, and 5 of the dissertation. Secondly, the five components constituting the bibliometric-based methodology are reintroduced. This corresponds to the presentation in Chapter 6. Thirdly, the chapter summarises the major empirical results of the methodological exploration. The empirical results are presented in Chapters 7 and 8. Focus is centered on the three research questions posed in Chapter 1. Finally, some recommendations to future work are suggested. Altogether, the summary of the treatment of these elements in the various chapters reveals the interrelated structure of the dissertation.

9.1 Summary of dissertation objectives and methodological components

Thesaurus construction approaches are typically separated into manual approaches and automatic approaches. As stated in Chapter 1 and discussed in depth in Chapter 2, some form of manual thesaurus construction is mandatory due to the relational complexities, semantic ambiguities, and dynamics, inherent in languages. However, manual construction and maintenance are complex and time consuming. It is therefore beneficial and necessary to combine manual approaches with automatic approaches due to the complementarity of the two approaches (e.g., Anderson & Pérez-Caraballo, 2001a; 2001b).

Automatic approaches reduce manual workload. The merit of these approaches is that they with ease can produce a basis of candidate thesaurus terms based on literary warrant in documents, or they can reveal significant corpus-based semantic relations between terms on a scale otherwise insurmountable for humans to process. Nevertheless, automatic approaches are deficient if used in isolation. Thus, a hybrid of less resource demanding, semi-automatic construction methods, is attractive, as noted by Soergel (1974).

When automatic approaches are used as a tool for thesaurus constructors, and not as a mean in itself, then we speak of semi-automatic thesaurus construction (Soergel, 1974). This is the foundation for the approach explored in the present dissertation. The semi-automatic thesaurus construction approach is conceived as a supplement to the resource demanding approach of manual intellectual thesaurus construction, and not as a construction approach in itself.

The semi-automatic thesaurus construction approach consists of a non-exhaustive exploratory methodology centered on bibliometric methods presented in Chapter 5, and supplemented with automatic methods and techniques presented in the Chapters 3 and 4.

Chapter 3 and 4 of the dissertation present a thorough analysis of the primary conditions for automatic thesaurus construction. The procedure of automatic thesaurus construction is centered on three crucial steps, which are 1) vocabulary construction, 2) term association, and 3) vocabulary organization.

In Chapter 3, the fundamental aspect of term selection is discussed in relation to vocabulary organization. From the analysis in Chapter 3, it becomes clear that automatic single term selection based on statistical modelling of simple term distributions is insufficient. While the 'bag of words' approach is still the most cost-effective approach to automatic indexing of full text documents, it is far too simple for

solid automatic thesaurus construction as it produces excessively many unimportant terms and therefore also many irrelevant relations.

As a result, one objective of the proposed semi-automatic thesaurus construction approach is to identify fewer, more contextual and less ambiguous terms, as such terms are considered to be important candidate thesaurus terms.

Three unconventional elements in relation to automatic term selection are accentuated in Chapter 3. These elements are chosen to support the objective of term selection in the proposed methodology of the semi-automatic approach. The elements include the use of noun phrases as candidate thesaurus terms, the utilization of the structure in scientific documents to locate important candidate thesaurus terms, and finally the utilization of ‘exemplary documents’ to further support the selection of contextually important terms. The rationale for application of these unconventional elements is discussed comprehensively in Chapter 3. Briefly summarised, noun phrases are the most suitable semantic units for indexing. Noun phrases selected in localized segments of domain specific documents are likely to be semantically related. Finally, ‘exemplary documents’ may play a special role in defining the terminology of a scientific domain.

Obviously, the treatment of these elements make the proposed methodology more resource demanding than a traditional ‘bag of words’ approach, however, it is believed that this may be counterbalanced by the production of fewer but more significant candidate thesaurus terms.

The two remaining steps in automatic thesaurus construction, term association and vocabulary organization, are meticulously analyzed in Chapter 4. Most importantly, the analysis demonstrates how recent research, by employing more sophisticated and resource demanding automatic methods and techniques, have been able to produce more refined thesaurus entries than hitherto, and at the same be able to improve performance (e.g., Crouch, 1990; Ruge, 1992; Qiu & Frei, 1993; Grefenstette, 1994; Chen et al., 1997; Schütze & Pedersen; 1997; Woods, 1997; Bookstein et al., 2003). These methods take into consideration such elements as specific characteristics of term distributions, location of terms in documents, domain terminology, noun phrases, the lexical-grammatical and lexical-semantic context of terms, second order co-occurrence analysis, asymmetric proximity measures, and alternative clustering and dimensionality reduction techniques. In other words, elements that go beyond the less sophisticated ‘bag of words’ approach to automatic indexing. The results demonstrate the inexpedience of reducing the costs of automatic thesaurus construction to a

minimum. Evidently, it leads to poor semantic term representations and consequently meagre performance results.

The application of bibliometric methods in the present dissertation closely resembles some of these more sophisticated automatic thesaurus construction approaches, nevertheless some noticeable differences separates the two approaches to thesaurus construction distinctively.

The motivation for investigating the applicability of bibliometric methods for semi-automatic thesaurus construction comes from the notion that highly cited references may act as concept symbols (Small, 1978). A cited reference acts as a concept symbol because it is characterized by consensus terminology in citation contexts of citing papers. Since a concept symbol is a cited reference, bibliometric methods are able to group related concept symbols and their consensus terminology based on co-citation analysis. The potential usefulness of concept symbols for thesaurus construction was first suggested by Rees-Potter (1987; 1989). In the present conception, a cited reference acting as a concept symbol resembles the notion of 'exemplary documents' suggested by Blair and Kimbrough (2002).

The dissertation reintroduce, modify, and expand former research approaches, as we propose a semi-automatic thesaurus construction approach based on bibliometric methods. The main objective for the dissertation is therefore to investigate the extent to which a number of bibliometric methods can identify candidate thesaurus terms, help identify equivalence, hierarchical, and associative relationships, as well as monitor or identify terminological and conceptual changes in a given specialty area. To pursue the main objective, a bibliometric-based methodology of five components is developed and explored.

Chapter 5 thoroughly introduces and discusses the bibliometric methods investigated in the proposed methodology. The bibliometric methods include document co-citation analysis, citation context analysis, co-word analysis, and bibliometric ageing methods. The chapter demonstrates the close resemblance between bibliometric methods based on co-occurrence analysis and the generic term co-occurrence methods presented in Chapters 3 and 4. Where they differ is on units of analysis. In the present application, bibliometric methods depend on distributions of citations in a subject literature, whereas term co-occurrence methods depend on distributions of terms in natural language text. This is the distinctive difference between the present bibliometric-based approach and traditional automatic thesaurus construction approaches.

The dyad relationship between cited and citing documents is the focal point of interest in relation to bibliometric-based thesaurus construction. A concept symbol is characterized by consensus terminology in citation contexts of citing papers. Consequently, the dyad relationship makes it possible to follow the indirect link from the concept symbol to its reference marker in citing papers. It is assumed that the localized citation contexts in domain specific citing papers contain contextual and agreed upon noun phrases that reflect upon the concept symbol. The most persistent of these noun phrases are conceived as important candidate thesaurus terms. Thus, the citation contexts become targets for term selection as they are subjected to shallow noun phrase parsing in the proposed semi-automatic thesaurus construction approach.

Most importantly, it is possible by use of document co-citation analysis to cluster semantically related highly cited references. The resulting group of concept symbols most likely refer to a common concept. The dyad relationship ensures that the noun phrases attached to the concept symbols and reflecting upon the common concept are grouped together regardless of their direct co-occurrence patterns. This we refer to as the *indirect approach*. Theoretically, the indirect approach implies that equivalent terminology used to characterize a concept can be brought together, an important task where traditional automatic thesaurus construction approaches often fail.

These considerations lead to the development of the exploratory methodology used to investigate the applicability of the selected bibliometric methods for semi-automatic thesaurus construction and maintenance in the present dissertation. Chapter 6 thoroughly introduces the exploratory methodology consisting of five components. The individual conceptions and purposes of the five components are restated below.

The **first component** concerns the *creation of a text corpus*, which is the basis for thesaurus construction. A sample text corpus is created by merging document representations from a domain dependent database with representations from a citation index. This is done by use of the so-called data set isolation method by Ingwersen and Christensen (1997). The creation of ‘enhanced document representations’ enables the utilization of cited references for bibliometric purposes, and domain specific descriptors for validation purposes.

In the present application, the outcome of the first component is the basis for the remaining four. However, it is not a requirement that the semi-automatic thesaurus construction approach commences with the first component. If, for example, ‘enhanced document representations’ are not considered essential, then the basis for semi-automatic thesaurus construction might as well be document representations from

a citation index. As the first component is not fundamental for thesaurus construction, it is not encompassed in the research questions of the present dissertation.

The **second component** concerns *vocabulary organization*. The purpose of the second component is to map and structure the text corpus by use of *document co-citation analysis*, utilizing a complete-link clustering algorithm. This creates concept groups containing potential concept symbols.

The second and third components of the methodology are mutually dependent. They are also the most important components in the proposed methodology. The applicability of document co-citation analysis for identification of candidate thesaurus terms is investigated in the first research question. In addition, document co-citation analysis plays an important role in creating the basis for the fourth component. The fourth component concerns term association and is investigated by the second research question. Thus, the ability of document co-citation analysis to group related concept symbols also plays an important role for the second research question.

Note that the second component concerns *vocabulary organization*, while the third component concerns *term selection*. The methodology thereby alters the ordering of the construction steps in relation to automatic thesaurus construction. The altering of the traditional steps in the present methodology introduces the *indirect approach* to the creation of concept groups. The indirect approach enables grouping of equivalent terminology otherwise difficult to cluster from traditional term co-occurrence patterns. It is assumed that terminology attached to concept groups is highly subject specific, contextual, and thus semantically related.

The **third component** concerns *term selection*. The bibliometric method of *citation context analysis*, supported by shallow noun phrase parsing, is investigated for its ability to select candidate thesaurus terms. As stated above, the second and third components of the methodology are mutually dependent. These components are also the most important in the proposed methodology. Their combined ability to select candidate thesaurus terms are investigated in the first research question of the dissertation.

Term selection is based on distributions of citation counts and the loosely defined 'document structure' of citation context. The assumption is that terminology used to describe a concept symbol in the citation contexts of citing papers is less ambiguous, contextual, and agreed upon. The term selection component of the exploratory methodology serves two closely related purposes. Firstly, to establish whether cited references act as concept symbols, and subsequently, to identify the common concepts

expressed by the concept groups in accordance with the concept symbols they contain. Secondly, to select noun phrases from the citation context that surrounds the individual concept symbols within the concept groups. As a result, a portfolio of noun phrases is attached to a concept symbol. From all portfolios attached to concept symbols within a concept group, important primary and secondary candidate thesaurus terms are selected based on frequency analysis. Consequently, the selected candidate thesaurus terms reflect the common concept expressed by the concept group. Noun phrase selection is performed automatically by use of a sophisticated shallow noun phrase parser.

The **fourth component** concerns *term association*. The component investigates the creation of a conceptual network of noun phrases within a concept group by use of *co-word analysis*. As a result, the fourth component depends on the outcome of the second and third components of the methodology. As stated above, the second component of vocabulary organization plays an important role for the ability of co-word analysis to disclose thesaural relationships in the fourth component. The fourth component investigates the second research question.

Frequently occurring noun phrases in a concept group are agreed upon, contextual, and most likely semantically related to each other. In addition, the indirect approach ensures that the noun phrases assigned to a concept symbol appear together because they share the same textual context surrounding the specific reference marker in citing papers. All noun phrases thereby become related to each other, either directly by occurring together in the same context, or indirectly through their common co-occurrence with the cited reference representing the concept symbol. Together all concept symbols and their assigned noun phrases refer to the common concept of the group. Consequently, first and second order co-occurrence analyses can be performed by use of co-word analyses in order to investigate its ability to disclose equivalence, hierarchical, and associative relationships. In addition, network analysis is used as a visual aid for the interpretation of the relational network structures.

The **fifth component** concerns maintenance of thesauri. The component investigates the applicability of a retrospective *concept profile analysis*, supported by a bibliometric ageing study, as a tool for monitoring terminological and conceptual changes within a specialty area for the purpose of thesaurus maintenance.

The fifth component to some degree depends on the outcome of the second and third components of the methodology. The component explores the third research question of the present doctoral dissertation. Persistent noun phrases in a concept

group constitute its *concept profile*. The citation and co-citation history of a concept group and its cited references are studied retrospectively for a given period. Note, a concept group reflects the common concept of its members. Concept profiles are created for earlier manifestations of the investigated concept group. The similarity between the different concept profiles is measured. The degree of similarity between the groups is used as an indication of terminological change. It is assumed that terminological change to some degree may be an indication of conceptual change. The similarity results are compared to the ageing patterns of core cited references that appear successively in the different manifestations of the concept group. The ageing pattern of such references may indicate the 'staying power' of a concept over time.

The five components are explored in the Chapters 7 and 8 of the dissertation. The first component is not encompassed in the research questions, as the component is not fundamental to the proposed semi-automatic thesaurus construction approach. Nevertheless, as in the present case study, we recommend the inclusion of the first component especially in cases where the domain dependent database contains a state-of-the-art controlled vocabulary. The creation of 'enhanced document representations' opens up for a number of comparison and validation possibilities in bibliometric analyses otherwise not feasible. As demonstrated in Chapter 8, the present dissertation successfully applies distributions of MeSH® descriptors to validate the created concept groups. Similar applications seem feasible in general co-citation analysis, where procedures for naming clusters have been debated for years.

The following section summaries the major empirical results of the methodological exploration of the four remaining components in relation to the three research questions posed in Chapter 1.

9.2 Summary of results

To reiterate, the second and third components concern the first research question. The fourth component concerns the second research question, and the fifth component concerns the third research question.

The first and most central research question investigates the basic characteristic of thesaurus construction, which is identification of candidate thesaurus terms:

1. Is it possible, by use of bibliometric methods, to detect candidate thesaurus terms in a specialty area within the life sciences given its disciplinary, publication, and terminological conditions?

The exploration of the second and third components in the case study of periodontology clearly demonstrates that the applied bibliometric methods of co-citation analysis and citation context analysis are able to select important candidate thesaurus terms. The exploration and results are presented in section 8.1 and 8.2 of Chapter 8. We believe that the special selection procedures inherent in the methodical steps of the two components ensure that a significant number of the selected primary candidate thesaurus terms turn out to be important index terms. Hence, the conclusion is that the applied bibliometric methods are very suitable for selection of candidate thesaurus terms in the specialty area of periodontology.

The key to this result is the focus on cited references as the primary unit of analysis. The citation context of a cited reference, acting as a concept symbol, is very interesting in relation to semi-automatic thesaurus construction. The terminology in the narrowly defined text window is highly contextual as it reflects on special characteristics in relation to the concept symbol. Likewise, consensus terminology is also found in the text windows, as the cited reference acts as a concept symbol. These features are very important in relation to selection of index terms. Consequently, the second and third components of the proposed methodology establish literary warrant for a significant number of important contextual and agreed upon primary candidate thesaurus terms. This is a fine basis for subsequent manual thesaurus construction.

As stated above, we believe that the special selection procedures inherent in the two components eventually warrant the promising result for the selected primary terms. One crucial selection procedure is the document co-citation clustering that creates the concept groups. There is no doubt that the unequivocal conceptual meanings among cluster members across all concept groups in the present case study of periodontology, is due to the implementation of the rigid complete-link clustering criterion, which produces clique-like clusters very suitable for conceptual purposes.

The results produced by the citation context analysis extend the findings of Rees-Potter (1987; 1989). We demonstrate the usefulness of citation context analysis for thesaurus construction purposes in a specialty area within the life sciences, dominated by journal papers. We can conclude that it is possible to identify concept symbols in citing journal papers in periodontology, and to extract noun phrases from their citation contexts, which are useful for thesaurus construction.

The investigation also demonstrates the usefulness of noun phrase parsing in citation context analysis. Noun phrase parsing alleviates the time consuming process of phrase identification. In the present approach to thesaurus construction, noun phrase parsing is especially suitable, as it extracts agreed upon terminology from which candidate thesaurus terms are eventually selected. In relation to the term selection method practiced by Rees-Potter (1987; 1989), the application of noun phrase parsing in the present methodology is more appropriate as it is less resource demanding, but still very sufficient.

The second research question investigates to what degree the applied methodology is able to help identify basic thesaural relationships between individual terms and concepts:

2. To what extent can the applied bibliometric methods, outlined in the exploratory methodology of the semi-automatic approach, help identify equivalence, hierarchical and associative relationships between individual terms and concepts within the concept groups?

The exploration of the fourth component in the case study of concept group 1 from the 2001 periodontology sample, verifies that co-word analysis is able to identify thesaural relationships, albeit with a certain error rate. As demonstrated in section 8.3, aggregate categories are needed in order to reveal empirically based thesaural relations by use of co-word analysis. Two categories proposed by Soergel (1974) are updated and applied in the present case study. The two categories are relationships of 'contextual contiguity' and definitional relationships. Relationships of 'contextual contiguity' are best detected by studying first order associations between noun phrases, while 'definitional relationships' are best detected by studying second order associations of noun phrases. First order associations include associative relationships and some types of hierarchical relationships, whereas second order associations include equivalence relationships and some other types of hierarchical relationships.

First order relationships of 'contextual contiguity' are easy to detect empirically and generally produce valid results. It is even possible in some cases to indicate statistically that a relationship is hierarchical. Nevertheless, in general it is a very difficult procedure, which probably often will result in numerous defective relations.

The exploration of the fourth component demonstrates that the indirect approach of the second and third components of the exploratory methodology, makes it possible to identify first and second order term associations from the same matrix by use of co-

word analysis. These findings verify that the proposed methodology is able to identify and represent semantically related terminology in a matrix, despite the fact that the terminology hardly ever co-occurs directly in citation contexts. Instead, the terminology is related indirectly because it shares common textual context with the same concept symbol. This is a major difference from traditional automatic term association analyses. As a result, the co-word analysis is able to detect significant second order definitional relationships, albeit with some margin of error.

As discussed in section 8.3, the problem of defective relationships relates to the frequency counts in the data set, as well as the proximity measures applied for the second order association analysis. Nevertheless, larger frequency counts are no guarantee for a valid result. Likewise, it is very difficult for a proximity measure to consider all vector aspects pertaining to 'association profiles', which are compared in a second order association analysis. Thus, in all inevitability the co-word analysis applied in the present methodology will detect valid as well as defective definitional relationships.

Bookstein and colleagues (2003) state that to detect the precise type of relationship between two terms by purely statistical means seems not to be feasible. However, such methods, and the results they produce, are very interesting if they are used in combination with manual intellectual analysis (Soergel, 1974; Bookstein et al., 2003). Bookstein and his team (2003) suggest that statistical detection can be used as a front end for an automatic thesaurus construction system. Once statistically interesting term associations are found, another procedure involving a human indexer can classify each association as to the exact semantic relationship between the terms.

This view is in accordance with the objective of the fourth component, which is the creation of conceptual networks of term associations. We do not expect to create a 'perfect' network with exact relations, nor do we claim that this is possible. The assumption investigated relates to the fact that existing relationships between noun phrases based on their meaning result in certain statistical patterns of the phrases in citation contexts. Therefore the aim is to detect and present such patterns, so that their corresponding thesaural relationships can be considered for manual thesaurus construction. This is clearly possible by the proposed fourth component of the exploratory methodology.

We conclude in relation to the second research question that the co-word method of the fourth component is able to detect the aggregate thesaural relationship categories introduced by Soergel (1974). The definition of the exact relationships, whether they are equivalence, hierarchical, or associative, must be established by human intellectual interpretation. Further, a PFNET is a solid visual presentation of a conceptual network

created by the co-word analysis. The PFNET reveals strong first order associations, but a manual intellectual inspection may also reveal second order associations. Thus, it is possible by use of co-word analysis, based on a matrix generated by the second and third components of the exploratory methodology, to produce a conceptual network useful for manual thesaurus construction.

The third research question concerns the applicability of bibliometric methods to monitor possible terminological and conceptual changes over a period of time in a specialty area:

3. To what extent can the applied bibliometric methods, outlined in the exploratory methodology of the semi-automatic approach, monitor and identify terminological and conceptual changes in a given subject specialty area over a given time period?

Based on the exploration of the fifth component in the special case study of concept group 2 from the 2001 periodontology sample, we have to conclude that in its present application the concept profile analysis is not able to monitor terminological changes over a period of time in periodontology.

As discussed in section 8.4, varying lengths of compared concept profiles, especially small versus long profiles, can lead to faulty interpretations of terminological changes, when this in fact is not the case.

It might be that the fifth component is more suitable for depicting conceptual stability rather than conceptual changes. As illustrated in the case study of concept group 2, core cited references appearing together in a successive number of annual concept groups indicate stability in the intellectual base of the specialty area of periodontology. When the core cited references have consensus terminology attached to them because they act as concept symbols, then it is a sign of conceptual stability or 'staying power' of the concept.

It is well known, and clearly illustrated in the present case study of periodontology, that co-citation analysis reflects the current 'hot topics' in a specialty area. The present study has demonstrated what this may imply for the composition of concept profiles over a period of time when the concept group is reduced to a few cited references. As discussed in Chapter 5, citation analysis is closely related with 'obsolescence-like' conditions of references over time. The fact that a cited reference becomes 'obsolete' or a concept group is gradually reduced and eventually disappears from the result of a cluster analysis, does not necessarily mean that the concept they may reflect has also become obsolete. This indicates that the proposed methodology explored in the

present dissertation, is best suitable for detecting ‘hot concepts’ reflected upon in research fronts. It also indicates that monitoring terminological changes quantitatively by use of co-citation analysis and concept profile analysis is very difficult and perhaps not appropriate. Thus, similar to the attempt made by Rees-Potter (1987; 1989), the present application of bibliometric methods for the monitoring of terminological and conceptual changes within a specialty area is inappropriate.

9.3 Discussion and recommendations to future work

In continuation of the dissertation summary, comments about future research as well as what could have been done differently in the present dissertation is addressed in this section.

Document co-citation analysis as applied in the present dissertation only focuses on a small number of the core intellectual base references for the 2001 citing sample. This produces few concept groups. The groups reflect the most visible research areas within periodontology, as reflected in the bibliographies of the 2001 citing papers. Several studies have suggested possible recall enhancing procedures in co-citation studies (e.g., Braam, Moed & van Raan, 1991a). However, in relation to the present application they do not seem very profitable. It gives no meaning to enhance the set of references by use of co-word analysis as suggested by Braam, Moed and van Raan (1991a), since the present methodology is strictly centered on concept symbols, i.e., highly cited references; this fact narrows the available references for analysis from the outset. More importantly, in order to create concept groups by use of the indirect approach co-citations are needed. What seems more profitable for future research, is the application of bibliographic coupling in combination with identification of concept symbols in order to broaden the basis for concept group creation and term selection.

While not deemed a serious problem in the present analysis, it is noticeable that noun phrase parsing creates a considerable number of phrase variants. As described in section 8.2, we believe that the methodology may profit by more elaborate phrase normalization procedures, yet, the question is whether it is worth the effort.

This leads to the last and most important consideration, whether the second, third, and fourth components of the methodology are ‘less resource demanding’, and whether the proposed methodology for semi-automatic thesaurus construction is practical or perhaps too detailed.

As presented here, the methodology may seem prolonged and time consuming, but remember we need to evaluate and discuss the components and their individual steps.

The aim with the methodology is that the three appropriate components (four if the first component is included) should be easy to apply. The only real time consuming step is the manual concept symbol analysis. However, the purpose of automatically identification of a 'consensus passage' and the extraction of commonly agreed upon noun phrases are specifically applied to reduce the otherwise time consuming process of concept symbol analysis. In addition, the implementation of the fourth component is very simple as the basis is created by the second and third components.

Eventually, it all comes down to the outcome of the approach. Please keep in mind it is time consuming to manually identify and verify candidate thesaurus terms, as well as investigate their potential relations. As demonstrated in the present dissertation, the second and third components of the proposed methodology are able to produce important candidate thesaurus terms with a literary warrant in the recent scientific literature. Moreover, with some errors, the fourth component is able to disclose two broad categories of thesaural relationships between the selected terms, including equivalent relationships that are otherwise very difficult to identify. In all probability, the semi-automatic thesaurus construction approach is less resource demanding than a manual analysis if its outcome is profitable. We therefore believe that the effort is worthwhile as the present outcome serves as a fine basis for manual thesaurus construction in periodontology. Consequently, the main contributions of the dissertation are:

- overall verification of the applicability of co-citation analysis, citation context analysis, and co-word analysis for semi-automatic thesaurus construction;
- demonstration of the ability of co-citation analysis and citation context analysis to specifically identify important candidate thesaurus terms among a number of potential noun phrases, such terms are important because they are contextual and agreed upon in the scientific community;
- demonstration of the ability of co-citation analysis and co-word analysis to detect thesaural relationships between terms that do not or rarely co-occur directly with each other in citation contexts. This is a direct result of the applied bibliometric based methodology.

We investigate the specialty area of periodontology within the life sciences. The terminology in the life sciences is often connected with stability and consistency. The results of the present dissertation obviously need to be assessed in relation to the usage of terminology in the specialty area. Thus, we cannot draw conclusions that go beyond the specialty area of periodontology. Nevertheless, we can speculate that the proposed

semi-automatic thesaurus construction approach may be suitable in specialty areas that resemble periodontology in relation to terminological usage, most likely other dentistry areas, or perhaps even the entire area of the life sciences.

Finally, we note that automatic methods can assist in, but not replace, the intellectual effort needed for the construction of an indexing language or a thesaurus. To quote Soergel (1974, p. 449) “[s]tatistics should not take precedence over human judgement in the evaluation of vocabulary, but these studies and other provide the basis for some useful decisions. In other words, the identification of terms and relationships by automatic methods should be considered as a kind of pre-processing of open-ended sources, especially abstracts and full-text documents, before transferring terms onto thesaurus forms. The results of this pre-processing are then used in the further steps of thesaurus building. Fully automatic thesaurus building may be attractive as an idea, but it is not feasible”. Hence, the research reported on in the present doctoral dissertation is a contribution to the development of ‘automatic methods’ as tools or supplements for manual intellectual thesaurus construction.

10. References

- Ahlgren, P., Jarneving, B. and Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6), 550-560.
- Aitchison, J., Gilchrist, A. and Bawden, D. (2000). *Thesaurus construction and use: A practical manual* (4th ed.). London: Aslib.
- Aldenderfer, M. S. and Blashfield, R. K. (1984). *Cluster analysis*. Newbury Park, CA.: SAGE Publications Inc.
- Al-Halimi, R. K. and Tompa, F. W. (2003). *Using word position in documents for topic characterization* (Technical Report No. cs-2003-36). Waterloo, Canada: University of Waterloo.
- Al-Hawamdeh, S. and Willett, P. (1989). Paragraph-based nearest neighbour searching in full-text document. *Electronic Publishing: Origination, Dissemination and Design*, 2, 179-192.
- Allen, J. (1995). *Natural Language Understanding* (2 ed.). Redwood City, CA: Benjamin/Cummings.
- Anderberg, M. (1973). *Cluster analysis for applications*. New York: Academic Press.
- Anderson, J. D. (1997). Organization of Knowledge. In J. Feather & P. Sturges (Eds.), *International Dictionary of Library and Information Science* (pp. 336-353). London: Routledge.
- Anderson, J. D. and Pérez-Carballo, J. (2001a). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing and Management*, 37(2), 231-254.
- Anderson, J. D. and Pérez-Carballo, J. (2001b). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machined indexing, and the allocation of human versus machine effort. *Information Processing and Management*, 37(2), 255-277.
- Anick, P. G. and Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Philidelhia, PA, 1997*, 314-323.
- ANSI. (2003). *ANSI/NISO Z39.19 - 2003: Guidelines for the construction, format, and management of monolingual thesauri*. Bethesda, Maryland: American National Standards Institute.
- Armitage, G. (2000). Classifying periodontal diseases: A long-standing dilemma. *Periodontology* 2000, 30, 9-23.
- Arppe, A. (1995). *Term extraction from unrestricted text*. Retrieved 10-03, 2004, from <http://www.lingosoft.fi/doc/nptool/term-extraction.html>
- Art and architecture thesaurus*. (2 ed.). New York, NY: Oxford University Press.
- Attar, R. and Fraenkel, A. S. (1977). Local feedback in full-text retrieval systems. *Journal of the ACM*, 24(3), 397-417.
- Augustson, J. G. and Minker, J. (1970). An analysis of some graph theoretical cluster techniques. *Journal of the Association for Computing Machinery*, 17(4), 571-586.
- Austin, D. (1984). *PRECIS: a manual of concept analysis and subject indexing* (2 ed.). London: British Library Bibliographic Services Division.
- Aversa, E. S. (1985). Citation patterns of highly cited papers and their relationships to literature ageing: A study of the working literature. *Scientometrics*, 7(3-6), 383-389.
- Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

- Baldi, S. (1998). Normative versus social constructivism processes in the allocation of citations: A network-analytic model. *American Sociological Review*, 63(3), 829-846.
- Batagelj, V. and Mrvar, A. (1998). Pajek: A program for large network analysis. *Connections*, 21(2), 47-57.
- Baxendale, P. B. (1958). Machine-made index for technical literature - An experiment. *IBM Journal of Research and Development*, 2(4), 354-361.
- Bazerman, C. (1988). *Shaping Written Knowledge*. Madison, Wisconsin: The University of Wisconsin Press.
- Bean, C. A. (1996). Analysis of non-hierarchical associative relationships among Medical Subject Headings (MeSH): Anatomical and related terminology. In *Advances in Knowledge Organization, proceedings of the international ISKO conference, Washington, DC, 1996* (Vol. 5, pp. 80-86).
- Belew, R. (2000). *Finding out about: A cognitive perspective on search engine technology*. Cambridge, UK: Cambridge University Press.
- Bernstein, L. M. and Williamson, R. E. (1984). Testing of a natural language retrieval system for a full text knowledge base. *Journal of the American Society for Information Science*, 35(4), 235-247.
- Bhattacharya, S. and Basu, P. K. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics*, 43(3), 359-372.
- Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5(4), 257-269.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.
- Bibexcel. www.umu.se/inforsk/Bibexcel/. A tool-box developed by Olle Persson, Inforsk, Umeå University, Sweden.
- Bishop, A. P. (1999). Document structure and digital libraries: how researchers mobilize information in journal articles. *Information Processing and Management*, 35, 225-279.
- Blair, D. C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier.
- Blair, D. C. (2003). Information retrieval and the philosophy of language. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 37, pp. 3-50). Medford, NJ: Information Today Inc.
- Blair, D. C. and Kimbrough, S. O. (2002). Exemplary documents: A foundation for information retrieval design. *Information Processing and Management*, 38, 363-379.
- Bloor, D. (1976). *Knowledge and Social Imagery*. London: Routledge & Kegan Paul.
- Boguraev, B. and Neff, M. (2000). Lexical cohesion, discourse segmentation and document summarization. In *Proceedings of RIAO, Paris, France, 2000* (pp. 962-979).
- Bonzi, S. (1984). Terminological consistency in abstract and concrete disciplines. *Journal of Documentation*, 40(4), 247-263.
- Bonzi, S. and Snyder, H. (1991). Motivations for citation: A comparison of self citation and citations to others. *Scientometrics*, 21(2), 245-254.
- Bookstein, A., Klein, S. T. and Raita, T. (1998). Clumping properties of content-bearing words. *Journal of the American Society for Information Science and Technology*, 49(2), 102-114.
- Bookstein, A., Kulyukin, V., Raita, T. and Nicholson, J. (2003). Adapting measures of clumping strength to assess term-term similarity. *Journal of the American Society for Information Science and Technology*, 54(7), 611-620.

- Bookstein, A. and Raita, T. (2001). Discovering term occurrence structure in text. *Journal of the American Society for Information Science and Technology*, 52(6), 476-486.
- Bookstein, A. and Swanson, D. R. (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5), 312-318.
- Bookstein, A. and Swanson, D. R. (1975). A decision theoretic foundation for indexing. *Journal of the American Society for Information Science*, 26(1), 45-50.
- Borgatti, S. P., Everett, M. G. and Freeman, L. C. (2002). *Ucinet for Windows: Software for Social Network Analysis*. Harvard: Analytic Technologies.
- Borgman, C. L. (1990). *Scholarly communication and bibliometrics*. London: Sage.
- Borgman, C. L. and Furner, J. (2002). Scholarly communication and bibliometrics. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 36, pp. 3-72). Medford, NJ: Information Today, Inc.
- Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10), 913-925.
- Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, 137, 8-96.
- Brin, S. and Page, L. (1998). Anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30, 1-7.
- Brookes, B. C. (1981). The foundations of information science. Part IV. Information science: The changing paradigm. *Journal of Information Science*, 3(1), 3-12.
- Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science*, 36, 223-229.
- Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37, 34-36.
- Brooks, T. A. (1988). Citer motivations. In A. Kent (Ed.), *Encyclopedia of Library and Information science* (Vol. 43). New York: Dekker.
- Braam, R. R., Moed, H. and van Raan, A. F. J. (1988). Mapping of science: Critical elaboration and new approaches, a case study in agricultural biochemistry. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88: Select proceedings of the 1st International Conference on Bibliometrics and Theoretical aspects of Information Retrieval* (pp. 15-18). August, 25-28, 1987, Diepenbeek, Belgium: Amsterdam: Elsevier.
- Braam, R. R., Moed, H. and van Raan, A. F. J. (1991a). Mapping of Science by combined Co-Citation and Word Analysis. I. Structural aspects. *Journal of the American Society for Information Science*, 42(4), 233-251.
- Braam, R. R., Moed, H. and van Raan, A. F. J. (1991b). Mapping of Science by combined Co-Citation and Word Analysis. II. Dynamical aspects. *Journal of the American Society for Information Science*, 42(4), 252-266.
- Buckland, M. (1991). Information as thing. *Journal of the American Society for Information Science*, 42(5), 351-360.
- Buckley, F. and Lewinter, M. (2002). *A Friendly Introduction to Graph Theory*. New Jersey, NJ: Prentice Hall.
- Budd, J. M. (1999). Citations and knowledge claims: sociology of knowledge as a case in point. *Journal of Information Science*, 25(4), 265-274.
- Burton, R. E. and Green, B. A. (1961). Technical reports in physics literature. *Physics Today*, 14(10), 35-37.

- Burton, R. E. and Kebler, R. W. (1960). The 'half-life' of some scientific and technical literatures. *American Documentation*, 11, 18-22.
- Buxton, A. B. and Meadows, A. J. (1977). The variation in the information content of titles of research papers with time and discipline. *Journal of Documentation*, 33(46-52).
- Börner, K., Chen, C. and Boyack, K. W. (2003). Visualizing knowledge domains. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 37, pp. 179-255). Medford, NJ: Information Today, Inc.
- Baayen, R. H. (2001). *Word frequency distributions*. Boston: Kluwer.
- Callan, J. (1994). Passage-level evidence in document retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994* (pp. 302-310).
- Callan, J. P. and Lewis, D. D. (1994). The efficiency issues workshop report. In D. K. Harman (Ed.), *The Second Text REtrieval Conference (TREC-2)* (pp. 303-304). Gaithersburg, MD.: NIST SP 500-215.
- Callon, M., Courtail, J. P. and Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1), 155-205.
- Callon, M., Courtail, J. P., Turner, W. A. and Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2), 191-235.
- Callon, M., Law, J. and Rip, A. (1986a). Qualitative scientometrics. In M. Callon, J. Law & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 103-123). London: The Macmillan Press Ltd.
- Callon, M., Law, J. and Rip, A. (Eds.). (1986b). *Mapping the Dynamics of Science and Technology: Sociology of Science in the real World*. London: The Macmillan Press Ltd.
- Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, 40(4), 284-290.
- Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999 (ACL-99)* (pp. 120-126).
- Card, S. K., MacKinlay, J. D. and Schneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco: Morgan Kaufmann.
- Carroll, J. M. and Roeloffs, R. (1969). Computer selection of keywords using word frequency analysis. *American Documentation*, 20(3), 227-233.
- Case, D. O. and Higgins, G. M. (2000). How can we investigate citation behaviour? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51(7), 635-645.
- Chen, C. (1999). *Information Visualization and Virtual Environments*. London: Springer-Verlag.
- Chen, C. (2003). *Mapping Scientific Frontiers - The Quest for Knowledge Visualization*. Heidelberg: Springer-Verlag.
- Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the United States of America*(101, suppl. 1), 5303-5310.
- Chen, C., Cribben, T., Macredie, R. and Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: Two case studies. *Journal of the American Society for Information Science and Technology*, 53(8), 678-689.

- Chen, C. and Morris, S. (2003). Visualizing Evolving Networks: Minimum Spanning Trees versus Pathfinder Networks. In *Proceeding of the IEEE Symposium on Information Visualization (InfoVis '03)* (pp. 67-74). Seattle, WA: IEEE Computer Society.
- Chen, C. and Paul, R. J. (2001). Visualizing a knowledge domain's intellectual structure. *Computer*, 34, 65-71.
- Chen, C., Paul, R. J. and O'Keefe, B. (2001). Fitting the jigsaw of citation: Information visualization in domain analysis. *Journal of the American Society for Information Science and Technology*, 52(4), 315-330.
- Chen, H. and Lynch, K. J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5), 885-902.
- Chen, H. and Ng, D. T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science*, 46(5), 348-369.
- Chen, H., Ng, T. D., Martinez, J. and Schatz, B. R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the Worm community system. *Journal of the American Society for Information Science*, 48(1), 17-31.
- Chen, H., Yim, T., Fye, D. and Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science*, 46(3), 175-193.
- Chen, Y. S. (1989). Analysis of Lotka's law: The Simon-Yule approach. *Information Processing and Management*, 25(5), 527-544.
- Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings from the conference on User-oriented content-based text and image handling (RIAO'88)*, Cambridge, MA, 1988 (pp. 609-623).
- Chubin, D. E. and Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5(4), 423-441.
- Church, K. and Gale, W. (1995). Poisson mixtures. *Journal of Natural Language Engineering*, 1(2), 163-190.
- Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22-29.
- Cole, J., & Cole, S. (1971). Measuring the quality of sociological research: Problems in the use of the Science citation Index. *American Sociologist*, 6, 23-29.
- Cole, J. R. and Cole, S. (1973). *Social Stratification in Science*. Chicago: University of Chicago Press.
- Cole, S. (1992). *Making science. Between nature and society*. Cambridge, MA.: Harvard University Press.
- Coulter, N., Monarch, I. and Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49(13), 1206-1223.
- Courtail, J. P. (1986). Technical issues and developments in methodology. In M. Callon, J. Law & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 189-210). London: The Macmillan Press Ltd.
- Courtail, J. P. (1998). Comments on Leydesdorff's article. *Journal of the American Society for Information Science*, 49(1), 98.
- Courtail, J. P. and Law, J. (1989). A co-word study of artificial intelligence. *Social Studies of Science*, 19, 301-311.

- Cozzens, S. E. (1982). Split citation identity: a case study from economics. *Journal of the American Society for Information Science*, 33(4), 233-236.
- Cozzens, S. E. (1985a). Comparing the sciences: Citation context analysis of papers from neuropharmacology and the sociology of science. *Social Studies of Science*, 15, 127-153.
- Cozzens, S. E. (1985b). Using the archive: Derek Price's theory of differences among the sciences. *Scientometrics*, 7, 431-441.
- Cozzens, S. E. (1989). What do citations count? The rhetoric first model. *Scientometrics*, 15(5-6), 437-447.
- Croft, W. B., Turtle, H. R. and Lewis, D. (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, 1991* (pp. 32-45).
- Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.
- Cronin, B. (1994). Tiered citation and measures of document similarity. *Journal of the American Society for Information Science and Technology*, 45(7), 537-538.
- Cronin, B., Snyder, H. and Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: A study of sociology. *Journal of Documentation*, 53(3), 263-273.
- Crouch, C. J. (1990). An approach to automatic construction of global thesauri. *Information Processing and Management*, 26(5), 629-640.
- Crouch, C. J. and Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, 1992* (pp. 77-88).
- Cunningham, S. J. and Bock, D. (1995). Obsolescence of computing literature. *Scientometrics*, 34(2), 255-262.
- Dagan, I. (2000). Contextual word similarity. In R. Dale, H. Moisl & H. Somers (Eds.), *Handbook of Natural Language Processing* (pp. 459-475). New York: Marcel Dekker.
- Damerau, F. (1965). An experiment in automatic indexing. *American Documentation*, 16(4), 283-289.
- Damerau, F. (1993). Evaluating domain-oriented multi-word terms from text. *Information Processing and Management*, 29(4), 433-447.
- De Beaugrande, R. A. and Dressler, W. U. (1981). *Introduction to Text Linguistics*. London: Longman.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science and Technology*, 41(6), 391-407.
- Dennis, S. (1965). The construction of a thesaurus automatically from a sample of text. In M. E. Stevens & V. E. Giuliano (Eds.), *Statistical association methods for mechanized documentation*. Washington, DC: National Bureau of Standards.
- Dennis, S. F. (1967). The design and testing of a fully automatic indexing-searching system for documents consisting of expository text. In G. Schecter (Ed.), *Information Retrieval: A Critical Review* (pp. 67-94). Washington D.C.: Thompson Book Company.
- Dextre Clarke, S. G. (2000). Thesaural relationships. In C. A. Bean & R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 37-52). Dordrecht, The Netherlands: Kluwer.
- Dillon, A. (1991). Reader's models of text structures: the case of academic articles. *International Journal of Man-Machine Studies*, 35, 913-925.
- Dillon, M. and Gray, A. S. (1983). FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2), 99-108.

- Ding, Y., Chowdhury, G. G. and Foo, S. (2000). Incorporating the results of co-word analyses to increase search variety for information retrieval. *Journal of Information Science*, 26(6), 429-451.
- Ding, Y., Chowdhury, G. G. and Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management*, 37(6), 817-842.
- Diodato, V. (1982). The occurrence of title words in parts of research papers: variation among disciplines. *Journal of Documentation*, 38, 192-206.
- Diodato, V. (1994). *Dictionary of bibliometrics*. New York: The Haworth Press.
- Doyle, L. B. (1961). Semantic road maps for literature searches. *Journal of the Association for Computing Machinery*, 8(4), 553-578.
- Doyle, L. B. (1962). Indexing and abstracting by association. *American Documentation*, ??, 378-390.
- Driver, H. E. and Kroeber, A. L. (1932). Quantitative expression of cultural relationships. *The University of California Publications in American Archaeology and Ethnology*, 31, 211-256.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.
- Earl, L. L. (1970). Experiments in automatic extracting and indexing. *Information Storage and Retrieval*, 6, 273-288.
- Edge, D. O. (1979). Quantitative measures of communication in science: A critical review. *History of Science*, 17(102-134).
- Edmundson, H. P. and Wyllys, R. E. (1961). Automatic abstracting and indexing - survey and recommendations. *Communications of the ACM*, 4(5), 226-234.
- Efthimiadis, E. (1996). Query expansion. In M. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 31, pp. 121-187). Medford, NJ: Information Today.
- Egghe, L. and Rousseau, R. (1990). *Introduction to Informetrics - Quantitative Methods in Library, Documentation and Information Science*. Amsterdam: Elsevier.
- Elkalifa, E. S. (1991). *The Effect of Collection Homogeneity on Term Association as a Method of Request Expansion in Information Retrieval*. Unpublished Ph.D, Case Western Reserve University.
- Ellis, D., Furner-Hines, J. and Willett, P. (1994). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management*, 3(2), 128-149.
- Evans, D. A., Ginther-Webster, K., Hart, M., Lefferts, R. G. and Monarch, I. A. (1991). Automatic indexing using selective NLP and first-order thesauri. In *Proceedings of RIAO, Barcelona, 1991* (pp. 624-644).
- Everitt, B. S. (1998). *Cluster Analysis* (Third ed.). London: Edward Arnold.
- Everitt, B. S. and Dunn, G. (2001). *Applied Multivariate Data Analysis* (2 ed.). London: Arnold.
- Everitt, B. S. and Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*. London: Arnold.
- Fagan, J. L. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40(2), 115-132.
- Fairclough, N. (1995). *Media Discourse*. London: Edward Arnold.
- Finney, B. (1979). *The reference characteristics of scientific text*. Unpublished M.Sc. dissertation, City University, London.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In F. Palmer (Ed.), *Selected Papers of J.R. Firth* (pp. 1-32). Oxford: Longman.
- Foskett, D. J. (1981). Thesaurus. In A. Kent (Ed.), *Encyclopedia of Library and Information Science* (Vol. 30). New York: Marcel Dekker.

- Fox, C. (1992). Lexical analysis and stoplists. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information Retrieval: Data Structures & Algorithms* (pp. 102-130). Englewood Cliffs, NJ: Prentice Hall.
- Frakes, W. B. (1992). Stemming algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information Retrieval: Data Structures and Algorithms* (pp. 131-160). Englewood Cliffs, NJ: Prentice Hall, Inc.
- Frakes, W. B. and Baeza-Yates, R. (1992). *Information Retrieval: Data Structures & Algorithms*. Upper Saddle River, NJ: Prentice Hall.
- Francis, W. and Kucera, H. (1982). *Frequency analysis of English usage*. Boston, MA: Houghton Mifflin Company.
- Freeman, L. C. (1979). Centrality in social networks: I. Conceptual clarification. *Social Networks*, 1, 215-239.
- Funk, M. E., Reid, C. A. and McGoogan, L. S. (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71, 176-183.
- Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30, 964-971.
- Gale, W., Church, K. and Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26, 415-439.
- Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122, 108-111.
- Garfield, E. (1965). Can citation indexing be automated? In M. E. Stevens & V. E. Giuliano (Eds.), *Statistical Association Methods for Mechanized Documentation* (pp. 189-192). Washington, DC: National Bureau of Standards.
- Garfield, E. (1970). Citation indexing, historio-bibliography, and the sociology of science. In *Proceedings of the Third International Congress of Medical Librarianship* (pp. 187-204).
- Garfield, E. (1974). The citation index as a subject index. *Current Contents*, May(18), 5-7.
- Garfield, E. (1975). The 'obliteration phenomenon' in science and the advantage of being obliterated! *Current Contents*, 51/52, 5-7.
- Garfield, E. (1977). The 'obliteration phenomenon' in science and the advantage of being obliterated. *Essays of an Information Scientist*, *Current Contents*, 2, 396-398.
- Garfield, E. (1979a). *Citation Indexing: Its theory and application in science, technology and humanities*. New York: John Wiley & Sons, Inc.
- Garfield, E. (1979b). Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1, 359-375.
- Garfield, E. (1984). The articles most cited in 1961-1982. 2. Another 100 citation classics highlight the Technology of Science. *Essays of an Information Scientist*, *Current Contents*, 7(29), 218-227.
- Garfield, E. (1986). Mapping the world of biomedical engineering: Alza lecture. *Annals of Biomedical Engineering*, 14, 97-108.
- Garfield, E. (1990). KeyWords Plus. *Current Contents*, 32, 3-7.
- Garfield, E. (1994). Research Fronts. *Current Contents*, 41(3-6).
- Garfield, E., Sher, I. H. and Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia: Institute for Scientific Information.
- Garfield, E. and Welljams-Dorof, A. (1992). Citation data: Their use as quantitative indicators for science and technology evaluation and policy-making. *Science and Public Policy*, 19, 321-327.
- Gauch, S. and Wang, J. (1996). Corpus analysis for TREC-5 query expansion. *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, 537-546.
- Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science*, 7(1), 113-122.

- Giuliano, V. E. (1965). The interpretation of word associations. In M. E. Stevens & V. E. Giuliano (Eds.), *Statistical Association Method for Mechanized Documentation* (pp. 25-33). Washington, DC: National Bureau of Standards.
- Giuliano, V. E. and Jones, P. E. (1963). Linear associative information retrieval. In P. W. Howerton & D. C. Weeks (Eds.), *Vistas in information Handling* (Vol. 1, pp. 30-46). Washington, DC: Spartan Books.
- Glossary of Periodontal Terms*. (4 ed.)(2001). Chicago, Illinois: The American Academy of Periodontology.
- Glänzel, W. and Schoepflin, U. (1994). Little scientometrics - big Scientometrics ... and beyond. *Scientometrics*, 30(2-3), 375-384.
- Glänzel, W. and Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, 21(1), 37-53.
- Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1), 27-57.
- Godby, C. J. (2002). *A computational study of lexicalized noun phrases in English*. Unpublished PhD, Ohio State University.
- Green, R. (1996). The development of a relational thesaurus. *Advances in Knowledge Organization, proceedings of the international ISKO conference, Washington, DC, 1996*, 72-79.
- Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, 1992* (pp. 89-97).
- Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Boston: Kluwer Academic Publisher.
- Griffith, B. C., Small, H. S., Stonehill, J. A. and Dey, S. (1974). The structure of scientific literatures II: Towards a macro- and microstructure for science. *Science Studies*, 4(4), 339-365.
- Grishman, R. and Kittredge, R. (Eds.). (1986). *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Hillsdale, NJ: Lawrence Erlbaum.
- Gross, P. L. K. and Gross, E. M. (1927). College libraries and chemical education. *Science*, 66, 385-389.
- Grossman, D. A. and Frieder, O. (1998). *Information retrieval: Algorithms and heuristics*. Boston: Kluwer Academic.
- Hagstrom, W. O. (1965). *The Scientific Community*. New York: Basic Books.
- Hahn, U. (1990). Topic parsing: Accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26(1), 135-170.
- Halliday, M. A. K. (1989). *Spoken and written language*. Oxford: Oxford University Press.
- Halliday, M. A. K. and Hassan, R. (1976). *Cohesion in English*. London: Longman.
- Hamers, L., Hemeryck, Y., Herweyers, M., Janssen, M., Keters, H., Rousseau, R. and Vanhoutte, A. (1989). Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. *Information Processing and Management*, 25(3), 315-318.
- Hargens, L. L. (2000a). Graphing micro-regions in the web of knowledge: A comparative reference-network analysis. In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honour of Eugene Garfield* (pp. 497-516). Medford, NJ: Information Today, Inc.
- Hargens, L. L. (2000b). Using the literature: Reference networks, reference contexts, and the social structure of scholarship. *American Sociological Review*, 65(12), 846-865.
- Harris, Z. S. (1968). *Mathematical structures of language*. New York: Wiley.

- Harter, S. P. (1975a). A probabilistic approach to automatic indexing: Part I. On the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science*, 25(4), 197-206.
- Harter, S. P. (1975b). A probabilistic approach to automatic indexing: Part II. An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, 25(5), 280-289.
- He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends*, 48(1), 133-159.
- Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING'92), Nantes, 1992* (pp. 539 – 545).
- Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1), 33-64.
- Hearst, M. A. and Plaunt, C. (1993). Subtopic structuring for full-length document access. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, 1993* (pp. 59-68).
- Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 43, 172-198.
- Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science*, 29, 308–310.
- Hindle, D. (1990). Noun classification from predicate argument structures. In *Proceedings of the Association for Computational Linguistics, Pittsburg, 1990*.
- Hinze, S. (1994). Bibliographical cartography of an emerging interdisciplinary discipline: The case of bioelectronics. *Scientometrics*, 29(3), 353-376.
- Hjørland, B. (1997). *Information seeking and subject representation. An activity-theoretical approach to information science*. London: Greenwood Press.
- Hjørland, B. (2002). Domain analysis in information science. Eleven approaches - traditional as well as innovative. *Journal of Documentation*, 58(4), 422-462.
- Hjørland, B. and Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, 46(6), 400-425.
- Hodge, G. (2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files*. Washington, DC: The Digital Library Federation.
- Hodges, T. (1972). *Citation Indexing: Its Potential for Bibliographic Control*. Unpublished PhD, University of California, Berkeley, CA.
- Holgerson, M. (1978). The limited value of cophenetic correlation as a clustering criterion. *Pattern Recognition*, 10, 287-295.
- Holm, B. E. and Rasmussen, L. E. (1961). Development of a technical thesaurus. *American Documentation*(7), 184-190.
- Hooten, P. A. (1991). Frequency and functional use of cited documents in information science. *Journal of the American Society for Information Science*, 42(7), 397–404.
- Houston, J. E. (Ed.). (1995). *Thesaurus of ERIC descriptors*. London: Oryx Press.
- Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews of the Cambridge Philosophical Society*, 57, 669-689.

- Hudon, M. (1996). Preparing terminological definitions for indexing and retrieval thesauri: a model. In *Advances in Knowledge Organization, proceedings of the international ISKO conference, Washington, DC, 1996* (pp. 363-369).
- Hyland, K. (2000). *Disciplinary Discourses: social interactions in academic writing*. Harlow, UK: Pearson Education.
- Ingwersen, P. (1992). *Information Retrieval Interaction*. London: Taylor Graham.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3-50.
- Ingwersen, P. (2001). Cognitive information retrieval. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 34, pp. 3-51). Medford, NJ: Information Today.
- Ingwersen, P. and Christensen, F. H. (1997). Data set isolation for bibliometric online analyses of research publications: Fundamental methodological issues. *Journal of the American Society for Information Science*, 48(3), 205-217.
- ISO. (1986). *ISO 2788: Guidelines for the establishment and development of monolingual thesauri* (2 ed.). Geneva: International Standards Organization.
- Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bulletin et la Société Vaudoise des Sciences Naturelles*, 37, 241-272.
- Jacob, E. K. (2001). The everyday world of work: two approaches to the investigation of classification in context. *Journal of Documentation*, 57(1), 76-99.
- Jacob, E. K. and Shaw, D. (1998). Sociocognitive perspectives on representation. In M. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 33, pp. 131-185). Medford, NJ: Information Today.
- Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999 (ACL-99)*, 341-348.
- Jacquemin, C. and Bourigault, D. (2003). Term extraction and automatic Indexing. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*, (pp. 599-615). Oxford, UK: Oxford University Press.
- Jardine, N. and Sibson, R. (1971). *Mathematical taxonomy*. New York: John Wiley.
- Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7, 217-240.
- Jelinek, F., Mercer, R. L. and Roukos, S. (1992). Principles of lexical language modelling for speech recognition. In S. Furui & M. M. Sondhi (Eds.), *Advances in Speech Signal Processing* (pp. 651-699). Amsterdam: Mercer Dekker.
- Jenkins, W. M. M. and Allan, C. J. (2001). *Periodontics: A synopsis*. Oxford: Butterworth-Heinemann.
- Jensen, S. B., Løe, H., C.R., S. and Theilade, E. (1968). Experimental gingivitis in man. IV. Vancomycin induced changes in bacterial plaque composition as related to the development of gingival inflammation. *Journal of Periodontal Research*, 3, 284-293.
- Johansson, S. and Holland, K. (Eds.). (1989). *Frequency Analysis of English vocabulary and grammar, based on the LOB corpus*. Oxford: Clarendon.
- Jonák, Z. (1984). Automatic indexing of full texts. *Information Processing and Management*, 20(5-6), 619-627.
- Jones, L. P., Gassie, E. W. J. and Radhakrishnan, S. (1990). INDEX: The Statistical basis for an automatic conceptual phrase-indexing system. *Journal of the American Society for Information Science*, 41(2), 87-97.

- Jones, P. E. and Curtice, R. M. (1967). A framework for comparing term association measures. *American Documentation*, 18, 153-161.
- Jones, W. P. and Furnas, G. W. (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the American Society for Information Science*, 38(6), 420-442.
- Jurafsky, D. and Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.
- Justeson, J. S. and Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1), 9-27.
- Järvinen, T. and Tapanainen, P. (1997). *A dependency parser for english* (Technical Report No. TR-1). Helsinki: Department of Linguistics, University of Helsinki.
- Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2), 259-289.
- Kaplan, N. (1965). The norms of citation behaviour: Prolegomena to the footnote. *American Documentation*, 16, 179-184.
- Kaszkiel, J. and Zobel, J. (1997). Passage retrieval revisited. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, 1997* (pp. 178-185).
- Katz, J. and Fordor, J. (1964). The structure of semantic theory. In J. Fordor & J. Katz (Eds.), *The Structure of Language: Readings in the Philosophy of Language* (pp. 479-518). Englewood Cliffs, NJ: Prentice-Hall.
- Katz, S. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1), 15-60.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14, 10-25.
- Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation*, 16, 223-233.
- Kilgariff, A. (1996). Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison. In *Proceedings of the ALLCACH Conference, 1996, Bergen, Norway*: AISB Workshop Series.
- Kilgariff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of the ACL SIGDAT workshop on very large corpora, Beijing and Hong Kong, August, 1997* (pp. 231-245).
- Kilgariff, A. and Rose, T. (1998). Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing, Granada, Spain, 1998* (pp. 46-52).
- Kim, W. and Wilbur, J. (2001). Corpus-based statistical screening for content-bearing terms. *Journal of the American Society for Information Science and Technology*, 52(3), 247-259.
- King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, 13(5), 261-276.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(604-632).
- Knaus, D., Mittendorf, E., Schäuble, P. and Sheridan, P. (1995). Highlighting relevant passages for users of the interactive SPIDER retrieval system. In *Proceedings of the Text Retrieval*

- Conference (TREC), Washington, WA, 1995* (pp. 233-243). Washington: National Institute of Standard and Technology.
- Knorr-Cetina, K. (1981). *The Manufacture of Knowledge: An Essay on the Constructivist and contextual Nature of Science*. Oxford: Pergamon Press.
- Koch, R. (1876). Die aetiologie der Milzbrand-Krankheit, Begrundet auf die Entwick lungsgeschichte des bacillus Anthracis. *Beitrage zur Biologie Pflanzen*, 2, 277-310.
- Kostoff, R. N., Eberhart, H. J. and Toothman, D. R. (1997). Database tomography for information retrieval. *Journal of Information Science*, 23(4), 301-311.
- Kowalski, G. J. and Maybury, M. T. (2000). *Information storage and retrieval systems. Theory and implementation*. Norwell, MA: Kluwer Academic Publishers.
- Kretser, O. D. and Moffat, A. (1999). Effective document presentation with a locality-based similarity heuristic. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California, 1999* (pp. 113-120).
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika*, 29, 1-27.
- Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, Ill.: University of Chicago Press.
- Kuhns, J. L. (1965). The continuum of coefficients of association. In M. E. Stevens & V. E. Giuliano (Eds.), *Statistical Association Methods for Mechanised Documentation* (pp. 33-39). Washington, DC: National Bureau of Standards.
- Kulyukin, V. and Settle, A. (2001). Ranked retrieval with semantic networks and vector spaces. *Journal of the American Society for Information Science and Technology*, 52(14), 1224-1233.
- Kwok, K. L. A. (1985a). A probabilistic theory of indexing and similarity measure based on cited and citing documents. *Journal of the American Society for Information Science*, 36(??), 342-351.
- Kwok, K. L. A. (1985b). A probabilistic theory of indexing using author-provided relevance information. *Proceedings of the American Society for Information Science*, 22, 59-63.
- Lalmas, M. and Ruthven, I. (1998). Representing and retrieving structured documents using the Dempster-Shafer theory of evidence: Modelling and evaluation. *Journal of Documentation*, 54(5), 529-565.
- Lancaster, F. W. (1986). *Vocabulary control for information retrieval* (2 ed.). Arlington, VA: Information Resources Press.
- Lancaster, F. W. (2003). *Indexing and abstracting in theory and practice* (3 ed.). London: Library Association Publishing.
- Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Latour, B. (1987). *Science in action*. Milton Keynes: Open University.
- Latour, B. and Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills, CA: Sage.
- Law, J. and Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics*, 23(3), 417-461.
- Lawrence, S., Giles, C. L. and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6), 67-71.
- Lesk, M. E. (1969). Word-word association in document retrieval systems. *American Documentation*, 20, 8-36.

- Lewis, D. D., Schapire, R. E., Callan, J. P. and Papka, R. (1996). Training algorithms for linear text classifiers. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Zürich, 1996* (pp. 298-306).
- Lewis, D. D. and Sparck Jones, K. (1996). Natural language processing for information retrieval. *Communications of the ACM*, 39(1), 92-101.
- Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics*, 11, 281-320.
- Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science*, 48(5), 418-427.
- Leydesdorff, L. (2004). Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information Science and Technology*, Forthcomming(retrived 27/8-2004: <http://users.fmg.uva.nl/lleydesdorff/jasist04/>).
- Liddy, E. (1991). The discourse level structure of empirical abstracts - an exploratory study. *Information Processing and Management*, 27(1), 55-81.
- Liddy, E. D. and Myaeng, S. H. (1993). DR-LINK's linguistic-conceptual approach to document detection. In D. K. Harman (Ed.), *The First Text REtrieval Conference (TREC-1)*. Gaithersburg, MD: National Institute of Standards and Technology.
- Lin, D. (1998). Automatic retrieval and clustering of similar words. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, 1998 (ACL-98)*, 768-774.
- Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1), 40-54.
- Lin, X., White, H. D. and Buzydlowski, J. (2003). Real-time author co-citation mapping for online searching. *Information Processing and Management*, 38, 689-706.
- Lindholm-Romantschuk, Y. and Warner, J. (1996). The role of monographs in scholarly communication: An empirical study of philosophy, sociology and economics. *Journal of Documentation*, 52(4), 389-404.
- Line, M. (1981). The structure of social science literature as shown by a large-scale citation analysis. *Social Science Information Studies*, 1(1), 67-87.
- Line, M. (1993). Changes in the Use of literature with time - obsolescence revisited. *Library Trends*, 41(4), 665-683.
- Line, M. B. and Sandison, A. (1974). Obsolescence and change in the use of literatur with time. *Journal of Documentation*, 30(3), 283-350.
- Liu, M. (1993). Progress in documentation: The complexities of citation practice - a review of citation studies. *Journal of Documentation*, 49(4), 370-408.
- Losee, R. M. (1988). Parameter estimation for probabilistic document-retrieval models. *Journal of the American Society for Information Science*, 39(1), 8-16.
- Losee, R. M. (1996). Text windows and phrases differing by discipline, location in document, and syntactic structure. *Information Processing and Management*, 32(6), 747-767.
- Losee, R. M. (1998). *Text retrieval and filtering: Analytic models of performance*. Boston: Kluwer.
- Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 16(12), 317-323.
- Lowe, H. J. and Barnett, G. O. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Informatics Association*, 27(14), 1103-1108.
- Luhn, H. P. (1957). A statistical approach to mechanised encoding and searching of library information. *IBM Journal of Research and Development*, 1, 309-317.

- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(159-165).
- Luukkonen, T. (1997). Why has Latour's theory of citations been ignored by the bibliometric community?: Discussion of sociological interpretations of citation analysis. *Scientometrics*, 38, 27-37.
- Lykke Nielsen, M. (2002). *The Word Association Method: a gateway to work-task based retrieval*. Unpublished PhD, Åbo Akademi University, Åbo.
- Löe, H. (1993). Periodontal diseases: a brief historical perspective. *Periodontology* 2000, 2, 7-12.
- Löe, H. and Silness, J. (1963). Periodontal disease in pregnancy. 1. Prevalence and severity. *Acta Odontologica Scandinavica*, 21, 533-551.
- Löe, H., Theilade, E. and Jensen, S. B. (1965). Experimental gingivitis in man. *Journal of Periodontology*, 36, 177-187.
- Löe, H., Theilade, E., Jensen, S. B. and Schiøtt, C. R. (1967). Experimental gingivitis in man. III. The influence of antibiotics on gingival plaque development. *Journal of Periodontal Research*, 2, 282-289.
- MacRoberts, M. H. and MacRoberts, B. R. (1984). The negational reference: Or the art of dissembling. *Social Studies of Science*, 14, 91-94.
- MacRoberts, M. H. and MacRoberts, B. R. (1989a). Another test of the normative theory of citing. *Journal of the American Society for Information Science*, 16, 151-172.
- MacRoberts, M. H. and MacRoberts, B. R. (1989b). Problems of citation analysis: a critical review. *Journal of the American Society for Information Science and Technology*, 40(5), 342-349.
- MacRoberts, M. H. and MacRoberts, B. R. (1996). Problems of Citation Analysis. *Scientometrics*, 36(3), 435-444.
- Malin, M. V. (1968). The science citation index: A new concept in indexing. *Library Trends*, 16(1), 376.
- Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3), 243-281.
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.
- Mantel, N. (1967). A technique of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209-220.
- Mantel, N. and Valand, R. S. (1970). A technique of nonparametric multivariate analysis. *Biometrics*, 26, 547-558.
- Marchionini, G. (1995). *Information seeking in electronic environments*. New York, NY: Cambridge University Press.
- Margulis, E. L. (1992). N-Poisson document modelling. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 1992* (pp. 177-189).
- Marshakova, I. V. (1973). A system of document connection based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2: Informatsionnye Protsessy I Sistemy [Scientific and Technical Information Serial of VINITI 6]*, 6(2), 3-8.
- Matsumoto, Y. (2000). Lexical knowledge acquisition. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 395-414). Oxford: Oxford University Press.
- McCain, K. W. (1986). Cocited author mapping as a valid representation of intellectual structure. *Journal of the American Society for Information Science*, 37(3), 111-122.
- McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433-443.

- McCain, K. W. (1991a). Core journal networks and cocitation maps: New bibliometric tools for serials research and management. *Library Quarterly*, 61(3), 311-336.
- McCain, K. W. (1991b). Mapping economics through the journal literature: an experiment in journal cocitation analysis. *Journal of the American Society for Information Science*, 42(4), 290-296.
- McCain, K. W. (1995). The structure of biotechnology R & D. *Scientometrics*, 32(2), 153-175.
- McCain, K. W. and Turner, K. (1989). Citation context analysis and ageing patterns of journal articles in molecular genetics. *Scientometrics*, 17(1-2), 127-163.
- McGill, M., Koll, M. and Noreault, T. (1979). *An evaluation of factors affecting document ranking for information retrieval systems* (NTIS report No. PB80-119506). Springfield, VA: US Department of Commerce.
- Merton, R. K. (1968). The Matthew Effect in science. *Science*, 159(5), 56-63.
- Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.
- Merton, R. K. (1979). Foreword to E. Garfield, Citation Indexing: Its Theory and Application in Science, Technology, and Humanities. In E. Garfield (Ed.), *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. (pp. v-ix). New York: Wiley.
- Merton, R. K. (1988). The Matthew Effect in science, II: Cumulative advantage and the symbolism of intellectual property. *ISIS*, 79, 606-623.
- Milás-Bracovic, M. (1987). The structure of scientific papers and their author abstracts. *Informatologia Yugoslavica*, 19, 51-67.
- Miller, S. (1984). *Experimental design and statistics* (2 ed.). London: Routledge.
- Miller, U. (1997). Thesaurus construction: Problems and their roots. *Information Processing and Management*, 33(4), 481-493.
- Miller, W. D. (1890). *Original investigations concerning pyorrhea alveolaris. The micro-organisms of the human mouth*. Philadelphia, PA: The S.S. White Dental Mfg. Co.
- Minker, J., Wilson, G. A. and Zimmerman, B. H. (1972). An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8(6), 329-348.
- Mitra, M., Buckley, C., Singhal, A. and Cardie, C. (1997). An analysis of statistical and syntactic phrases. In *Proceedings of Intelligent Multimedia Information Retrieval Systems and Management (RIAO'97), Montreal, 1997* (pp. 200-214).
- Moens, M. F. (2000). *Automatic indexing and abstracting of document texts*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Molholt, P. (1996). Standardization of inter-concept links and their usage. *Advances in Knowledge Organization, proceedings of the international ISKO conference, Washington, DC, 1996*, 65-71.
- Moravcsik, M. J. and Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5, 86-92.
- Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 21-48.
- Morris, S. A. and Yen, G. (2004). Crossmaps: visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Science of the United States*, 101(suppl.1), 5291-5296.
- Morris, S. A., Yen, G., Wu, Z. and Asnake, B. (2003). Timeline visualization of research fronts. *Journal of the American Society for Information Science and Technology*, 55(5), 413-422.

- Motylev, V. M. (1989). The main problems of studying literature aging. *Scientometrics*, 15(1-2), 97-109.
- Mullins, N., Snizek, W. and Oehler, K. (1988). The structural analysis of a scientific paper. In A. F. J. Van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 85-101). Amsterdam: Elsevier Science Publishers.
- Mullins, N. C., Hargens, L. L., Hecht, P. K. and Kick, E. L. (1977). The group structure of cocitation clusters: A comparative study. *American Sociological Review*, 42(4), 552-562.
- Nadel, E. (1981). Citation and co-citation indicators of a phased impact of the BCS theory in the physics of superconductivity. *Scientometrics*, 3, 203-221.
- Naranan, S. (1971). Power law relations in science bibliography - a self-consistent interpretation. *Journal of Documentation*, 27(2), 83-97.
- Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ: Computer Horizons.
- Narin, F. and Hamilton, K. S. (1996). Bibliometric performance measures. *Scientometrics*, 36, 293-310.
- Newman, M. G., Takei, H. H. and Carranza, F. A. (2002). *Carranza's Clinical Periodontology* (9 ed.). Philadelphia, PA: W.B. Saunders Company.
- Noyons, E. C. M. (1999). *Bibliometric mapping as a science policy and research management tool*. Leiden University, The Netherlands: DSWO Press.
- Noyons, E. C. M., Moed, H. F. and Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science*, 50(2), 115-131.
- Noyons, E. C. M. and van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science*, 49(1), 68-81.
- Oberski, J. E. J. (1988). Some statistical aspects of co-citation cluster analysis and a judgement by physicists. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 431-462). Amsterdam: Elsevier Science Publishers.
- Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletins of the Japanese Society for Scientific Fisheries*, 22, 526-530.
- O'Connor, J. (1982). Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management*, 18, 125-131.
- O'Connor, J. (1983). Biomedical citing statements: computer recognition and use to aid full-text retrieval. *Information Processing and Management*, 19(6), 361-368.
- Paice, C. D. and Jones, P. A. (1993). The identification of important concepts in highly structured technical papers. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, WA, 1993* (pp. 69-78).
- Pao, M. L. (1988). Term and citation searching: a preliminary report. *Proceedings of the American Society for Information Science*, 25, 177-180.
- Pao, M. L. and Worthen, D. B. (1989). Retrieval effectiveness by semantic and citation searching. *Journal of the American Society for Information Science*, 40(4), 226-235.
- Peat, H. J. and Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5), 378-383.
- Pejtersen, A. M. (1980). Design of a classification scheme for fiction based on an analysis of actual user-librarian communication, and use of the scheme for control of librarians' search strategies.

- In *Proceedings of the second International Research Forum on Information Science, 1980* (pp. 146-159).
- Pereira, F., Tishby, N. and Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the Association for Computational Linguistics, Columbus, 1993* (pp. 183-190).
- Peritz, B. C. (1983a). Are methodological papers more cited than theoretical or empirical ones? The case of sociology. *Scientometrics*, 5, 211-218.
- Peritz, B. C. (1983b). A classification of citation roles for the social sciences and related fields. *Scientometrics*, 5, 303-312.
- Peritz, B. C. (1992). On the objectives of citation analysis: Problems of theory and method. *Journal of the American Society for Information Science*, 43(4), 448-451.
- Persson, O. (1994). The intellectual base and research front of JASIS 1986-1990. *Journal of the American Society for Information Science*, 45(1), 31-38.
- Pinto Molina, M. (1995). Documentary abstracting: Toward a methodological model. *Journal of the American Society for Information Science*, 46(3), 225-234.
- Popping, R. (2000). *Computer-assisted Text Analysis*. London: Sage Publications.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.
- Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 149, 510-515.
- Price, D. J. D. S. (1970). Citation measures of hard science, soft science, technology, and nonscience. In C. E. Nelson & D. K. Pollock (Eds.), *Communication among scientists and engineers* (pp. 3-22). Lexington, MA.: Heath.
- Price, D. J. D. S. (1986). *Little science, big science...and beyond*. New York: Columbia University Press.
- Qiu, Y. and Frei, H. P. (1993). Concept based query expansion. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 1993*, 160-169.
- Qiu, Y. and Frei, H. P. (1995). *Improving the Retrieval Effectiveness by a Similarity Thesaurus* (Technical Report No. 225). Zurich, Switzerland: Department of Computer Science, Swiss Federal Institute of Technology (ETH).
- Raghavan, V. V. and Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5), 279-287.
- Ranganathan, S. R. (1987). *Colon classification* (7 ed.). Bangalore, India: Sarada Ranganathan Endowment for Library Science.
- Rasmussen, E. (1992). Clustering algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, NJ: Prentice Hill.
- Rau, L. F., Jacobs, P. S. and Zernik, U. (1989). Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing and Management*, 25, 419-428.
- Rees-Potter, L. K. (1987). *A bibliometric analysis of terminological and conceptual change in sociology and economics with the application to the design of dynamic thesaural systems*. Unpublished Ph.D. dissertation, University of Western Ontario, London, Ontario.
- Rees-Potter, L. K. (1989). Dynamic thesaural systems: a Bibliometric study of terminological and conceptual change in sociology and economics with the application to the design of dynamic thesaural systems. *Information Processing and Management*, 25(6), 677-691.
- Resnick, A. (1961). Relative effectiveness of document titles and abstracts for determining the relevance of documents. *Science*, 134, 1004-1005.

- Ries, J. E., Su, K., Peterson, G., Sievert, M. C., Patrick, T. B., Moxley, D. E. and Ries, L. D. (2001). Comparing frequency of content-bearing words in abstracts and texts in articles from four medical journals: an exploratory study. *Medinfo*, 10(1), 381-384.
- Riloff, E. (1995). Little words can make a big difference for text classification. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, 1995* (pp. 130-136).
- Rip, A. J. (1988). Mapping of science: Possibilities and limitations. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 253-273). Amsterdam: Elsevier Science Publishers.
- Roberts, C. W. (Ed.). (1997). *Text Analysis for the Social Sciences - Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, New Jersey: Lawrence Erlbaum associates, publishers.
- Robertson, S. E. and Walker, S. (1994). Some simple effective approximation to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994* (pp. 232-241).
- Robertson, S. E., Walker, S., Hancock-Beaulieu, M. and M., G. (1994). Okapi in TREC3. In *Text Retrieval Conference TREC-3* (pp. 109-126). Gaithersburg, USA: U.S. National Institute of Standards and Technology.
- Rohlf, F. J. (1970). Adaptive hierarchical clustering schemes. *Systematic Zoology*, 19, 58-82.
- Rohlf, F. J. and Fisher, D. R. (1968). Test for hierarchical structure in random data sets. *Systematic Zoology*, 17, 407-412.
- Rolling, L. N. (1965). The role of graphic display of concept relationships in indexing and retrieval vocabularies. In *Classification Research: Proceedings of the Second International Study Conference, Copenhagen, 1965* (pp. 295-320).
- Rorvig, M. (1999). Images of similarity: A visual exploration of optimal similary metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, 50(8), 639-651.
- Rotto, E. and Morgan, R. P. (1997). An exploration of expert-based text analysis techniques for asserting industrial relevance in U.S. engineering dissertation abstracts. *Scientometrics*, 40(1), 83-102.
- Rowley, J. E. (1994). The controlled versus natural indexing language debate revisited: A perspective on information retrieval practice and research. *Journal of Information Science*, 20(2), 108-119.
- Ruge, G. (1992). Experiments on linguistically-based term associations. *Information Processing and Management*, 28(3), 317-332.
- Ruge, G. (1999). Combining corpus linguistics and human memory models for automatic term association. In T. Strzalkowski (Ed.), *Natural language information retrieval* (pp. 75-98). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Sailsbury, B. A. and Stiles, H. E. (1969). The use of the B-coefficient in information retrieval. In *Proceedings of the American Society for Information Science* (Vol. 6, pp. 265-268).
- Salton, G. (1963). Associative document retrieval techniques using bibliographic information. *Journal of the Association for Computing Machinery*, 10, 440-457.
- Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.
- Salton, G. (1971). *The SMART retrieval system: Experiments in automatic document processing*. Englewoods Cliffs, NJ: Prentice-Hall.

- Salton, G. (1972). Experiments in automatic thesaurus construction for information retrieval. *Information Processing*, 71, 115-123.
- Salton, G. (1975). *Dynamic Information and Library Processing*. Englewood Cliff, NJ: Prentice-Hall.
- Salton, G. and Bergmark, D. (1979). A citation study of computer science literature. *IEEE Transactions on Professional Communication*, PC-22(3), 146-158.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 25, 513-523.
- Salton, G. and Buckley, C. (1991). Global text matching for information retrieval. *Science*, 253, 1012-1015.
- Salton, G., Buckley, C. and Smith, M. (1990). On the application of syntactic methodologies in automatic text analysis. *Information Processing and Management*, 26(1), 73-92.
- Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Salton, G., Wong, A. and Yang, S. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18, 613-620.
- Salton, G. and Yang, C. S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4), 351-372.
- Salton, G., Yang, C. S. and Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1), 33-44.
- Salvador, M. R. and Lopez-Martinez, R. E. (2000). Cognitive structure of research: Scientometric mapping in sintered materials. *Research Evaluation*, 9, 189-200.
- Sanderson, M. and Croft, W. B. (1999). Deriving concept hierarchies from text. In *Proceedings of the ACM SIGIR Conference on research in information retrieval, 1999*. (pp. 206-213).
- Sandison, A. (1980). Aging in scientific literature. *Journal of Documentation*, 36(4), 342-343.
- Sandison, A. (1987). Studies of citations and of 'obsolescence'. *Journal of Information Science*, 13, 371-372.
- Schmitz-Esser, W. (1999). Thesaurus and beyond. An advanced formula for linguistic engineering and information retrieval. *Knowledge Organization*, 26(1), 10-29.
- Schneider, J. W. and Borlund, P. (2002). Preliminary study of the potentiality of bibliometric methods for the construction of thesauri. In H. Bruce, R. Fidel, P. Ingwersen & P. Vakkari (Eds.), *Proceedings of the Fourth International Conference on Conceptions of Library and Information Science, Seattle, WA, USA July 21-25, 2002* (pp. 151-165). Greenwood Village, CO: Libraries Unlimited.
- Schneider, J. W. and Borlund, P. (2004). Introduction to bibliometrics for construction and maintenance of thesauri: methodical considerations. *Journal of Documentation*, 60(5), 524-549.
- Schubert, A. and Glänzwil, W. (1986). Mean Response Time. A new indicator of journal citation speed with application to physics journals. *Czechoslovakian Journal of Physics*, B 36(121-125).
- Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder Associative Networks: Studies in Knowledge Organization*. Norwood, New Jersey: Ablex Publishing Corporation.
- Schwarz, C. (1990). Automatic syntactic analysis of free text. *Journal of the American Society for Information Science*, 41(6), 408-417.
- Schütze, H. and Pedersen, J. O. (1997). A co-occurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, 33(3), 307-318.
- Scott, J. (2000). *Social Network Analysis - A Handbook* (2nd ed.). London: Sage Publications.

- Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, 43(9), 628-638.
- Seglen, P. O. (1994). Causal relationship between article citedness and journal impact factor. *Journal of the American Society for Information Science*, 45(1), 1-11.
- Seglen, P. O. (1996). Quantification of scientific article contents. *Scientometrics*, 35(3), 355-366.
- Seglen, P. O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. *Acta Orthopaedica Scandinavica*, 69(3), 224-229.
- Shadish, W. R., Tolliver, D., Gray, M. and Sen Gupta, S. K. (1995). Author judgments about works they cite: Three studies from psychology journals. *Social Studies of Science*, 25, 477-498.
- Shaw, W. M. (1990). Subject indexing and citation indexing. *Information Processing and Management*, 26, 693-718.
- Sheridan, P. and Smeaton, A. F. (1992). The application of morpho-syntactic language processing to effective phrase matching. *Information Processing and Management*, 28(3), 349-369.
- Siegel, S. and Castellan, N. J. (1988). *Nonparametric statistics: For the behavioral sciences* (2 ed.). New York, NY: McGraw-Hill Book Company.
- Silness, J. and Løe, H. (1964). Periodontal disease in pregnancy. 2. Correlation between oral hygiene and periodontal condition. *Acta Odontologica Scandinavica*, 24, 747-759.
- Singhal, A., Buckley, C. and Mitra, M. (1996). Pivoted document normalization. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Zürich, 1996* (pp. 21-29).
- Singhal, A., Mitra, M. and Buckley, C. (1997). Learning routing queries in a query zone. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, 1997* (pp. 25-32).
- Singhal, A., Salton, G., Mitra, M. and Buckley, C. (1996). Document length normalization. *Information Processing and Management*, 32(5), 619-633.
- Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), 143-177.
- Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265-269.
- Small, H. (1977). A co-citation model of a scientific speciality: A longitudinal study of collagen research. *Social Studies of Science*, 7(2), 139-166.
- Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8(3), 327-340.
- Small, H. (1979). Co-citation context analysis: The relationship between bibliographic structure and knowledge. In R. D. Tally & R. R. Deultgen (Eds.), *Proceedings of the American Society for Information Science Annual Meeting, Minneapolis, USA, 1979* (Vol. 16, pp. 270-275). White Plains, NY: Knowledge Industry Publications.
- Small, H. (1980). Co-citation context analysis and the structure of paradigms. *Journal of Documentation*, 36(3), 183-196.
- Small, H. (1982). Citation context analysis. In B. Dervin & M. J. Voigt (Eds.), *Progress in Communication Sciences* (Vol. 3, pp. 287-310). Norwood, N.J.: Ablex.
- Small, H. (1985). The lives of a scientific paper. In K. S. Warren (Ed.), *Selectivity in Information Systems: Survival of the Fittest* (pp. 83-97). New York, NY: Praeger.
- Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society for Information Science*, 37(3), 97-110.
- Small, H. (1993). Macro-level changes in the structure of co-citation clusters: 1983-1989. *Scientometrics*, 26(1), 5-20.

- Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics*, 38, 275-293.
- Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9), 799-813.
- Small, H. and Garfield, E. (1985). The geography of science: disciplinary and national mappings. *Journal of Information Science*, 11(4), 147-159.
- Small, H. and Greenlee, E. (1980). Citation context analysis of a co-citation cluster: Recombinant DNA. *Scientometrics*, 2(4), 277-301.
- Small, H. and Greenlee, E. (1989). A co-citation study of AIDS research. *Communication Research*, 16(5), 642-666.
- Small, H. and Griffith, B. C. (1974). The structure of scientific literature I: Identifying and graphing specialities. *Science Studies*, 4(1), 17-40.
- Small, H. and Sweeney, E. (1985). Clustering the science citation index using co-citations. I. A comparison of methods. *Scientometrics*, 7(3-6), 391-409.
- Small, H., Sweeney, E. and Greenlee, E. (1985). Clustering the science citation index using co-citations. II. Mapping science. *Scientometrics*, 8(5-6), 321-340.
- Smeaton, A. (1992). Progress in the application of natural language processing to information retrieval tasks. *Computer Journal*, 35(3), 268-278.
- Smeaton, A. (2004). Indexing, Browsing and Searching of Digital Video. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 38, pp. 371-407). Medford, NJ: Information Today, Inc.
- Smeaton, A. F. and Rijsbergen, C. J. (1983). The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3), 239-246.
- Smeaton, A. F. and Sheridan, P. (1991). Using morpho-syntactic language analysis in phrase matching. In *Proceedings of RIAO, Barcelona, 1991* (pp. 414-429).
- Smith, L. C. (1981). Citation analysis. *Library Trends*, 30(1), 83-106.
- Sneath, P. and Sokal, R. (1973). *Numerical taxonomy : The principles and practice of numerical classification*. San Francisco, CA: W. H. Freeman.
- Snyder, H., Cronin, B. and Davenport, E. (1995). What's the use of citation: Citation analysis as a literature topic in selected disciplines of the social-sciences. *Journal of Information Science*, 21(2), 75-85.
- Soergel, D. (1974). *Indexing languages and thesauri: construction and maintenance*. Los Angeles, CA: Melville.
- Soergel, D. (1985). *Organizing information*. Orlando, FL: Academic Press.
- Soergel, D. (1999). The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science and Technology*, 50(12), 1119-1120.
- Sokal, R. and Rohlf, F. (1962). The comparison of dendograms by objective methods. *Taxon*, 11, 33-40.
- Sokal, R. and Rohlf, F. J. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research*. New York, NY: W. H. Freeman & Company.
- Sparck Jones, K. (1970). Some thoughts on classification for retrieval. *Journal of Documentation*, 26, 89-101.
- Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworths.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 111-121.

- Sparck Jones, K. (1973). Index term weighting. *Information Storage and Retrieval*, 9(11), 619-633.
- Sparck Jones, K. (1991). The role of artificial intelligence in information retrieval. *Journal of the American Society for Information Science*, 42(8), 558-565.
- Sparck Jones, K. (1992). Thesaurus. In S. C. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence* (2 ed., Vol. 2, pp. 1605-1613). New York, NY: John Wiley & Sons.
- Sparck Jones, K. and Barber, E. O. (1971). What makes an automatic keyword classification effective? *Journal of the American Society for Information Science*, 22(3), 166-175.
- Sparck Jones, K. and Jackson, D. M. (1970). The use of automatically obtained keyword classifications for information retrieval. *Information Storage and Retrieval*, 5(4), 175-201.
- Sparck Jones, K. and Needham, R. M. (1968). Automatic term classifications and retrieval. *Information Storage and Retrieval*, 4(2), 91-100.
- Sparck Jones, K. and Tait, J. I. (1984). Automatic search term variant generation. *Journal of Documentation*, 40(1), 50-66.
- Sperber, D. and Wilson, D. (1995). *Relevance: Communication and Cognition* (2 ed.). Oxford: Basil Blackwell.
- Spiegel, J. and Bennett, E. M. (1965). A modified statistical association procedure for automatic document content analysis and retrieval. In M. E. Stevens & V. E. Giuliano (Eds.), *Statistical Association Method for Mechanized Documentation* (pp. 47-61). Washington, DC: National Bureau of Standards.
- SPINES thesaurus: A controlled and structured vocabulary of science and technology for policy making, management and development.* (1976). Paris: UNESCO.
- Srinivasan, P. (1990). On generalizing the two-poisson model. *Journal of the American Society for Information Science*, 41(1), 61-66.
- Srinivasan, P. (1992). Thesaurus construction. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information Retrieval: Data structures & algorithms* (pp. 161-218). Upper Saddle River, NJ: Prentice Hall.
- Stewart, J. A. (1983). Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces*, 62, 166-184.
- Stiles, H. E. (1961). The association factor in information retrieval. *Journal of the Association for Computing Machinery*, 8(2), 271-279.
- Stone, D. C. and Rubinoff, M. (1968). Statistical generation of a technical vocabulary. *American Documentation*, 19(4), 411-412.
- Strzalkowski, T. (1994). Document indexing and retrieval using natural language processing. *Proceedings of RIAO, Conference Proceedings Intelligent Text and Image Handling*, 131-143.
- Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing and Management*, 31(3), 397-417.
- Strzalkowski, T., Lin, F., Wang, J. and Perez-Carballo, J. (1999). Evaluating natural language processing techniques in information retrieval. In T. Strzalkowski (Ed.), *Natural language information retrieval* (pp. 113-147). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Strzalkowski, T., Guthrie, L., Karlgren, J., Leistenstider, J., Lin, F., Perez-Carballo, J., Straszheim, T., Wang, J. and Wilding, J. (1997). Natural language information retrieval. In D. K. Harman & E. Voorhees (Eds.), *Information Technology: The Fifth Text REtrieval Conference (TREC-5)* (pp. 291-313). Gaithersburg, MD: NIST.
- Sullivan, D., White, D. H. and Barboni, E. J. (1977). The state of a science: Indicators in the specialty of weak interactions. *Social Studies of Science*, 7(2), 167-200.

- Svenonius, E. (1997). Definitional approaches in the design of classification and thesauri and their implications for retrieval and for automatic classification. In *Knowledge organization for information retrieval, proceedings of the International Study Conference on Classification Research, London, 1997* (pp. 12-16).
- Swales, J. M. (1990). *Genre Analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.
- Száva-Kováts, E. (2002). Unfounded attribution of the 'half-life' index-number of literature obsolescence to Burton and Kebler: A literature science study. *Journal of the American Society for Information Science and Technology*, 53(13), 1098-1105.
- Tabah, A. (1999). Literature dynamics: Studies on growth, diffusion, and epidemics. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 34, pp. 249-286). Medford, NJ: Information Today Inc.
- Tague, J. (1969). Association trails. In A. Kent (Ed.), *Encyclopedia of Library and Information Science* (Vol. 2, pp. 55--81). New York: Marcel Dekker.
- Tanimoto, T. T. (1958). *An elementary mathematical theory of classification and prediction* (IBM Internal Report).
- Temmerman, R. (2000). *Towards new ways of terminology description*. Amsterdam, The Netherlands: John Benjamins Publishing Company.
- Teufel, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Text*. Unpublished PhD, University of Edinburgh.
- Theilade, E., Wright, W. H., Jensen, S. B. and Løe, H. (1966). Experimental gingivitis in man. II. Longitudinal clinical and bacteriological investigation. *Journal of Periodontal Research*, 1, 1–13.
- Tijssen, R. J. W. (1992). *Cartography of Science: scientometric mapping with multidimensional scaling methods*. Leiden: DSWO Press, Leiden University.
- Tokunaga, T., Iwayama, M. and Tanaka, H. (1995). Automatic thesaurus construction based on grammatical relations. In *Proceedings of IJCAI* (pp. 1308-1313).
- Tufte, E. (1983). *The Visual Display of Quantitative Information*. Chelshire, CT: Graphics Press.
- Turner, W. A., Chartron, G., Caville, F. and Michelet, B. (1988). Packaging information for peer review: New co-word analysis techniques. In A. F. J. v. Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 291-323). Amsterdam: Elsevier Science Publications.
- Tzoukerman, E., Klavans, J. L. and Jacquemin, C. (1997). Effective use of natural language processing techniques for automatic conflation of multi-word terms: the role of derivational morphology, part of speech tagging, and shallow parsing. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, 1997* (pp. 148-155).
- Vakkari, P. (1997). Information seeking in context: A challenging metatheory. In *Proceedings of an international conference on research in information needs, Seeking and use in different contexts, Tampere, Finland, 1996* (pp. 451-464).
- van Dalen, H. P. and Henkens, K. (2001). What makes a scientific article influential? The Case of demographers. *Scientometrics*, 50(3), 455-482.
- van Dijk, T. A. (1997). The study of discourse. In T. A. van Dijk (Ed.), *Discourse as structure and process* (pp. 1-34). London-New Delhi: Sage.
- van Dijk, T. A. and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

- van Rijsbergen, C. J. (1977). A theoretical basis for use of co-occurrence data in information retrieval. *Journal of Documentation*, 12(2), 95-100.
- van Rijsbergen, C. J. (1979). *Information Retrieval* (2 ed.). London: Butterworths.
- van Raan, A. F. J. (1988). *Handbook of quantitative studies of science and technology*. Amsterdam: Elsevier.
- van Raan, A. F. J. (1997). Scientometrics: State of the art. *Scientometrics*, 38, 205-218.
- van Raan, A. F. J. (1998). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics*, 43(1), 129-139.
- van Raan, A. F. J. (2000). The Pandora's box of citation analysis: Measuring scientific excellence: The last evil? In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A Festschrift in honor of Eugene Garfield* (pp. 301-319). Medford, NJ: Information Today.
- Vechtomova, O., Robertson, S. and Jones, S. (2003). Query expansion with long-span collocates. *Information Retrieval*, 6, 251-273.
- Vinkler, P. (1998). Comparative investigation of frequency and strength of motives toward referencing: The reference threshold model. *Scientometrics*, 43(1), 107-127.
- Voorhees, E. M. (1986). Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing and Management*, 22(6), 465-476.
- Voutilainen, A. (1993). NPTool, a detector of English noun phrases. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Ohio State University, Columbus, Ohio, 1993* (pp. 48-57).
- Wang, P. and White, M. D. (1999). A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages. *Journal of the American Society for Information Science*, 50, 98-114.
- Wang, Z. W., Wong, S. K. M. and Y.Y., Y. (1992). An analysis of vector space models based on computational geometry. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, 1992* (pp. 152-160).
- White, H. D. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology*, 52(2), 87-108.
- White, H. D. (2003a). Author Cocitation Analysis and Pearson's r. *Journal of the American Society for Information Science and Technology*, 54(131), 250-259.
- White, H. D. (2003b). Pathfinder Networks and Author Cocitation Analysis: A Remapping of Paradigmatic Information Scientists. *Journal of the American Society for Information Science and Technology*, 54(5), 423-434.
- White, H. D. (2004a). Citation analysis and discourse analysis revisited. *Applied Linguistics*, 25, 89-116.
- White, H. D. (2004b). Reward, persuasion, and the Sokal Hoax: A study in citation identities. *Scientometrics*, 60(1), 93-120.
- White, H. D. and Griffith, B. C. (1981). Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163-171.
- White, H. D. and Griffith, B. C. (1982). Authors as markers of intellectual space: Co-citation in studies of science, technology and society. *Journal of Documentation*, 38(4), 255-272.
- White, H. D., Lin, X. and Buzydlowski, J. (2001). The endless gallery: Visualizing authors' citation images in the 463 humanities. In *Proceedings of the American Society for Information Science and Technology, Washington, DC, 2001* (pp. 182-189).
- White, H. D. and McCain, K. W. (1989). Bibliometrics. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 24, pp. 119-186). Amsterdam: Elsevir.

- White, H. D. and McCain, K. W. (1997). Visualization of literatures. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 33, pp. 99-168). ??: ??
- White, H. D. and McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4), 327-355.
- White, M. D. and Wang, P. (1997). A qualitative study of citing behaviour: Contributions, criteria, and metalevel documentation concerns. *Library Quarterly*, 67, 122-154.
- Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science*, 19(3), 473-496.
- Wilkinson, R. (1994). Effective retrieval of structured documents. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994*, 311-317.
- Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5), 577-597.
- Wilson, C. S. (1999). Informetrics. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 34, pp. 107-247). Medford, NJ: Information Today, Inc.
- Wolfram, D. (2000). Applications of informetrics to information retrieval research. *Informing Science*, 3(2), 77-82.
- Wolfram, D. (2003). *Applied Informetrics for Information Retrieval Research*. Westport, CT: Libraries Unlimited, Inc.
- Woods, W. A. (1997). *Conceptual Indexing: A better way to organize knowledge* (No. TR-97-61). Palo Alto, CA: Sun Labs Technical Report.
- Wouters, P. (1999). *The Citation Culture*. Unpublished PhD, University of Amsterdam.
- XLSTAT. www.xlstat.com. A statistical tool box developed by Thierry Fahmy.
- Yitzhaki, M. (2002). Relation of the title length of a journal article to the length of the article. *Scientometrics*, 54(3), 435-447.
- Zhai, C., Tong, X., Milic-Frayling, N. and Evans, D. A. (1997). Evaluation of Syntactic Phrase Indexing - CLARIT NLP Track Report. In E. Voorhees & D. K. Harman (Eds.), *The Fifth Text REtrieval Conference (TREC-5), NIST Special Publication 500-238* (pp. 347-358). Gaithersburg, MD,: National Institute of Standards and Technology.
- Ziman, J. M. (1968). *Public knowledge: An essay concerning social dimension of science*. Cambridge: Cambridge University Press.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA.: Addison-Wesley.
- Zobel, J., Moffat, A., Wilkinson, R. and Sacks-Davis, R. (1995). Efficient retrieval of partial documents. *Information Processing and Management*, 31(3), 361-377.
- Zuckerman, H. (1987). Citation analysis and the complex problem of intellectual influence. *Scientometrics*, 12, 329-338..