Collocation and Term Extraction Using Linguistically Enhanced Statistical Methods

Dissertation
zur Erlangung des akademischen Grades
Doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Philosophischen Fakultät
der Friedrich-Schiller-Universität Jena

von **Joachim Wermter**, MA (Master of Arts)
geboren am 10.08.1969 in Horb am Neckar

Gutachter

1. Prof. Dr. Udo Hahn

2. Prof. Dr. Rüdiger Klar

3. Prof. Dr. Adrian P. Simpson

Tag des Kolloquiums: 04.08.2008

Acknowledgements

My heartfelt thanks go out to Prof. Dr. Udo Hahn (Professor of Computational Linguistics at Jena University) who greatly supported me during the course of my work in a unique way. Without his advice and assistance this work probably would not have seen the light of day. I would also like to thank Prof. Dr. Rüdiger Klar and PD Dr. med. Stefan Schulz from the Department of Medical Informatics at Freiburg University Hospital who greatly helped me at the early stages of my research within the DFG-funded MorphoSaurus project.

Next, I would like to thank Sabine Demsar, Kristina Meller, and Konrad Feldmeier, who did a great job at annoting the PNV triple candidates as a collocation gold standard. Many thanks also to the developers of the Umls Metathesaurus for setting up and maintaing such a great terminological resource. Both efforts made the evaluative part of this work possible in the first place.

Many thanks also to Dr. med Peter Horn (Hannover Medical School) who greatly assisted me in selecting the right MeSH terms for HSCT and immunology.

My dearest thanks go out to my wife Holly. Without her constant support and encouragement, particularly in difficult times, the whole enterprise would not have been worth while.

Contents

1 Introduction **1**

1.1 Main Objectives and Contributions 4

1.2 Structure of this Thesis 6

2 Defining Collocations and Terms **9**

2.1 Defining Collocations 10

2.1.1 Defining Collocations from the Lexicographic Perspective 11

2.1.1.1 The Meaning-based Lexicographic Approach to Collocations 12

2.1.1.2 Other Lexicographic Accounts of Collocations 16

2.1.2 Defining Collocations from the Frequentist Perspective 18

2.1.2.1 Firth's Model of Language Description 19

2.1.2.2 The Collocational Layer of Firth's Model 20

2.1.2.3 Neo-Firthian Developments: Halliday and Sinclair . . . 21

2.1.3 Defining Collocations from a Computational Linguistics Perspective 23

2.1.4 Linguistic Properties of Collocations Adopted 27

2.1.4.1 Four Basic Characteristic Properties 28

2.1.4.2 Demarcation Line: Collocations vs. Free Word Combinations 31

2.2 Defining Terms 34

2.2.1 General Theory of Terminology 36

2.2.2 Beyond General Theory of Terminology 38

2.2.3 Conventional Definitions of Terms 39

2.2.4 Problems with the Classical and Conventional Approaches . . . 40

2.2.5 Pragmatic Definitions of Terms 42
2.2.6 Terms and Sublanguage 43
2.2.7 (Computational) Linguistic Definitions of Terms 46
2.3 Assessment of Linguistic Definitions for Collocations and Terms 49

3 Approaches to the Extraction of Collocations and Terms 53
3.1 Approaches to Collocation Extraction 54
3.1.1 Berry-Rogghe 55
3.1.2 Smadja 56
3.1.3 Lin 59
3.1.4 Evert and Krenn 62
3.2 Approaches to Term Extraction 65
3.2.1 Justeson and Katz 66
3.2.2 Daille 68
3.2.3 Frantzi and Ananiadou 69
3.2.4 Jacquemin 73
3.3 Lexical Association Measures and their Application 75
3.3.1 Statistical Foundations 77
3.3.2 T-test 82
3.3.3 Log-Likelihood 84
3.3.4 Mutual Information 85
3.3.5 Extensions to Larger-Size N-Grams 86
3.3.6 Shortcomings and Linguistic Filtering 88
3.3.7 Frequency of Co-Occurrence 90
3.3.8 C-value 91

4 Linguistically Enhanced Statistics To Measure Lexical Association 93
4.1 Statistical Requirements 95
4.1.1 Avoidance of Non-Linguistic Assumptions 96
4.1.2 Extensibility of Size 97
4.1.3 The Frequency Of Co-Occurrence Factor 98
4.1.4 Output Ranking 98
4.2 Linguistic Requirements 100

4.2.1 Firth as Linguistic Frame of Reference . . . 101
4.2.2 Linguistic Requirements for Collocations . . . 103
4.2.3 Linguistic Requirements for Terms . . . 105
4.3 Limited Syntagmatic Modifiability for Collocation Extraction . . . 108
4.3.1 Defining Limited Syntagmatic Modifiability . . . 109
4.3.2 Illustrating Limited Syntagmatic Modifiability . . . 111
4.4 Limited Paradigmatic Modifiability for Term Extraction . . . 114
4.4.1 Defining Limited Paradigmatic Modifiability . . . 114
4.4.2 Illustrating Limited Paradigmatic Modifiability . . . 117
4.5 Evaluation Setting . . . 119
4.5.1 General Requirements for Evaluation . . . 120
4.5.1.1 Assembling the Text Corpus . . . 121
4.5.1.2 Extracting and Counting the Candidates . . . 121
4.5.1.3 Classification of the Targets . . . 123
4.5.1.4 Quantitative Performance Evaluation . . . 125
4.5.1.5 Baselines and Significance . . . 128
4.5.1.6 Qualitative Performance Evaluation . . . 130
4.5.2 Evaluation Setting for Collocation Extraction . . . 131
4.5.2.1 Text Corpus and Linguistic Filtering . . . 131
4.5.2.2 Target Structure and Candidate Sets . . . 133
4.5.2.3 Classification of Candidate Set and Quality Control . . . 135
4.5.3 Evaluation Setting for Term Extraction . . . 139
4.5.3.1 Text Corpus and Linguistic Filtering . . . 139
4.5.3.2 Target Structures and Candidate Sets . . . 141
4.5.3.3 Classification of Candidate Sets . . . 143

5 Experimental Results 147
5.1 Experimental Results for Collocation Extraction . . . 148
5.1.1 Quantitative Results . . . 148
5.1.1.1 Results on Performance Metrics . . . 149
5.1.1.2 Results on Significance Testing . . . 157
5.1.2 Qualitative Results . . . 158
5.1.2.1 Results on the Static Criteria . . . 159

5.1.2.2 Results on the Dynamic Criteria 161
5.1.3 Limited Syntagmatic Modifiability Revisited 165
5.2 Experimental Results for Term Extraction 167
5.2.1 Quantitative Results 167
5.2.1.1 Results on Performance Metrics 168
5.2.1.2 Results on Significance Testing 180
5.2.2 Qualitative Results 183
5.2.2.1 Results on the static criteria 184
5.2.2.2 Results on the dynamic criteria 189
5.3 Assessment of Experimental Results 196

6 Conclusions and Outlook **201**

7 Summary **205**

A Collocation Classification Manual **207**
A.1 Definitionen aus der einschlägigen Literatur 207
A.2 Unsere Richtlinien für die Verwendung des Begriffs ‘Kollokation’ 208
A.3 Details zur Klassifizierung 210

B MeSH Terms and UMLS Source Vocabularies **213**
B.1 MeSH Terms 213
B.2 UMLS Source Vocabularies 215

List of Tables

3.1 Observed and marginal frequencies 78
3.2 Expected frequencies and their computation from marginal frequencies . . . 79
3.3 Observed and marginal frequencies for a German PNV collocation. 80
3.4 Expected frequencies for the same German PNV collocation. 80

4.1 Collocational and non-collocational PNV Triples with Associated Syntagmatic Attachments 112
4.2 Support Verb Construction PNV Triple with Associated Syntagmatic Attachments 113
4.3 Possible selections for $k = 1$, $k = 2$ and $k = 3$ for a trigram noun phrase 115
4.4 LPM and k-selection modifiabilities for $k = 1$ and $k = 2$ for the trigram term "*open reading frame*" 118
4.5 LPM and k-selection modifiabilities for $k = 1$ and $k = 2$ for the trigram non-term "*t cell response*" 118
4.6 Context of quantitative performance evaluation. 126
4.7 The McNemar significance test of differences for comparing two lexical association measures (LAMs). 129
4.8 Frequency distribution of PNV triple tokens and types for our 100-million-word German newspaper corpus 134
4.9 Frequency distribution of PNV triple tokens and types for 10 million words of German newspaper corpus 134
4.10 Proportion of actual PNV triple collocations and sub proportions of the three collocational categories. 136
4.11 Ranges of the Kappa coefficient and designated strengths of agreement . . . 137
4.12 Observed coarse-grained collocation classifications by two annotators 137

4.13 Overview of agreement rates and Kappa values for coarse-grained classification. . . . 138
4.14 Overview of agreement rates and Kappa values for fine-grained classification. 138
4.15 Frequency distribution for n-gram noun phrase term candidate tokens and types for the 100-million-word MEDLINE text corpus 142
4.16 Frequency distribution for n-gram noun phrase term candidate tokens and types for the 10-million-word MEDLINE text corpus 143
4.17 Proportion of actual terms among the bigram, trigram and quadgram term candidates in the large and small corpus. . . . 144

5.1 Precision scores of association measures for collocation extraction on the 114 million word (upper table) and 10 million word (lower table) German newspaper corpus. . . . 150
5.2 Recall scores of association measures for collocation extraction on the 114 million word (upper table) and 10 million word (lower table) German newspaper corpus. . . . 152
5.3 F-scores of association measures for collocation extraction on the 114 million word (upper table) and 10 million word (lower table) German newspaper corpus. . . . 154
5.4 Fallout scores of association measures for collocation extraction on the 114 million word (upper table) and 10 million word (lower table) German newspaper corpus. . . . 156
5.5 Collocation extraction: significance testing of differences using the two-tailed McNemar test at 99% confidence interval on the large and the small corpus . 158
5.6 Results on the two static criteria for upper-portion targets (Ts) and lower-portion non-targets (NTs) on the large corpus. . . . 159
5.7 Results on the two static criteria for upper-portion targets (Ts) and lower-portion non-targets (NTs) on the small corpus. . . . 160
5.8 Results on the dynamic qualitative criteria 3 for upper-portion non-targets (NTs) on the large corpus. . . . 161
5.9 Results on the dynamic qualitative criteria 4 for lower-portion targets (Ts) on the large corpus. . . . 163
5.10 Results on the two dynamic qualitative criteria for upper-portion non-targets (NTs) and lower-portion targets (Ts) on the small corpus. . . . 164

5.11 Bigram precision scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. . . . 169
5.12 Trigram precision scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. . . . 170
5.13 Quadgram precision scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. . . . 172
5.14 Bigram recall scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. 173
5.15 Trigram recall scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. 175
5.16 Quadgram recall scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. . . . 176
5.17 Bigram fallout scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. 178
5.18 Trigram fallout scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. . . . 179
5.19 Quadgram fallout scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus. . . . 180
5.20 Bigram term extraction: significance testing of differences using the two-tailed McNemar test at 99% confidence interval on the large and the small MEDLINE corpus . . . 181
5.21 Trigram term extraction: significance testing of differences using the two-tailed McNemar test at 99% confidence interval on the large and the small MEDLINE corpus . . . 182
5.22 Quadgram term extraction: significance testing of differences using the two-tailed McNemar test at 99% confidence interval on the large and the small MEDLINE corpus . . . 183
5.23 Results on the two static qualitative criteria for bigram term extraction on the large MEDLINE corpus. . . . 184

5.24 Results on the two static qualitative criteria for bigram term extraction on the small MEDLINE corpus. 185
5.25 Results on the two static qualitative criteria for trigram term extraction on the large MEDLINE corpus. 186
5.26 Results on the two static qualitative criteria for trigram term extraction on the small MEDLINE corpus. 187
5.27 Results on the two static qualitative criteria for quadgram term extraction on the large MEDLINE corpus. 187
5.28 Results on the two static qualitative criteria for quadgram term extraction on the small MEDLINE corpus 188
5.29 Results on the two dynamic qualitative criteria for bigram term extraction on the large MEDLINE corpus. 190
5.30 Results on the two dynamic qualitative criteria for bigram term extraction on the small MEDLINE corpus. 191
5.31 Results on the two dynamic qualitative criteria for trigram term extraction on the large MEDLINE corpus. 192
5.32 Results on the two dynamic qualitative criteria for trigram term extraction on the small MEDLINE corpus. 194
5.33 Results on the two dynamic qualitative criteria for quadgram term extraction on the large MEDLINE corpus. 195
5.34 Results on the two dynamic qualitative criteria for quadgram term extraction on the small MEDLINE corpus 196

List of Figures

2.1 The lexical-collocational layer of Firth's model of language description. . . . 20

4.1 The lexical-collocational layer of Firth's model of language description. . . . 102

5.1 Collocation precision on 114 million word corpus 149
5.2 Collocation precision on 10 million word corpus 149
5.3 Collocation recall on 114 million word corpus 151
5.4 Collocation recall on 10 million word corpus 151
5.5 Collocation F-score on 114 million word corpus 153
5.6 Collocation F-score on 10 million word corpus 153
5.7 Collocation ROC curve on 114 million word corpus 155
5.8 Collocation ROC curve on 10 million word corpus 155
5.9 Qualitative criterion 3: non-targets moved from upper to lower portion (Left: LSM rank compared to frequency rank. Right: t-test rank compared to frequency rank). 162
5.10 Qualitative criterion 4: targets moved from lower to upper portion (Left: LSM rank compared to frequency rank. Right: t-test rank compared to frequency rank). 163
5.11 Distribution of syntagmatic attachments for collocations and non-collocations. The x- and y-axes are log-scaled to improve visibility. 165
5.12 Distribution of syntagmatic attachments for the three collocation categories. The x- and y-axes are log-scaled. 166
5.13 Bigram term precision on 100 million word corpus 169
5.14 Bigram term precision on 10 million word corpus 169
5.15 Trigram precision on 100 million word corpus 170
5.16 Trigram precision 10 million word corpus 170

5.17 Quadgram term precision on 100 million words . . . 171
5.18 Quadgram term precision on 10 million words . . . 171
5.19 Bigram term recall on 100 million words . . . 173
5.20 Bigram term recall on 10 million words . . . 173
5.21 Trigram term recall on 100 million words . . . 174
5.22 Trigram term recall on 10 million words . . . 174
5.23 Quadgram term recall on 100 million words . . . 176
5.24 Quadgram term recall on 10 million words . . . 176
5.25 Bigram term ROC on 100 million words . . . 177
5.26 Bigram term ROC on 10 million words . . . 177
5.27 Trigram term ROC on 100 million words . . . 178
5.28 Trigram term ROC on 10 million words . . . 178
5.29 Quadgram term ROC on 100 million words . . . 179
5.30 Quadgram term ROC on 10 million words . . . 179
5.31 Criterion 3 for t-test trigrams on large corpus. . . . 193
5.32 Criterion 3 for LPM trigrams on large corpus. . . . 193
5.33 Criterion 4 for t-test trigrams on large corpus. . . . 193
5.34 Criterion 4 for LPM trigrams on large corpus. . . . 193

Chapter 1

Introduction

By the time John Rupert Firth (Firth, 1957) framed his famous slogan "*You shall know a word by the company it keeps!*", he may not have known that he not only set the stage for a whole school of linguistic research, British empiricist contextualism, but also drew the attention of many computational linguists to the language phenomena he was explicitly and implicitly referring to – collocations and terms. But why do computational linguists even need to worry about these two kinds of linguistic expressions? The answer is that collocations and terms are pervasive in natural language and, for this reason, any language processing application has to find ways to tackle them. What makes these two types of expressions different – although to different degrees – is that their recognition, extraction and interpretation in natural language text falls outside the realm of standard procedures applied to the "typical" language constructions which obey the rules of syntax and semantic compositionality and which typically encompass natural language processing (NLP) engines such as part-of-speech (POS) taggers, syntactic parsers, and semantic interpreters. In fact, it is typically the case that collocations and terms as multi-word expressions need to be treated by language processing modules as a sort of atomic linguistic units which need not further be analyzed as they already denote stand-alone linguistic or conceptual entities.

Although Firth (1957) did not explicitly refer to the notion of "term" (or "terminological expression") as a distinctive linguistic unit, we will see in this thesis that the linguistic property deducible from his slogan – frequency of co-occurrence – applies both to collocations and to terms. In fact, this property has turned out to be so

prominent that almost all of the computational linguistics research dedicated to the tasks of collocation and term extraction from natural language text data employs dedicated statistical machinery – *lexical association measures* – which to varying degrees capitalize on this property and utilize it in various, sometimes quite sophisticated ways. While the reason for this prominence may certainly be sought in the empirical turnaround that field underwent in the mid-1990s, it has the effect that various statistical and linguistic aspects are ignored or never even considered. On the statistical side, much of the statistical machinery employed both for the extraction of collocations and terms – having been originally devised for completely different tasks such as significance testing for various experimental design set-ups – relies on assumptions that are typically not borne out by the probability distributions of natural language data (cf. section 3.3 of this thesis). Admittedly, it may be justified to overlook such rather theoretical concerns if standard statistical association measures[1] were to have a formidable application performance in extracting collocations and terms from text. Unfortunately, there is more than just spurious evidence in both the research literature on collocation extraction (cf. section 3.1) and on term extraction (cf. section 3.2) which indicates that *plain* frequency of co-occurrence counting of collocation and term candidates appears to perform equally well.

In case, at this point, the impression may be conveyed that measuring statistical association is an enterprise not worth undertaking, some clarifications are in order. First of all, measuring the *lexical association* between words is in fact essential in any attempt to isolate collocations and terms from their non-specific (i.e. non-collocation and non-term) counterparts in text. The reason for this may be sought in the primary task of a lexical association measure, *viz.* to determine the degree of *collocativity* or *termhood* of a certain *collocation candidate* or a certain *term candidate*.[2] In fact, lexical associations in the form of collocativity or termhood have time and again been described as the procedural backbones of applications tackling collocation ex-

[1] As we will see in section 3.3, strictly speaking, not all of these association measures are "statistical" in the sense of testing for some null hypothesis (e.g. the t-test), but some do derive their theoretical underpinnings from information theory (e.g. mutual information).

[2] Here, the legitimate question may be how collocation and term candidates are actually obtained in the first place. In fact, most approaches perform various degrees of linguistic processing on the text corpus data from which collocations and terms are to be extracted, ranging from part of speech tagging to full syntactic parsing. From the linguistic structures assigned in this way, collocation and term candidates may be identified (see subsection 3.3.6).

traction (Evert, 2005; Manning & Schütze, 1999) and term extraction (Jacquemin, 2001) from natural language text. Second of all, measuring lexical association between words may already inherently be conceived of as a statistical task which needs to be performed on the basis of empirical observations on natural language data. The labor-intensive and costly alternative to this would be to set up collocation or term lists completely manually – either through manual text corpus analysis or through linguistic introspection. Although such (electronic) resources do, of course, exist in forms of collocation lexicons or term databases, they tend to be notoriously incomplete, as has also been noticed by studies on collocation extraction (Lin, 1998b) and on term extraction (Daille, 1994). The reason for this incompleteness is to be sought in the productivity and creativity of natural language – one of its fundamental properties – which obviously also holds for collocations and terms. Thus, since such linguistic expressions are constantly being coined, it is almost impossible not to resort to some form of automatic text corpus-based statistical machinery.

The question which then naturally follows from our previous observations is not whether lexical association measures should be based on statistical procedures and computations, but whether these may not be utilized in such a way that – instead of computing their scores based on criteria from the realm of statistical hypothesis testing or information theory – they employ more linguistically based parameters. The source from which such linguistic parameters need to be fed naturally lies in natural language text data itself or, to be more exact, in observable and quantifiable properties of natural language text data. Now, the issue then is what such observable and quantifiable linguistic properties may be in the case of collocations and terms, besides the frequency of co-occurrence property already adduced above. In fact, an examination of the linguistic research literature on collocations and on term shows various directions to investigate these questions and, therefore, it will constitute one of the major foci of this work (cf. chapter 2). What this thesis aims to show is that there is indeed a linguistic property which fulfills the criterion of being a valid linguistic paramter, *viz. limited modifiability*. As collocations and terms are different kinds of linguistic expressions, however, this property is manifested differently in the two types of constructions, i.e. whereas it is expressed *syntagmatically* in the case of collocations, it is done so *paradigmatically* in the case of terms (cf. chapter 4). In fact, this property may also well be motivated within the lexical-collocational layer of Firth's (1957) model of language description which serves as an appropriate linguistic

frame of reference.[3] But even if we are able to motivate and establish both observable and quantifiable linguistic properties for collocations and for terms and, furthermore, also incorporate these into linguistically motivated statistical association measures, the whole enterprise would be futile if our newly coined lexical association measures were not able to perform better than their standard statistical competitors at the task they are designed for – extracting collocations and terms from text. Hence, an integral part of this work will be to show whether this is the case and, for this purpose, to establish and carry out a thorough and comparative performance evaluation (cf. chapter 5).

1.1 Main Objectives and Contributions

The main contributions of this work are centered around five objectives which will be outlined in the following. Each of these objectives is motivated by the gaps that research on the deployment of lexical association measures for the tasks of automatically extracting collocations and terms from natural language text corpora exhibit. While a preview of these shortcomings has already been given, the research goals established from them will be taken on either in particular sections or throughout the whole of this thesis.

1. We will substantiate and define two new linguistically motivated statistical association measures in a language- and domain-independent manner. While their task will be identical compared to their standard statistical and information-theoretic competitors – the computations of lexical association scores to determine the degree of collocativity and termhood of candidate items – their defining parameters will be based on actual linguistic properties of the targeted linguistic constructions, *viz.* collocations and terms.

2. We will show that there are linguistic differences between collocations and terms that need to be considered both for the task of isolating observable and quantifiable linguistic properties and for establishing an appropriate evaluation setting. In particular, it will become clear that while collocations are general-language

[3] One should keep in mind that also the linguistic property already mentioned, frequency of co-occurrence, may be motivated within Firthian linguistics.

constructs which may surface in a wide variety of syntactic expressions, terms are basically confined to subject-specific sublanguage domains and mainly appear in noun phrases.

3. The linguistically observable and quantifiable property isolated for both collocations and terms – limited modifiability – will be structured within an appropriate linguistic frame of reference, *viz.* the lexical-collocational layer of Firth's (1957) model of language description. With its help, it will be possible to account for the linguistic differences and the distinct kinds of syntactic environments in which collocations and terms surface.

4. We will establish a comprehensive performance evaluation setting in which we will be able to compare the linguistically enhanced association measures for collocation extraction (limited syntagmatic modifiability – LSM) and for term extraction, (limited paradigmatic modifiability – LPM) against their standard frequency-based, statistical and information-theoretic competitors. In particular, while our evaluation will be run on a wide array of standard quantitative performance metrics, we will also contribute a new qualitative performance evaluation metric that compares the output rankings of an association measure to a challenging baseline – frequency of co-occurrence.

5. Finally, we will show that our linguistically enhanced term and collocation association measures outperform their competitors by large margins at every aspect of performance evaluation considered. Hence, lexical association measures which base their statistical computations on linguistic parameters instead of standard statistical ones not only exhibit conceptual but also empirical superiority.

Some preliminary discussions and results of the research presented here have already been published in the following conference proceeding papers: the linguistically enhanced association measure LSM for collocation extraction in Wermter & Hahn (2004); the linguistically motivated association measure LPM for term extraction as well as evaluation aspects in Wermter & Hahn (2005c), Wermter & Hahn (2005b) and Wermter & Hahn (2005a); the comparative qualitative evaluation setting in Wermter & Hahn (2006).

1.2 Structure of this Thesis

As it is first mandatory to substantiate in detail the characteristic features of collocations and terms, chapter 2 will zoom in on their linguistic properties, which have been put forth in the scientific literature. We will focus on the observations that, from a conceptual and linguistic point of view, collocations and terms denote different linguistic entities and surface in different linguistic contexts, both syntactically and pragmatically. At the same time, however, it will become clear that there is a linguistic property – limited modifiability – which both collocations and terms share but which is manifested differently in both kinds of linguistic expressions, i.e. while it surfaces *syntagmatically* in collocations, it is manifested *paradigmatically* in terms.

Chapter 3 gives an extensive overview over the most representative and influential approaches to collocation and term extraction that have been proposed in the computational linguistics research literature. Although we divide the computational approaches into those tackling collocation extraction, on the one hand, and term extraction, on the other hand, the methodological boundaries between them are not always as clear-cut as the boundaries in the linguistic literature between collocations and terms, as discussed in chapter 2. We will see that the processing machinery applied to both kinds of linguistic expressions – both in terms of linguistic processing and the lexical association measures applied – is similar if not even equal in both cases. Because it is essential for understanding their shortcomings, this chapter will also feature an extensive discussion on the underlying statistical properties of the standard frequency-based, statistical and information-theoretic association measures. In addition, this discussion will highlight the fact there is already one prominent linguistic property of collocations and terms which all standard measures exploit to various degrees – frequency of co-occurrence.

As the centerpiece of this thesis, chapter 4 will motivate, define and illustrate the two linguistically enhanced approaches to statistically measure lexical association for collocations and for terms, *viz.* limited syntagmatic modifiability (LSM), for the case of collocation extraction, and limited paradigmatic modifiability (LPM), for term extraction. This will be done only after having formulated both their statistical and their linguistic requirements which will derived from the observations established in chapters 2 and 3. It will be shown that, on the statistical side, we have to make sure that we do not make any assumptions that run contrary to the properties of natural

language in general as well as collocations and terms in particular. On the linguistic side, we will ensure that we utilize observable properties suitable to be formalized and quantified in a such manner that they may be used by a statistical procedure. This chapter will also extensively lay out and implement the requirements for constructing an extensive comparative testing ground in order to thoroughly evaluate both linguistic measures against their competitors. For collocation extraction, in particular, the evaluation setting will be on German-language preposition-noun-verb collocation candidates, while for term extraction it will be on English-language noun phrase term candidates from the biomedical domain.

Then, chapter 5 will report on the experimental results obtained for both the collocation extraction and the term extraction tasks as were outlined in the evaluation settings established in chapter 4. Both the quantitative and the qualitative performance evaluations for the collocation extraction and term extraction tasks will show that the linguistically motivated association measures outperform the standard frequency-based, statistical and information-theoretic association measures by large margins in every respect. Importantly, an extensive analysis of the results will summarize the commonalities and differences between our linguistically motivated association measures at their respective tasks.

Finally, chapter 6 draws the main conclusions from the research presented in this thesis and points out further directions of research stemming from this work.

Chapter 2

Defining Collocations and Terms

Since the main goals of this thesis are the definition, implementation and evaluation of statistical association measures which incorporate linguistic properties of collocations and terms, it is first mandatory to substantiate in detail the characteristic features of these linguistic expressions which have been put forth in the scientific literature. One can imagine that the research literature on the issue of collocations and of terms is vast and that any attempt to provide an overview will necessarily have to zoom in on the main aspects, in particular within the context of a computational approach like this one. As the first two sections on defining the notion of collocations (section 2.1) and the notion of terms (section 2.2) will show, these two kinds of linguistic expressions have received quite different treatments in the respective research literature. At first sight, this is not astonishing because from a conceptual and linguistic point of view, collocations and terms denote different linguistic entities and surface in different linguistic contexts. What is remarkable though (and will be discussed extensively in chapter 3) is the fact that the computational approaches to their automatic extraction from unrestricted text have been very similar in terms of the association measures and extraction procedures applied.

One of the insights that this chapter aims to articulate is that, in terms of linguistic discourse, the notion of collocations preferably needs to be located in the area of general, largely subject-independent language whereas the notion of terms falls into the area of domain-specific sublanguage of a certain subject field. Another finding is that, from a syntactic point of view, collocations surface in different kinds of syntactic expressions whereas terms are mainly confined to noun phrases. However, what will

also become clear in the course of this chapter is that there is indeed a linguistic property – *viz.* the property of *limited modifiability* – which both collocations and terms share and which may be derived from the discussion and insights of the respective research strands (as will be described and assessed in section 2.3). The linguistic frame of reference within which this linguistic property may be located is the collocational layer of Firth's model of language description. This model – although from a historical perspective located in our discussion on collocations in section 2.1 – will help us define our linguistically motivated statistical association measures for both collocation and term extraction. What Firth's lexical-collocational layer is able to capture is the observation that the linguistic property of limited modifiability is manifested differently, i.e. while it surfaces *syntagmatically* in collocations, it is manifested *paradigmatically* in terms.

2.1 Defining Collocations

That words in natural language are neither randomly combined into phrases and sentences nor that they are only constrained by the rules of syntax had been known by linguists for quite some time. Curiously, this basic fact about collocations and, at the same time, their rather diverse and apparently idiosyncratic behavior, has been taken out of focus by a substantial part of contemporary mainstream linguistics which has been primarily concerned with examining language from a theoretical perspective. In particular, generative linguistics in the Chomskyan tradition (Chomsky (1965) or Chomsky (1995)) demote all lexical and syntactic idiosyncracies safely into the realm of the lexicon.[1]

By pointing out that "*You shall know a word by the company it keeps!*", it is Firth (1957), who commonly gets the credit for first intoducing the notion of collocation into contemporary linguistics (see also Bartsch (2004)) and who thus coined probably one of the most well-known slogans in 20th century linguistics. Still, as e.g. also Lehr (1996) points out, already Firth used to be rather vague about a precise definition of the concept, and hence it is not surprising that there has been a rather enormous

[1]It is actually only with the advent of phrase structure grammar theories which also were concerned with aspects of language computability, such as Head-Driven Phrase Structure Grammar (HPSG) (Pollard & Sag, 1994), when collocations again received at least some interest in theoretical linguistics, as can e.g. be witnessed in the work of Krenn (1994)

conceptual diversity surrounding the idea of collocation in linguistic research up to today. Drawing a very rough dividing line, two lines of linguistic research may be identified in the last half-century and we will describe them in some detail in the first two subsections below. On the one hand, there is the *structural-lexicographic approach* which is mainly concerned with adequate representation forms of collocations within linguistic lexicons and dictionaries (subsection 2.1.1). On the other hand, there is the *frequentist corpus-based approach* to collocations which was initiated and significantly influenced by Firth's linguistic research (subsection 2.1.2) and which is dedicated to an empirically grounded analysis of natural language.

As the field of computational linguistics and natural language processing (NLP) is also in need of linguistic definitions, computational linguists – if their research or application task is to extract collocations from unrestricted text – typically acknowledge that there is a wide array of diverse definitions provided by the two lines of linguistic research (subsection 2.1.3). Besides the property of co-occurrence, however, these only have minimal or no influence in how the algorithms and procedures for an extraction task are defined, as we will also see in more detail in section 3.1 in the next chapter. The consequence of this is that, in general, insights about linguistic properties of collocations are not incorporated in computational implementations. Obviously, this constitutes one of the gaps that this thesis aims to fill. For this purpose, we will assemble and assess the linguistic properties of collocations adopted from the various linguistic research strands in subsection 2.1.4. On the one hand, we will focus on four characteristic linguistic properties of collocations which have the capacity to be algorithmically formalized from a computational perspective. On the other hand, we will capitalize on linguistic properties that will help us to draw a linguistic demarcation line between collocations and non-collocations and also to establish different linguistic subtypes of collocations.

2.1.1 Defining Collocations from the Lexicographic Perspective

The kinds of linguists who typically have a profound interest in examining collocations and their linguistic properties are lexicologists and lexicographers. This is of course due to the fact that lexicographers have to worry about how to represent information about collocations in a linguistic dictionary or lexicon. This subsection will therefore

describe two kinds of representative strands of work in this vein. One kind, represented by Hausmann (1985) and Mel'čuk (1995a), places their lexicographic descriptions of collocations into a broader linguistic and meaning-based framework (subsubsection 2.1.1.1) whereas the other kind, represented by Benson et al. (1986b) and Benson (1989), confines itself to more or less "theory-free" lexicographic descriptions (subsubsection 2.1.1.2).

2.1.1.1 The Meaning-based Lexicographic Approach to Collocations

Meaning-based approaches to collocations are characterized by their often close connection to applicative areas such as lexicography and foreign-language learning. Mel'čuk, a prominent lexicographically oriented linguist, has embedded his approach to collocations into a complete linguistic framework, *viz.* Meaning-Text Theory, which attempts to account for relations between lexical items language-independently. Within this framework, Mel'čuk (1995a) and Mel'čuk (1998) attempt to come to terms with the idiosyncrasy of collocations by embedding them into a more semantically oriented layer of description. In the Meaning-Text Theory (MTT) lexical relations are used as a means of describing so-called institutionalized lexical relations. Such relations are defined as holding between two lexical items with a constant meaning linked to their combination. Although these meanings, referred to as Lexical Functions, explain the relations between lexical items mostly on the semantic level, phonological and syntactic descriptions are not excluded per se.

Lexical Functions (LFs) aim at coping with the problem of lexical choices. For Mel'čuk, this boils down to go from a given semantic representation to a corresponding (deep) syntactic representation. In this process, the speaker has to select lexical units, i.e. lexical lexemes and phrasemes to build sentences.[2] Although LFs are taken as a particular device to systematically describe the relations between two lexical units across various languages, they are far more encompassing than the notion of collocations. A composite formulaic notation[3] including phonological, syntactic and semantic features is given to cover various syntagmatic relations between lexical items. Thereby, it is assumed that all languages, in different ways, realize the meanings postulated by LFs and that the main difference lies in the language-specific ways in

[2]See Wanner (1996) and Bartsch (2004) for a detailed description on the aspects of lexical choices.

[3]Mel'čuk (1995a) actually parallels them to mathematical functions represented by the following standard expression $f(x) = y$.

which the combination of given lexical items is used to arrive at various LF meanings.

There are 36 syntagmatic LFs which are distinguished by their syntactic part of speech. Mel'čuk (1996) provides some examples and their English realizations:

Verbal Lexical Functions:

1. Degrad [Lat. degradare (to degrade, worsen)]
   a. Degrad(clothes) = to wear off
   b. Degrad(house) = to become dilapidated
   c. Degrad(temper) = to fray

Adjectival Lexical Functions:

2. Magn [Lat. magnus (big, great)]
   a. Magn(belief) = staunch
   b. Magn(thin[person]) = as a rake
3. Bon [Lat. bonus (good)]
   a. Bon(aid) = valuable
   b. Bon(proposal) = tempting

Nominal Lexical Functions:

4. Centr [Lat. centrum (the center/culmination of)]
   a. Centr(crisis) = the peak (of the crisis)
   b. Centr(desert) = the heart (of the desert)

As can be seen from the above examples, the semantic radar of LFs is far more extensive and comprehensive than just natural language expressions that are typically assumed to fall under the notion of collocation.[4] In particular, Mel'čuk's MTT is

[4]It has been argued (Bartsch, 2004) that Mel'čuk's set of formulaic descriptions of syntagmatic relations between lexical units may be beneficial for translating collocations from one language to another because they generally apply across languages in which such relations are realized by different lexical elements.

aimed at providing a complete linguistic framework for the mapping from the content or meaning of an utterance to its form or text, with collocations being one particular (i.e., idiosyncratic) lexical surface realization. The overall lexicographic goal of MTT is the creation of so-called Explanatory Combinatorial Dictionaries (ECDs) (cf. (Bartsch, 2004)) displaying the combinatorial properties of word combinations in a language.

In the area of German linguistics, research on collocations is founded on a completely different conceptualization, i.e. one derived from a phraseological-semantic[5] point of view. In particular Hausmann (1985) and Burger (2003), besides focusing on prescriptive correctness of collocational language use, categorize collocations according to the semantic specificity of their constituents. Thus, content words (i.e. verbs, adjectives, nouns) play a central role as components of collocations. The different constituents in a collocation do not have an equal status, but rather, their relationship is a directed one. The *collocational base* is defined as the dominant constituent while the *collocate* is dominated by the base. In particular, the base is the semantically autonomous part, which, however, needs the collocate to obtain its full meaning. This is illustrated in the following preposition-noun-verb (PNV)[6] collocations from German (and their English translations):

5. a. "*zur Verfügung stellen*" (to make available)

   b. "*in Erwägung ziehen*" (to take into consideration)

Here, the collocational base "*Verfügung*" (availability) is completed by the meaning of the collocate "*stellen*" (to place) and, in the English translation, the meaning of the collocational base "*available*" is completed by the meaning of the collocate "*make*". Central to Hausmann's definition of collocations is the directionality from the base to the collocate in that the base as the dominant constituent is the element which is semantically more stable and which thus exerts a stronger influence in a way that it can dominate the collocate.[7] Hence, collocations consist of at least two component parts, with at least one component part either having kept or lost its literal meaning.

[5]The technical term *phraseologism* appears to have been coined by this line of collocational research to set it apart from the Firthian approach.

[6]All examples are taken from the German-language newspaper text corpus collected to run the experiments for the automatic extraction of PNV collocations as described in subsection 4.5.2.

[7]Hausmann's definitions have been criticized for being too narrow by Bartsch (2004).

Another central distinction in Hausmann's conception of collocations is concerned with the degree of fixedness between the different constituents of a collocational expression. On the one hand, there are fixed word combinations under which mainly idioms can be found and for which the above definition for base and collocate hardly applies. These fixed expressions are referred to as *fully idiomatic expressions* in which every component is void of its literal meaning, as is exemplified by the following idiomatic expressions:

6. a. "*ins Gras beißen*" (literal: to bite the grass; actual: to bite the dust, die)

   b. "*auf der Hand liegen*" (literal: to lie on the hand; actual: to be obvious)

In contrast, less fixed *partly idiomatic expressions* (teilidiomatisierte Wendungen) are expressions in which some component part, typically the base in Hausmann's conception, still keeps its literal meaning, such as the nouns "*Druck*" (pressure) and "*Geltung*" (importance) in the following examples:

7. a. "*unter Druck geraten*" (to get under pressure)

   b. "*zur Geltung kommen*" (to become important)

These are also the types of expressions that Hausmann (1985) refers to as *collocations*. On the other end of the continuum there are free word combinations where all components keep their literal meaning, thus making the expression fully compositional:

8. a. "*auf einen Baum klettern*" (to climb up a tree)

   b. "*an die Zukunft glauben*" (to believe in the future)

In the linguistic classification task to derive a gold standard for German preposition-noun-verb (PNV) collocations (as described in subsubsection 4.5.2.3), a distinction along these lines has turned out to be quite operational for the human classification of collocation candidates. We will return to this issue in subsection 2.1.4 below in which we assemble our adopted linguistic properties of collocations.

2.1.1.2 Other Lexicographic Accounts of Collocations

This subsubsection reviews two accounts of collocations which may be mainly described as applicational as they are primarily concerned with collecting and representing collocational entries in a lexicon or dictionary. Whereas the first account (Benson et al., 1986b) still offers some theoretical underpinnings, the second one (Dudenredaktion, 2002) is purely applicative in nature but needs to be discussed as it consitutes the only such type of work for the German language.

The first dedicated and large-scale lexicographic study of collocations was undertaken for the English language by Benson et al. (1986b), Benson (1989) and Benson (1990), which led to the publication of the *BBI Combinatory Dictionary of English: A Guide to Word Combinations* (in short: BBI) (Benson et al., 1986a).[8] Benson et al. (1986a) outline the motivation for a dictionary of word combinations and the kinds of information included in it.[9] The goal is to provide information on the general combinatorial possibilities of an entry word. Various types of combinatorial preferences are listed, such as e.g. whether there are any combinatorial preferences of verbs for nouns (e.g. "*[to adopt, enact, apply] a regulation*") or what the possible adverbial combinations (i.e. modifications) of a verb are (e.g. "*to regret [deeply, very much]*".

These combinatorial preferences are classified into two types of collocations, i.e., grammatical collocations and lexical collocations. Grammatical collocations are phrases consisting of a dominant word (e.g. noun, adjective, verb) and a preposition or grammatical structure such as an infinitive or a clause, as exemplified by the following expressions:

9. a. "*account for*"

   b. "*adjacent to*"

   c. "*dependent on*"

   d. "*the fact that + clause*"

[8]The current edition of this dictionary is Benson et al. (1997).

[9]From the viewpoint of embedding the BBI into a linguistic framework, it has to be noted that Benson et al. (1986b), Benson et al. (1986a) and Benson et al. (1986a) make references to Mel'čuk's Meaning-Text Theory.

Lexical collocations, on the other hand, are classified by the BBI approach according to their part-of-speech patterns, such as verb-(preposition)-noun, adjective-noun or noun-noun, as exemplified by the following expressions:

10. a. "*compose music*"
"*launch a missile*"
"*set an alarm*" (verb-noun pattern)

b. "*strong tea*"
"*chronic alcoholic*" (adjective-noun pattern)

c. "*a swarm of bees*"
"*a flock of sheep*" (noun-noun pattern)

Although some of these expressions describe lexically determined co-occurrences and thus are more in line with what is commonly understood as collocation, it can be seen that others again are fairly compositional from a semantic perspective in that all constituents still keep their literal meaning and thus probably would not be labelled "collocation" by approaches such as Hausmann's (outlined in subsubsection 2.1.1.1 above). A look at the intended audience, however, explains the extensiveness of the BBI approach to word combinations and collocations since Benson et al. (1997) explicitly target their dictionary towards foreign-language learners. Still, the BBI dictionary is the most comprehensive lexicographic resource of word combinations in any language to date and thus deserves attention.

As far as dictionaries and lexicographic resources for German-language[10] collocations and idiomatic expressions are concerned, Volume 11 of the *Duden* series (Dudenredaktion, 2002) may be regarded as the main representative. This dictionary, however, differs from the BBI dictionary in several respects. As already the title "*Redewendungen*" (figures of speech, sayings) suggests, the focus of this volume is rather on idiomatic speech figures than on the allowable and preferred combinatorial properties of words. Hence, each entry in the dictionary is accompanied by etymological information rather than by lexico-grammatical one. Still, in practice this dictionary has a broader definition of what is considered to fall under the notion of "*Redewendungen*". In their introductory remarks, for example, the *Duden* editorial staff points out that, besides idiomatic expressions, they also count *Funktionsverbgefüge* (support

[10]The language under investigation for collocations in this thesis.

verb constructions) to the class of collocations. As will be seen later on, these types of syntactic constructions will play a prominent role in the set of collocational candidates in our experimental study described in subsection 4.5.2. In particular, we will focus on their surface realization as preposition-noun-verb (PNV) constructions.

2.1.2 Defining Collocations from the Frequentist Perspective

The notion of *collocation* in its original meaning is almost inseparably tied to the linguistic tradition of British contextualism and its founder, John R. Firth. But as was already hinted at above, Firth does not only have to be credited for having drawn the attention to the concept collocation in linguistics but his work also laid the groundwork for the *frequentist* or *empiricist* tradition of British (corpus) linguistics with its main representatives Michael A. K. Halliday and John Sinclair (subsubsection 2.1.2.3). The central notion in their research, in extension to Firth, was that the empirical, even statistical, side of language use in text corpora could serve as a framework to describe and explain natural language.[11] Indeed many of the roots of the empirically motivated and statistical methodology in contemporary computational linguistics may be sought in this linguistic tradition.[12] In particular, the notion of *co-occurrence*, which runs like a thread through the corpus linguistics tradition, has come to be a defining property in almost all applications to collocation extraction in computational linguistics.

But first we will lay out Firth's model of language description and, in particular, its lexical-collocational layer (subsubsections 2.1.2.1 and 2.1.2.2), as it will play a central role in providing a suitable linguistic frame of reference for the linguistically motivated statistical association methods presented in this thesis – not only for the extraction of general-language collocations but also of domain-specific terms.

[11] This focus on actual empirical language use is in stark contrast to the structuralist and Chomskyan generative tradition in linguistics which introspectively relies on so-called "grammaticality judgments" of language speakers – mostly the researcher himself – in order to describe and explain linguistic constraints.

[12] This can also be seen in various accounts on contemporary statistical NLP (Manning & Schütze, 1999, p.6)

2.1.2.1 Firth's Model of Language Description

Firth's model of language description crucially relies on the notion of *linguistic context*. For Firth (1957), this meant a frame of reference for isolated words or sentences.[13] Thereby, linguistic context was divided into four descriptive layers, each of which was founded on the same textual basis, *viz.* one or more situationally dependent texts:

1. The phonetic layer examines the relationship of single phones to other phones or phonetic sequences.

2. The morphological layer examines the relationship of single morphemes to other morphemes or morpheme sequences.

3. The syntactic layer examines the relationship of grammatical classes to each other. Grammatical classes are derived from text words by means of an empirically obtained inventory of grammar rules.[14]

4. The lexical layer examines words in relationship to other words or word sequences.

The *semantic level*[15] of Firth's model of language description is actually located within the *situative context* (hence not the linguistic context), whereby situative context refers to the context of textual production. On all four layers, there are two contextually descriptive axes, the syntagmatic axis and the paradigmatic axis (*syntagmatic context or structure* and *paradigmatic context or system* (see Firth (1968)).

For example, on the lexical level the syntagmatic *structure* of a text results from the sequence of subsequent words, whereas the paradigmatic *system* is obtained by empirically determined substitutional classes. This principle of structural and systemic contexts on various language layers is referred to as contextualization in Neo-Firthian-style linguistics. The following section will outline in detail the relevance of the syntagmatic and the paradigmatic axes on the lexical layer of Firth's model as they will play a central role for collocations and terms, respectively.

[13] See Lehr (1996) for an extensive overview of British contextualism and the Firthian approach to linguistics in particular.

[14] Firth (1968) refers to them as *collogations of generalized categories*; see also Lehr (1996).

[15] In contextualism, the notion of *meaning* is equalled with *function in a context* and, hence, is present on all levels of Firth's model of language description.

2.1.2.2 The Collocational Layer of Firth's Model

Firth (1957) renames the *lexical layer* of his model of language description with the term *collocational layer*, which illustrates the prominence that collocations take in his model of language description. Lehr (1996) describes this lexical-collocational layer in great detail, from which an apt graphical representation may be derived in figure 2.1.

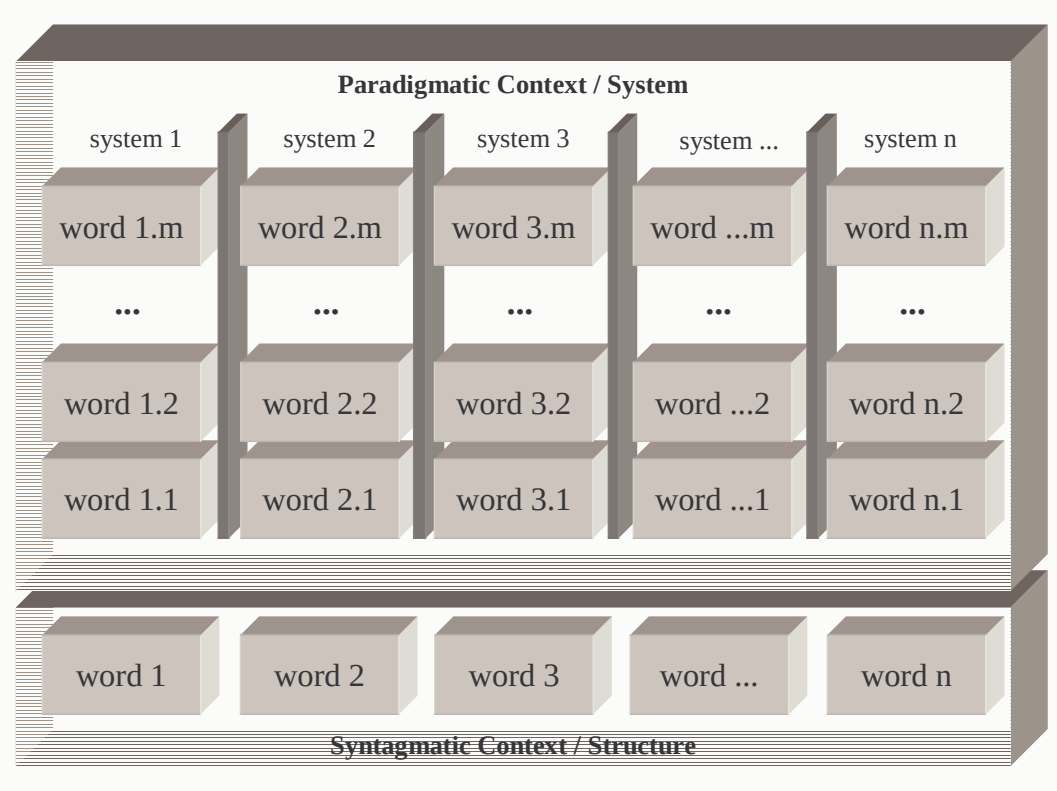

Figure 2.1: The lexical-collocational layer of Firth's model of language description.

With *text* being the central notion of language utterance in Firth's model, the words within a syntagmatic context (*structure words*), $word_1$ to $word_n$, constitute elements of a concrete textual structure. Those words which are elements in the paradigmatic context (*system words*), $system_1$ through $system_n$, only have virtual character in that they do not appear in the current text but can be empirically made accessible from other texts.

Collocations are occurrences of words in the syntagmatic context which are constituted of two or more structure words. How the boundaries of collocations within a text are determined remains unclear in Firth (1957) (see also Lehr (1996)). On the paradigmatic axis, system words may appear in place of the structure words of the current text in that they function as potential substitutes. It is important to note that this can only occur within the predefined substitutional frame of a system.

While Firth's approach to describe collocations on a contextualized lexical level has been further refined by his successors in British contextualism (see next subsubsection 2.1.2.3),[16] the notions of syntagmatic context and paradigmatic context remain foundational to his approach. By coupling these notions with the linguistic property of *limited modifiability* and by putting it on a quantifiable basis, we will introduce new approaches to the linguistic design of statistical association measures for the automatic extraction of collocations and of terms from unrestricted text (see sections 4.3 and 4.3). Thereby, it should be noted that although the concept of *collocation* obviously plays a central role in Firth's linguistic conceptualization, the concept of *term* or *technical term* is not mentioned. This, however, is mainly due to historic reasons as in traditional linguistics the difference between these two notions is not very clear-cut and very often they were lumped together under the common heading *collocation*. Only with the growing importance of and the concurrent linguistic research interest in domain-specific (sub)language use (see subsection 2.2.6) has the notion of *technical terminology* gained its place next to the notion of *collocation*.[17]

2.1.2.3 Neo-Firthian Developments: Halliday and Sinclair

Firth's formulation of the collocational-lexical level of his model of language description was, to some respect, incomplete. At least concerning the methodological angles of concrete language analysis, his descriptions are more intuitive than systematic. It was up to his contextualist successors to form a coherent and methodologically sound model of analysis for the lexical level of language description. In particular Halliday (1966) and Sinclair (1966) elaborated on Firth's (1957) thought of *meaning by collo-*

[16] At the same time, it has also been vigorously disputed by linguists who were working within the generative paradigm at the time (e.g. by Langendoen (1968)).

[17] As will be seen in the next subsection 2.1.3 in contemporary computational linguistics, the linguistic differences are often ignored or, as in Manning & Schütze (1999), *terminological expressions* are described as a subclass of collocations.

cation on the lexical level by introducing the notion that patterns of collocation can form the basis for a lexical analysis of language and are alternative to, and independent of, the grammatical analysis. These two levels of analysis are regarded as being complementary, with neither of the two being subsumed by the other.

In parallel to phraseological perspectives on collocations (cf. subsubsection 2.1.1.1), contextualist linguists (e.g. Halliday (1966)) also advance the observation that the different constituents of collocations do not have an equal (i.e., unstructured) status but display a hierarchical structuring, which they refer to as *nodal item* or the *collocant* (i.e., the collocational base in Hausmann's (1985) terminology[18]) and the *collocate*. The node and the collocate are in a directed relationship (*node* → *collocate*, i.e. the node collocates with the collocate but not vice versa), with the collocate further specifiying the meaning of the node. Recognizing the necessity to extract such structures from text to make them quantifiable in the first place, post-Firthian linguists also attempted to specify a procedure to determine the distance between a node and and its potential collocate (the collocational span), in order to be able to locate the latter one.

> Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c ... (Halliday, 1969, p. 276)

This passage illustrates two interesting points. First, it underlines the basic assumption held by British contextualists that natural language exhibits empirical and quantifiable properties, which puts them into opposition to the mainstream generative and structuralist linguists at that time. Second, as the necessary linguistic machinery to adequately compute such a collocational span from unrestricted natural language text was basically missing at that time, it led Sinclair (1966, p. 415), one of the early proponents of a corpus-based approach to linguistics, to the following assessment:[19]

[18]It should be noted, however, that, other than with respect to the internal structure of collocations, Hausmann's (1985) and other phraseologists' normative approach to collocations was developed in explicit contradistinction to the frequentist and empirical approach taken by the British contextualists.

[19]See also Lehr (1996) for a similar assessment.

> The extent of the span is at present arbitrary, and depends mainly on practical considerations; at a late stage in the study we will be able to fix the span at the optimum value, but we start with little more than a guess.

Perhaps the most influential effect of post-Firthian British linguists on collocation (and term) extraction research in computational linguistics was their observation that the constituents of collocations follow the basic pattern of *co-occurrence*.

> This tendency to co-occurrence is the basic formal pattern into which lexical items enter. It is known as "collocation", and an item is said to collocate with another item or items. (Halliday et al., 1965, p. 33)

In this respect, it should be mentioned, however, that already Firth took notice of the pattern of co-occurrence which is reflected, to some extent, in his *recurrence criterion* (Firth, 1957).

Being the most recent one in the line of Neo-Firthian linguistic research, it is finally Sinclair (1991) who grounds these ideas into the notion of *co-occurrence*-based corpus analysis and states that evidence from large corpora suggests that grammatical generalizations do not rest on a rigid foundation, but are the accumulation of the patterns of hundreds of individual words and phrases. Two principles are proposed in order to explain the way in which meaning arises from language text. The grammatical level is represented by the the so-called *open-choice principle*, which sees language text as the result of a very large number of complex choices, with the only constraint being grammaticality. The *idiom principle* represents the lexical level and accounts for the constraints that are not captured by the open-choice model – with collocations being part of the idiom principle.

Up until today, the notion of co-occurrence runs like a thread through the corpus and computational linguistics literature on collocations (see e.g. (Manning & Schütze, 1999, p.153–157)) and can be said to be one of the defining quantifiable linguistic properties of collocations (see subsubsection 2.1.4.1 below).

2.1.3 Defining Collocations from a Computational Linguistics Perspective

The various approaches, previously described, to define and pinpoint collocations from a linguistic perspective had a mixed impact on the research on automatic procedures to

extract collocations from machine-readable natural language text. Early approaches, such as Berry-Rogghe (1973) (see subsection 3.1.1 for an extensive discussion of this approach), adhered quite closely to the theoretical specifications on collocations which linguistic research had come up with and, as a consequence, attempted to examine linguistic theories, or even aimed at developing them further. More current approaches, while still describing and emphasizing the groundwork done by linguists, are more driven by the requirements of computability and applicability and, hence, the notion of collocation is defined and used in a much broader and practical sense than in linguistics.

In their widely used and acclaimed textbook on statistical NLP, Manning & Schütze (1999) dedicate a complete chapter (chapter 5) to the topic of collocation extraction from text corpora. Without doubt, this prominence reflects the fact that recognizing collocations in a natural language processing pipeline is an important processing step, which – ideally – should be situated somewhere before the semantic module (cf. subsection 3.1.3 for more discussion of this point). On the other hand, however, due to the empirical turnaround of the 1990s in computational linguistics, collocations, due to their frequentist properties framed by the British contextualists, have also turned out to be an ideal linguistic construction to apply and adapt common statistical machinery and measures to problems in natural language processing.

Nonetheless, current approaches to collocation extraction in computational linguistics also need, at least, a working definition of their notion of collocation. For quite a few researchers, such a definition turns out to be rather operational, such as in Choueka (1988):

> A collocation is defined as a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components.

The definition has two parts, with the first part describing the presumed surface token representation of collocations in natural language text and the second part stating a semantic property with respect to the component parts of a collocation. The first part of this definition on the surface token representation is pertinent to Choueka's (1988) method for identifying potential collocations in text[20] while the second, more

[20]The application setting of this work is information retrieval.

linguistically tuned part merely serves as an illustrating point.

This kind of procedure is quite characteristic for a lot of research on collocation extraction in NLP. Typically, both linguistic definitions and linguistic properties of collocations are laid out, or, at least, referred to. For example, Manning & Schütze (1999) dedicate around six pages to defining the notion of collocation from a linguistic viewpoint and reference foundational linguistic research, such as Benson (1989), in working out three essential linguistic properties of collocations, *viz. non-compositionality*, *non-substitutability*, and *non-modifiability*.[21] However, this is not reflected in the algorithmic methods or lexical association measures (i.e., their computational implementation) of a corresponding collocation extraction procedure. The extraction machinery presented is typically rather unrelated to the linguistic properties outlined, with one exception, though, *viz.* the notion of *co-occurrence*. As we already described in the previous subsubsection 2.1.2.3, co-occurrence is taken to be a defining linguistic property of collocations, at least for the British contextualist linguists. And in fact, *frequency* of co-occurrence, as it turns out, actually plays an important role for almost all of the standard statistical and information-theoretic asssociation measures employed for collocation extraction.[22]

In this respect, it is illuminating to elaborate on the kinds of subclasses that Manning & Schütze (1999) actually consider to be collocations, the vast majority of which are in line with much of the NLP research on collocations but some of which would certainly be controversial in the linguistics research community.

- **Idioms** are defined as *frozen expressions* in which "there is just one way of saying things and any deviation will completely change the meaning of what is said" (Manning & Schütze, 1999, p. 186)

- **Support Verb Constructions** are characterized by the little semantic content their *light verbs* have, such as in "*make a decision*", "*do a favor*", in which there is hardly any meaning on its own in the verbs "*do*" and "*make*".

- **Phrasal Verbs**: Such verbal constructions (e.g. "*tell off*" or "*go down*") are an important part of the lexicon in English and consist of a combination of main verb and particle. These verbs often correspond to single lexemes in other languages.

[21] These will be explained in detail in the next subsection.

[22] These association measures will be described in depth in section 3.3.

- **Terminological Expressions**: For Manning & Schütze (1999) these are phrases which refer to concepts and objects in technical domains. It is noted that such expressions are often fairly compositional (contrary to general-language collocations), such as in the case of "*hydraulic oil filter*", but it is noted to be important that they be treated consistently throughout technical texts.

The first three collocational subclasses would probably be uncontroversial among linguists or lexicographers.[23] However, Manning & Schütze (1999)'s fourth collocational subclass, *terminological expressions* (or short *technical term* or *terms*), would most probably not be regarded as a collocation by linguists. On the one hand, they would not fall under the definitional status postulated by the meaning-oriented or lexicographic approaches to collocations, and certainly not by their phraseological representatives (see subsection 2.1.1 above). In the case of contextualist approaches to collocations, the situation may be less clear. As laid out in the previous subsection 2.1.2, Neo-Firthian linguistic approaches (e.g. (Halliday, 1966) or (Sinclair, 1966)) attempted to flesh out Firth's model of language description concerning the methodological angles of concrete language analysis and in this respect, the notion of collocation plays a major role although rather from a general-language perspective. Still, Firth's original basic notion of the lexical (or collocational) level of language does of course not exclude technical terminology per se.[24]

Nevertheless, the task of finding and compiling terminological expressions from specialized technical domains is a comparatively new need which has been put to the foreground, both in linguistics and even more so in computational linguistics more recently and which has arisen due to the increasing amount of textual databases in specialized technical domains. Hence, it is not astonishing that this issue was not pressing at the time when most linguistic definitions on collocations were formulated. At the same time, however, it is not surprising either that computational linguists define and use the notion of collocations in a much broader and more practical sense in that certain types of natural language expressions, such as technical terms, which are challenging from an application perspective, are included because the automatic association measures and extraction methods developed for general-language collocations

[23] Although *phrasal verbs* would probably be considered to be a kind of collocational expression peculiar to the English language.

[24] Indeed, Firth does not exclude any kind of natural language expression.

have turned out to be applicable to them as well. Thus, Manning & Schütze (1999, p. 152) even go so far to consider technical terms as a special case of collocations:

> There is a considerable overlap between the concept of *collocation* and notions like *term*, *technical term*, and *terminological phrase*. As these names suggest, the latter three are commonly used when collocations are extracted from technical domains (in a process called *terminology extraction*).

Although this statement would most probably not be subscribed to by linguists working on collocations or by terminologists in the theoretical vein of terminology research (see section 2.2), it illustrates that, for the tasks of automatically finding both collocations and terms in a text corpus, for computational linguists these two types of natural language expressions appear to show statistical and distributional similarities which warrant the use of similar or even common methods and association measures[25] – with frequency of co-occurrence being the most salient one.

2.1.4 Linguistic Properties of Collocations Adopted

As we witnessed in the previous two subsections, collocations are not easily defined. We showed that this is reflected by the great variety of definitional approaches that were developed in linguistic research. In the following, we will synthesize two essential perspectives on collocations which have been laid out by linguistics research and, because they exhibit formalizable and partly even quantifiable linguistic features and observations, were picked up by computational linguistics research on collocation extraction. On the one hand, these are concerned with four basic characteristic linguistic properties of collocations, and, on the other hand, with linguistic criteria that draw the demarcation line between collocational expressions and non-collocational expressions.

[25] McKeown & Radev (2000, p. 527) make a similar point by stating that "by applying the same algorithm to different domain-specific corpora, collocations specific to a particular sublanguage can be identified."

2.1.4.1 Four Basic Characteristic Properties

With respect to basic linguistic properties of collocations, it has already been noted on several occasions that, while most computational linguistics research references and describes such properties, they do not necessarily flow into the methodological considerations when designing collocation extraction algorithms. Indeed, only the first one of the basic linguistic properties below, (frequency of) lexical co-occurrence, has been a widely used one, if not the most influential one in this respect. The other three, non-compositionality, non-substitutability, and non-modifiability, while having been highlighted by Manning & Schütze (1999),[26] have had limited or no influence on the design of collocation extraction algorithms.[27]

Lexical Co-occurrence. As described in subsection 2.1.2, a recurrent observation in Firthian and Neo-Firthian linguistics was that collocations follow the basic patterns of *co-occurrence* with respect to their constituent parts. Due to its inherently quantitative and empirically verifiable nature, this property exerted a great influence on many of the proposed methods for collocation extraction (see section 3.3 below for a detailed account), especially after the empirical turn in computational linguistics. In a way, it is even safe to say that many of these methods are basically variations on the common theme of co-occurrence.

Non- or limited compositionality. One of the fundamental principles of semantic theory is the principle of compositionality, which states that the meaning of a natural language expression is a function of the meaning of its parts.[28] For collocations, however, the meaning is not a straightforward composition of

[26] It should be noted that Manning & Schütze (1999) highlighted these three properties to the computational linguistics research community; in linguistics research on collocations, of course, they are taken for granted (cf. (Benson, 1989))

[27] As will be discussed in detail in subsection 3.1.3, although the collocational procedures devised by Lin (1999) and Lin (1998b) make use of the properties of non-compositionality and non-substitutability, these methods are not applied to separate collocations from non-collocations but rather to fine-classify an already acquired set of collocations in order to identify the idiomatic ones.

[28] In semantic theory (cf. (Cann, 1993)) this principle, sometimes referred to as Fregean Principle of Compositionality, accounts for the fact how the lexical meanings of individual words contribute to the overall meaning of a phrase or a sentence, i.e. more generally speaking, how the meanings of smaller expressions contribute to the meanings of larger ones that contain them. The notion of "function" is essentially an operation that derives a single result given a specified input.

its parts. The collocational subtype of idioms is at the extreme end of this property (see subsubsection 2.1.1.1 above) in that the meaning is completely different from its (usually also existing) meaning as a free word combination. This property will serve as an adequate criterion to draw the demarcation line between collocations and free word combinations and will be further illustrated in the respective subsubsection 2.1.4.2 below.

Non- or limited substitutability. The components of a collocation cannot be substituted by other words, neither syntactically nor semantically, and keep their collocational meaning. This is even the case if a substitute word has the same part of speech and a similar meaning in that context, as shown in the following examples from German preposition-noun-verb combinations:[29]

11. a. "*im Raum stehen*" (to be unsolved or undone)
    b. *"*im Raum posieren*"
    c. *"*im Zimmer stehen*"

In this example, first the verbal constituent "*stehen*" (to stand) and then the nominal constituent "*Raum*" (room) have been replaced by syntactically equal (in terms of part of speech) and semantically similar words (with the verb "*to posture*" and the noun "*chamber*", respectively). As a consequence, the expression adopts a rather non-sensical meaning and looses its status as a collocation. As will be described in subsection 3.1.3, although it requires a wide-coverage and resource-intensive thesaurus, this property has been computationally implemented by Lin (1999) to fine-classify an already existing set of collocations into compositional and non-compositional ones.

Non- or limited modifiability. This property describes the syntagmatic effect that many collocations cannot be modified freely with additional lexical material or through other kinds of grammatical transformations. Very often, linguistic research (Benson, 1989)[30] notes, at least intuitively, that this is particularly the case for idiomatic expressions, such as the following ones:

[29] Following standard notational practice in linguistics research, ill-formed or odd expressions are marked with an asterisk (*).

[30] This is also expressed by Manning & Schütze (1999).

12. a. "*jmdn auf die Schippe nehmen*" (to lampoon somebody)
    b. *"*jmdn auf die alte Schippe nehmen*"
    c. *"*jmdn auf die hölzerne Schippe nehmen*"

As can be seen, the nominal constituent of this collocation, "*Schippe*" (shovel), cannot be modified by additional lexical material, such as the modifying adjective "*alt*" (old), at least not without completely loosing its collocational meaning and adopting a completely different, rather odd one.[31] Concerning other subtypes of collocations, e.g. support verb constructions (cf. subsubsection 2.1.4.2 below), standard linguistic testing seems to allow for some modification:

13. a. "*jmdn zur Verantwortung ziehen*" (to make s.o. responsible)
    b. "*jmdn zur politischen Verantwortung ziehen*" (to make s.o. politically responsible)
    c. "*jmdn zur alleinigen Verantwortung ziehen*" (to make s.o. solely responsible)

Given these examples, it looks, superficially at least, as if the property of non- or limited modifiability does not equally hold for support verb constructions. Whether or not all these observations with respect to the property of non- or limited modifiability can also be empirically verified will be discussed extensively in subsection 5.1.3 below. Crucially, however, in section 4.2, we will see how this property of non- or limited modifiability may be related to the lexical level of the Firthian model of language description, in particular its view on syntagmatic context in which the constituent words of a collocation are located[32] and thus how this may serve as the linguistic basis in defining our linguistically motivated statistical association measure for collocation extraction (see section 4.3).

[31] This linguistic judgment, of course, is derived in a rather intuitive and introspective way, which is the standard methodology of mainstream (i.e., structural and generative) linguistics. Neo-Firthian linguistics, on the other hand, would also ask whether such a judgment can be empirically verified.

[32] This has been explained in detail in in subsubsection 2.1.2.2 and depicted in figure 2.1 above.

2.1.4.2 Demarcation Line: Collocations vs. Free Word Combinations

Another linguistic perspective on collocations, which has been picked up by computational linguistics research, concerns the demarcation line between collocations and their various subtypes, on the one hand, and so-called free word combinations, on the other hand. There are various linguistic possibilities on how and where to draw this line[33] but the most common and accepted way is to do this on the semantic layer, i.e. with respect to the compositionality (cf. the previous subsubsection 2.1.4.1) between the component parts of a natural language expression. Naturally, since the demarcation line is located on the semantic level of natural language, this linguistic perspective on collocations has mainly been shaped by the meaning-based account of collocations described in subsection 2.1.1.1. From these sources, three major subtypes of collocations can be derived, *viz. idiomatic phrases*, *support verb constructions* or *narrow phrases*, and *fixed phrases*, all with varying degrees of and contributions to semantic compositionality between their lexical constituent parts. They are all different from so-called *free word combinations* in which every component part fully contributes to the overall meaning of the expression, which makes them fully compositional from a semantic perspective. As we will also see in subsubsection 4.5.2.3, these kinds of semantic criteria are helpful for linguistically informed human judges to classify natural language expressions as collocational or non-collocational and thus to construct a gold-standard test data set to evaluate the quality of collocation extraction methods.

Idiomatic Phrases. The semantically most intransparent subtype of collocations, in terms of the constituent parts contributing to the overall meaning, are *idiomatic phrases* or *idioms*. In their case, *none* of the lexical components involved contribute to the overall meaning in a semantically transparent way, which makes the meaning of the expression metaphorical or figurative.[34] Some examples of these have already been given above (in subsubsections 2.1.1.1 and 2.1.4.1). For example, the literal meaning of the German preposition-noun-verb combination "*[jemanden] auf die Schippe nehmen*" is "'*to take [someone] onto the shovel*", whereas its completely intransparent figurative meaning is "'*to lampoon some-*

[33] McKeown & Radev (2000) give an overview of different approaches.

[34] In Hausmann's (1985) conception of collocations this is referred to as the degree of fixedness between the different constituents of a collocational expressions.

body". Some further adjective-noun and verb-noun idioms from the English language are given below:

14. a. "*red tape*"
    b. "*to kick the bucket*"
    c. "*to bite the dust*"

Support Verb Constructions / Narrow Collocations. The second subtype of collocations, *support verb constructions*,[35] contains expressions which are partly compositional in that *at least one* component contributes to the overall meaning in a semantically transparent way and thus constitutes its semantic core. For example, in the support verb construction "*zur Verfügung stellen*", in which the literal meaning is "*to put to availability*" and the actual collocational meaning is "*to make available*", the noun "*Verfügung*" (availability) is the semantic core of the expression, while the verb only has a support function with some impact on argument structure, causativity or lexical aspect. Besides the German examples in subsubsections 2.1.1.1 and 2.1.4.1 above, some more verb-preposition-noun and verb-noun collocations from the English language are given below:

15. a. "*to put at risk*"
    b. "*to come to an end*"
    c. "*to do a favor*"

There are, however, also (preposition)-noun-verb constructions in which not the noun but the verb is the semantic core contributing to the overall meaning of the collocational expression, as shown in the example below:

16. a. "*aus eigener Tasche bezahlen*"
    b. literal: to pay out of one's own pocket
    c. actual: to pay oneself

In a strict linguistic sense, of course, these expressions are not support verb constructions. Still, because they are also characterized by the fact that only one

[35] The equivalent linguistic term in German is *Funktionsverbgefüge*.

component part, which happens to be the verb in this case, contributes to the overall meaning in a semantically transparent way, these so-called *narrow collocations* (McKeown & Radev, 2000) are put in the same collocational subtype as true support verb constructions.[36]

Fixed Phrases. Very often, so-called *fixed phrases* which are right at the border to free word expressions, are adduced as a third subtype of collocations in linguistic treatments of collocations (Benson, 1989).[37] In their case, all basic lexical meanings of the components involved contribute to the overall meaning in a semantically quite transparent way. There are two basic patterns of compositionality involved here. They are either not as completely compositional as to classify them as free word combinations (as in example 17a) or, although compositional, their combination is very fixed and retracted (as in example 17b).

17. a. "*im Koma liegen*" (literal: to lie in coma; actual: to be comatose)
    b. "*Zähne putzen*" (to clean teeth)

Although in example 17a all the basic lexical meanings of the different lexical components somehow contribute to the overall meaning of the expression, this contribution is not as compositionally complete as in the case of free word combinations. Example 17b would most probably be regarded as collocational by linguists subscribing to the Neo-Firthian tradition because this expression satisfies one of its essential requirements to classify a collocation, *viz.* the tendency of the lexical component parts to co-occur (see the previous subsubsection 2.1.4.1).

Free Word Combinations. Outside the three previously described collocational subtypes there are free word combinations, which are characterized by completely adhering to the principle of compositionality in that the meaning of every natural language expression is a function of the meaning of its component parts.[38] Cowie (1981) defines a free word combination as a maximally variable type of composite unit which is characterized by the openness of combinability

[36] What also unifies these two types is the fact that they both function as predicates.

[37] This subtype is controversial among linguists in that, for example, phraseologists such as Hausmann (1985), would not count fixed phrases to the linguistic class of collocation, due to their quasi-compositional nature.

[38] The complete meaning of a linguistic expression is of course not solely dependent on the meanings

of each element in relation to the other or others. For example, in a verb-noun combination such as "*to manage a business*", nouns are freely combinable with the verb "*to manage* and, vice versa, verbs a freely combinable with the noun "*business*". In particular, as shown in the following example from German, this may also be tested by means of the property of substitutability (see subsubsection 2.1.4.1) above):

18. a. "*auf einen Baum klettern*" (to climb a tree)
    b. "*auf einen Baum steigen*" (to climb up a tree)
    c. "*auf eine Tanne klettern*" (to climb a fir tree)

As can be seen, in this case of the free word (noun-verb) combination "*auf einen Baum klettern*" (to climb a tree), it is very well possible to replace both the verb and the noun by a semantically similar item.

2.2 Defining Terms

Terms are pervasive in the document collections of scientific and technical domains. Their identification is a vital issue for any application dealing with the analysis, understanding, generation, or translation of such documents. As pointed out by Jacquemin & Bourigault (2003), this need arises, in particular, because of the ever-growing mass of specialized documentation on the world wide web, in industrial and government archives and document collections or in various digital libraries, to name just a few of these fast-growing document stores. Hence, the identification and extractions of relevant technical terminology is essential for such purposes as information retrieval, document indexing, translation aids, document routing or summarization.

The definition of what actually constitutes a term, however, substantially differs between computational approaches to term identification, on the one hand, and the classical notion of terminology, as particularly elaborated by Eugen Wüster (as outlined in the following next subsection 2.2.1), on the other hand. This has to do with the fact that the characterization of terms in a computational framework must take into account novel dimensions of termhood in order to be able to tackle application

of its lexical component parts. The syntactic structure (of the component parts) of an expression is also relevant to the derivation of its meaning.

tasks, such as terminology extraction from text corpora. Contrary to that, traditional approaches to defining terminology heavily focus on the conceptual and even philosophical aspects of terms. As pointed out by Sager (1990) and elaborated on by Pearson (1998), the notion of terminology[39] itself is rather polysemous and may be referred to in three different yet related ways as:

1. a theory, i.e., the set of premises, arguments and conclusions required for explaining the relationships between concepts and terms;

2. a vocabulary of a special subject field.

3. the set of practices and methods for the collection, description, and presentation of terms;

According to the first notion, terminology may be a scientific theory, in fact even a discipline in its own right, whose object of investigation is to illuminate the relationship between concepts and terms. As already hinted at, Wüster's General Theory of Terminology (see subsection 2.2.1 below) as well as its contemporary offsprings (see subsection 2.2.2) and other related approaches (subsection 2.2.3) may be seen as the main proponents and will be described accordingly. What they all share is that terms are primarily defined from a conceptual perspective and only marginally, if at all, from a linguistic one. The problems that this strand of research has with respect to the increasing cross-disciplinarity of subject fields and with respect to the requirements for computational approaches to terminology extraction will be discussed in subsection 2.2.4. The second notion of terminology takes a rather pragmatic stance in that it is not tied to a particular theory or framework of terminology but rather views terminology from a quite utilistic perspective as specialized vocabularies for particular subject fields or simply the stock of words associated with a particular discipline (subsection 2.2.5).

The third notion points out that terminology may be used to describe procedures to collect and process terms. This may be done manually by a standardization body in making recommendations for an existing terminology. For example, as a result of Wüster's strive to establish and standardize the study of terminology in an international setting, the International Organization for Standardization (ISO), as an

[39]Several aspects of the study of terminology discussed in the following subsections are based on Pearson (1998).

international standard-setting body, includes the following definition of term in ISO 1087 (1990):[40]

> 5.3.1.2 **term**: Designation (5.3.1) of a defined concept (3.1) in a special language by a linguistic expression.
>
> NOTE – A term may consist of one or more words (5.5.3.1) [i.e. simple term (5.5.5) or complex term (5.5.6)] or even contain symbols (5.3.1.1).

The procedures, however, to collect and process terms may of course also be automated and hence constitute the core of computational approaches to automatic term extraction from text corpora – one of the two major foci of this thesis. As one of the seminal works in this respect, Justeson & Katz (1995) extensively motivate and elaborate on defining the properties of terms from a linguistic perspective, something which other approaches to terminology have not done systematically (subsection 2.2.7). However, the crucial groundwork towards defining terms from a linguistic perspective and establishing the respective properties that may be utilizable for computational approaches has been carried out by research on the notion of sublanguage, i.e. the linguistic properties which make language use in specialized domain different from general every-day language use (subsection 2.2.6).

2.2.1 General Theory of Terminology

Terminology gradually began to emerge as a separate discipline when one of its main proponents, Eugen Wüster (see Wüster (1974) and Wüster (1979) as well as Pearson (1998)),[41] contended that terms should be treated differently from general-language words in three respects. First, in contrast to lexicology or lexicography in which the lexical unit is the natural starting point, Wüster's work on terminology sets out from the notion of "concept".[42] As a consequence, a concept should be considered independent of its label or term, even independent of any particular language. For

[40] The dotted numbers in this quotation denote cross-references to other ISO definitions.

[41] This influential approach to the notion of terminology and termhood originated from the positivist movement during the inter-war period, and emerged from the so-called Vienna Circle, a group of philosophers who gathered at Vienna University.

[42] The notion of "concept" as an abstract idea or a mental construct, of course, may warrant a whole discussion on its own because of its many facets depending on the scientific discipline looking at it (i.e., philosophical, ontological, cognitive, etc.).

terminologists such as Wüster, this definition also has a cognitive aspect in that concepts are the product of mental processes in which objects and phenomena in the actual words are first perceived and postulated as mental constructs.

The second point is that terminologists are only interested in vocabulary alone and hence are not concerned with linguistic questions regarding lexis, morphology or syntax. As also noted by Pearson (1998), this seems to suggest that the General Theory of Terminology, and Wüster in particular, perceive terms as being separate from words not not only with respect to their meaning but also with respect to their nature and use. As a separate class, there is a one-to-one correspondence between terms as labels and concepts as mental constructs, and in an ideal world, a term uniquely maps to one concept within a given subject field or domain. As labels, terms are set apart from language in use and enjoy a sort of protected status. Traditional terminologists in this vein took a rather prescriptive stance and thus, in principle, were not concerned with terms in textual use but only with what they represented.[43]

Wüster flanked his goal of establishing a General Theory of Terminology, the classical stance on the study of terminology, by the following tasks:

1. The development of standardized international principles for the description and recording of terms.

2. The formulation of general principles of terminology.

3. The creation of an international center for the collection, dissemination, and coordination of information about terminology, which developed into *Infoterm*[44] and is sponsored by the UNESCO.

Wüster (1979) attempts to draw a clear distinction between terminology and linguistics in order to arrive at an autonomous discipline. There, the objects considered are no longer considered as units of natural language, but rather concepts as clusters of internationally unified features which are expressed by means of equivalent signs of

[43]In fact, Wüster (1974) and Wüster (1979) are concerned with imposing normative guidelines for terminological language use which should be used to establish fixed and standardized meanings of term concepts in order to avoid terminological confusion in technical communication. On the cognitive side, these standardized terms were to serve as a means to represent conceptual structures of particular subject domains.

[44]`http://www.infoterm.info/`

different linguistic and non-linguistic systems. Central to these postulations is the assumption that a concept is universal and its only variation is given by surface forms in different languages. As a matter of fact, Wüster (1979) attempts to prescribe that the language experts and users of a certain domain (i.e., scientists and technicians) characterize their subject field in the same way so that the only possible differences arising would be due to their different languages or their use of alternative linguistic designations for the same object (i.e. interlingual synonymy and intralingual synonymy). Since these divergences could disrupt professional communication, Wüster (1979) was a staunch advocate of a single language for scientific and technical communication, although the efforts to promote terminology standardization on an international level was considered as a more attainable short-term goal.

2.2.2 Beyond General Theory of Terminology

Current terminologists in the vein of Wüster's General Theory of Terminology appear to have loosened the strict division to linguistics. The focus, however, is still on the pre-linguistic (and in this respect also pre-textual) notion that domain experts in an area of knowledge have terms as conceptualized constructs in their mind.[45] Still, Cabré Castellví (2003) slowly approaches the theoretical study of terminology to linguistics in that it is assumed that the elements of a terminology are terminological units and that these are units of knowledge, units of language, and units of communication. Admitting that these are not distinctive features with respect to other linguistic units, such as words or lexical items in general usage, Cabré Castellví (2003) defines the following linguistic conditions which distinguish terminological units from other ones:

1. terminological units are lexical units, either through their lexical origin or a through a process of lexicalization.

2. they may have lexical and syntactic structure, which, however, tends to be more constrained than for general lexical units.

3. regarding word class, they occur as nouns, adjectives, verbs and adverbials, although there is a strong tendency towards nominal and adjectival structures.

[45] What also certainly plays a role here is the strive of terminological researchers to establish (or maintain – depending on the viewpoint) the study of terminology as a discipline in its own right.

4. they may belong to one of the following semantic categories: entities, events, properties or relations.

5. their meaning is self-contained within a special subject.

6. their syntactic combinability is restricted on the basis of the combinatory principles of lexical items, although, in general, it is more restrictive.

Although these conditions show that terminological units, in several respects, exhibit linguistic properties similar to general lexical units, they highlight the contrasts when these properties diverge. On the semantic or discourse level (see conditions 4 and 5), it is noted that their meaning is tied to a particular (technical) domain and that they fall into all of the semantic categories used to describe linguistic structure. On the lexical and syntactic level (see conditions 2 and 6), it is noted that terminological units have a tendency to occur as adjectives and nouns and that they are more constrained with respect to their syntactic structure. As can be seen, this may already be considered a hint at the linguistic property of limited modifiability of terms. In fact, this condition is in line with observations made by research on domain-specific sublanguage use (see 2.2.6 below) and computational approaches to automatic term extraction from natural language corpora, in particular Justeson & Katz (1995) (see subsection 2.2.7 for a more detailed account).

2.2.3 Conventional Definitions of Terms

Newer but still conventional approaches to the definition of terms try to distinguish between terms on the one hand and words on the other hand. In particular, Sager (1990, p. 19) attempts to frame the boundary of terminology to linguistics from this angle:

> The lexicon of a special subject reflects the organizational characteristics of the discipline by tending to provide as many lexical units as there are concepts conventionally established in the subspace and by restricting the reference of each such lexical unit to a well-defined region. Besides containing a large number of items which are endowed with the property of a special reference, the lexicon of a special language also contains items

> of general reference which do not usually seem to be specific to any discipline or disciplines and whose referential properties are uniformly vague or generalized. The items which are characterized by a special reference within a discipline are the "terms" of that discipline, and collectively, they form its "terminology"; those which function in general reference over a variety of sublanguages are simply called "words", and their totality the "vocabulary".

The first assertion, *viz.* that there are as many lexical units as there are concepts, seems to be in line with Wüster's idealistic goal of a terminology reflecting the conceptual structure of a subject domain. Rather problematic is the point that the lexicon of a special language contains two classes of entries, the ones with special reference and the ones with general reference. The fact that the latter kind of items presumably has reference to a variety of sublanguages indicates that they may not merely constitute general-language words used in everyday communication. Still, there are no examples given for the *words* and the *vocabulary* of a particular subject domain and, hence, this attempt to arrive at a distinction between terms on the one hand and words on the other hand remains superficial and poorly motivated (for a similar point why such a distinction may not hold see Pearson (1998)).

2.2.4 Problems with the Classical and Conventional Approaches

Both the classical view (i.e. based on Wüster's General Theory of Terminology) and, with some qualifications previously outlined, its modern successors emphasize the cognitive role of *conceptual maps* in the mind of domain experts. However, even if this were the case, this assumption may turn out to be rather misleading because domain experts would not build up such conceptual maps from introspection.[46] On the contrary, domain experts and terminologists alike constantly refer to textual data and analyze lexical elements for the purpose of acquiring and validating "conceptual" descriptions.

More current but still conventional definitions follow the prescriptive stance of

[46] Jacquemin (2001) provides a similar counter-argument.

the classical view and pursue a one-to-one correspondence between concept and term. Although such a reduction of ambiguity may be useful to ease the burden of communication bottlenecks and to facilitate the compilation of standardized terminologies, it is difficult if not impossible to apply in a computational environment in which one deals with occurrences of terms in text.

Another classical assumption is that terminology is only used by a closed expert community and that each subject domain more or less has its own discrete terminology. Conventional approaches qualify this assumption to a certain respect in that they try to distinguish between terms on the one hand and words on the other hand, whereby the notion of words is used as an all-inclusive category for those lexical items which do not fit elegantly into the classical classification scheme for terms. As a consequence, technical terms which are used in a single subject field are set apart from general terms which are used in more than one subject domain.

Both from a text-based and a computational perspective, the problem with such an approach lies in its "closed-world assumption" with respect to subject domains. Already the subject field in which this work is to be located, *viz.* computational linguistics, provides obvious counter-evidence that such an assumption would be feasible in practice. If a (standardized) terminology of the field would have been compiled 15 years ago, it would have looked conspicuously different from a terminology of computational linguistics compiled nowadays. Because of the surge in using statistical and machine-learning-based methodology in computational linguistics, a lot statistics terms[47] would have to be included in such a terminology, although, in a strict sense, these are terms from another, separate subject field. Obviously, an approach that would classify these cross-disciplinary terms simply as "words", as opposed to the genuine computational linguistics terms (as suggested by Sager (1990) – see subsection 2.2.3 above), would ignore the relevance such terminology has attained in current computational linguistics research. It should also be noted that a computational approach to automatic term recognition from text, as described below and pursued in this work, would definitely output a lot of statistics terminology, in particular if the document collection for such a procedure would include material from the last ten years. This example also illustrates the problems one would run into with such an approach in the light of the increasing cross-disciplinarity, in which the traditional

[47] For example, statistics terms such as "*maximum likelihood estimate*", "*maximum entropy*", "*hidden markov model*", "*mutual information*", "*hypothesis testing*", "*likelihood ratio*" etc.

demarcation lines between subject fields are becoming blurred and there is often a considerable terminological overlap between them.

2.2.5 Pragmatic Definitions of Terms

"Pragmatic" definitions of terms are understood as approaches which are not committed to a particular theory or framework of terminology, but still are concerned with finding a sort of "working definition" for their purposes.[48] The setting for such definitions is typically such that a language is used for special purposes, such as a second language curriculum. What they have in common is that they try to address the above described shortcomings of the classical approaches to terminology. For example, Hoffmann (1985) suggests that within a specialized vocabulary, there are three categories of terms in a domain-specific terminology: subject-specific terms, non subject-specific specialized terms and general vocabulary. Thus, terms with a special reference in only one domain are distinguished from such which have a special reference in more than one domain. Such an approach certainly acknowledges the fact that a domain-specific terminology cannot be conceived as a static block.[49] A slightly different stance is taken by Trimble (1985) who distinguishes between three types of terms, i.e. highly technical terms, a technical term bank, and subtechnical terms. Highly technical terms are, more or less, roughly equivalent to Hoffmann's (1985) subject-specific terms whereas the technical term bank seems to have an approximate equivalent to the non-subject-specific vocabulary. The third category, subtechnical terms, is constituted by common words that may have adopted special meanings in certain subject fields.[50]

Although such categorizations attempt to accommodate the fact the terminology of a particular domain may be divided into different "semantic" portions, the criteria for membership differ from terminologist to terminologist, and thus make it difficult to arrive at a consistent, let alone standardized way for such a procedure. In fact, Hoffmann (1985) himself concedes that a systematic way of distinguishing between such types of terms is not only difficult in practice but also questionable with respect to its intended purpose in the first place. Furthermore, like it was the case with

[48] See also Pearson (1998) for a similar definition.

[49] It certainly would address the question of computational linguistics vs. statistics terminology discussed above.

[50] Examples given in this respect are *control*, *operation*, *positive*, etc.

respect to the classical approach described above, such categorizations do not give any indications on how it may be helpful for the computational tractability of terms from natural language texts.

2.2.6 Terms and Sublanguage

The notion of *sublanguage* appears to have been made an issue, in particular, by researchers on the boundaries between computational language analysis and its applications in various specialized language domains, such as e.g. the medical domain. In other, more linguistically oriented research, also the terms *language for specific purposes*, *specialized language*, or *scientific language* are encountered. No matter, however, what the label may be, sublanguage always plays an implicit or explicit role when talking about domain-specificity, subject-specificity or a subject field. Therefore, it is important to describe, on the one hand, what the properties of sublanguages are and, on the other hand, if and to what extent these properties may contribute to the definition of terms. As the ultimate goal, we hope to arrive at a better understanding of how to isolate the domain-specific terms of a particular subject domain.

Harris (1968, p. 152), who may be considered as having introduced the concept of sublanguage in the first place, attempts to define it in terms of mathematical (set-theoretic) properties of sentences:

> Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it.

In this respect, a sublanguage would display some sort of mathematical closure, i.e. a finite set of sentences. The same, however, would not hold for the grammar of a sublanguage (Harris, 1968, p. 155):

> Thus the sublanguage grammar contains rules which the language violates and the language grammar contains rules which the sublanguage never meets. It follows that while the sentences of such science object-languages are included in the language as a whole, the grammar of these sublanguages intersects (rather than is included in) the grammar of the language as a whole.

From a current linguistic point of view, such a different treatment of sublanguage sentences, on the one hand, and sublanguage grammar, on the other hand, may seem rather odd. Thus, Sager (1982, p. 9), one of the early investigators of sublanguage in the medical context, defines it as follows:

> The discourse in a science subfield has a more restricted grammar and far less ambiguity than has the language as a whole. We have found that research papers in a given science subfield display such regularities of occurrence over and above those of the language as a whole that it is possible to write a grammar of the language used in the subfield, and that this specialized grammar closely reflects the informational structure of discourse in the subfield. We use the term sublanguage for that part of the whole language which can be described by such a specialized grammar.

The focus of this definition is on the information structure of sublanguages which is reflected in specialized grammatical structures. It should be noted that this rather represents a "top-down" approach in that a (however described) information structure dictates allowable grammatical structures. In the practice of the Linguistic String Project (LSP), however, the specialized sublanguage grammar often did not fit the domain language used (i.e., clinical narratives).[51] This observation is related to the fact that it is not sufficient to talk about a domain-specific sublanguage (or even sublanguage grammatical structures) per se, and thus simply imply the existence of a (sub)language of physics, aeronautics, medicine, etc. (as e.g. Lehrberger (1982) and Lehrberger (1988) seem to suggest). Using such a definition of sublanguage, one at first glance might be tempted to label the linguistic context of the LSP as "language of medicine". This, however, would ignore the fact that such a domain-specific language of medicine itself is not a monolithic block but rather itself may be comprised of different "sublanguages" in form of text categories or genres, such as clinical narratives, textbooks for medical students, scientific publications etc.

In a slightly different vein which is of particular interest with respect to terminology and terms, Hirschman & Sager (1982, p. 27) characterize sublanguage as follows:

> We define sublanguage here as the particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles

[51] As a consequence, e.g., parses output by the LSP system had to be hand-edited (Macleod et al., 1987).

> on a technical specialty or science subfield), in which the authors of documents share a common vocabulary and common habits of word usage. As a result, the documents display recurrent patterns of word co-occurrence that characterize discourse in this area and justify the term sublanguage.

Here it is suggested that language used in restricted domains exhibits recurrent word co-occurrence patterns or habits of word usage[52] which may be utilized as what is termed as the "informational" content of the text. Harris (1988, p. 40) states a similar observation:

> When the word combinations of a language are described most efficiently, we obtain a strong correlation between differences in structure and differences in information. This correlation is stronger yet in sublanguages.

To find evidence for his assumption, Harris (1988) developed a sublanguage grammar for a collection of scientific articles on medicine[53] by recording how words occurred with each other in sentences of the articles and by collecting words with similar combinability into classes. Because Harris wanted to establish mathematical properties of sublanguage and language use, he actually replaced words by symbols thus creating a sort of string grammar, in which he then used string analysis to determine patterns of substring combinability.

At this point, we may conclude that there is tight interconnection between domain-specificity, on the one hand, and sublanguage, on the other hand. Although from a current linguistic point of view, various inadequacies may be invoked, both with respect to the theoretical basis and the methodological repertoire of the description of sublanguage properties, various sublanguage researchers have observed that there appear to be limitations and constraints on the grammatical and lexical structure of sublanguages. In subsection 4.2.3, we will see that in fact we will be able to identify such a limiting property for terms, based on the notion of limited modifiability, and that we will be able to use it for their automatic identification in domain-specific text and thus distinguish them from general-language words – a concern that has long been a point of scientific debate among terminologists.

[52] It should be noted these observations are mostly based on manual text data analysis using a KWIC (keyword in context) computer program, or even by manual text corpus data analysis.

[53] On immunology, to be exact.

2.2.7 (Computational) Linguistic Definitions of Terms

Somewhat in parallel to computational linguists doing research on collocation extraction (see subsection 2.1.3), NLP researchers in the realm of term extraction from (domain-specific) text data appear to acknowledge that there is indeed a body of research on definitions and properties of terms. At the same time and in a similar vein, however, there appears to be little impact of this on the development of respective term extraction procedures. Jacquemin (2001) may have the most radical utilitistic stance in that, in the context of corpus-based terminology, he defines a term as the *output* of a procedure of terminological analysis. Such an approach may very well be the reaction to the position of classical and conventional terminologists to draw a clear distinction between terminology, on the one hand, and linguistics, on the other hand.

While acknowledging that the notion of terminology (or technical terminology) may neither have nor need a formal definition in the context of automatic term extraction, Justeson & Katz (1995), in their seminal work on automatic term extraction from corpora, emphasize that there are *linguistic properties* of terms which may be derived, either by careful manual analysis of domain-specific dictionaries or text corpora or by both. In their corpus and dictionary analysis from the several subject domains (i.e. fiber optics, medicine, physics, mathematics, psychology), one first important syntactic property derived by Justeson & Katz (1995)'s analysis is that the vast majority of terms actually are or occur within noun phrases (NPs), which are referred to as *terminological noun phrases*. They are distinguished from "other" NPs in that they are *lexical*,[54] i.e. they are supposed to be of limited compositionality and thus are distinctive entities requiring inclusion in the lexicon because their meanings are not unambiguously derivable from the meanings of the words that compose them. The presumed property of limited compositionality, however, is not investigated any further, and it may not be as straightforward, as it is in the case of collocations in the first place. In their discussion of collocational subclasses, Manning & Schütze (1999) (see subsection 2.1.3), for example, note that terms may often be fairly compositional (cf. their example "*hydraulic oil filter*").

Justeson & Katz (1995) outline two general linguistic properties of terminological noun phrases, of which the first one is vaguely described as "statistical" and the

[54]Hence, they are also sometimes referred to as *lexical NPs*.

second one as "structural". These two properties are then incorporated into their terminology identification algorithm to different degrees (see subsection 3.2.1). Although the former (statistical) property is also labeled as "repetitive" , a closer look at it reveals that it boils down to the – by now well known – frequency of co-occurrence property (see the discussion on collocations in subsection 2.1.4 above), which is also set to hold for terms by stating that terms occur more frequently than non-terms. What is interesting in Justeson & Katz (1995)'s discussion is that, for terminological noun phrases, this property is linked to a more restricted range and extent of modifier variation as well as to repeated references to the entities designated. Repetition in case of non-terms, on the other hand, is much more restricted [p.11]:

> Repetition including the modifiers of a nonlexical (i.e., non-terminological) NP can be appropriate pragmatically, when repetition of the specifying function is motivated ... The more modifiers are involved, the less likely such possibilities are. Even when repetition of the full NP might be pragmatically appropriate, precise repetition can create a tedious or monotonous effect, the more so the NP and the more recently the repeating phrase was used; some sort of stylistic variation is usual. Exact repetition of nonlexical NPs is expected to occur primarily either when they are widely separated in relatively large texts or else as an accidental effect.

In the case of terminological NPs, on the other hand, the property of repetition is natural:

> ... omission of modifiers from a lexical NP normally involves reference to a different entity. Lexical NPs – even those with compositional semantics – are much less susceptible to the omission of modifiers. When a lexical NP has been used to refer to an entity, and that entity is subsequently reintroduced after an intervening shift of topic, the reintroduction to it is very likely to involve the use of the full lexical NP, especially when the lexical NP is terminological. Lexical NPs are also far less susceptible than nonlexical NPs to other types of variation in the use of modifiers. Modifying words and phrases can be inserted or exchanged within a nonlexical NP but not, without a change of referent within a lexical NP. Similarly,

> the precise words comprising a nonlexical NP can be varied without a change of referent, but usually not in a lexical NP. Variations either in the choice of some words or in the presence vs. absence of some words in terminological NPs reflect distinct terms, often differentia of a noun or NP head.

Here, it appears as if the property of repetition (i.e., frequency of occurrence) is tied to another property, i.e. limited or lack of variation.[55] Thus, in this respect, there appears to be mounting evidence that also terms seem to exhibit certain forms of limitations on their variability or, to put it in different words, their modifiability – a notion of a linguistic property which also has been hinted at by Cabré Castellví (2003) above in subsection 2.2.2. Although such a property typically would be perceived as being different from frequency of co-occurrence, Justeson & Katz (1995) consider it as a sort of prerequisite for it. This means that, due to limited variability (or modifiability) within lexical noun phrases because of their terminological status, terms are often repeated in text and thus exhibit a high frequency of occurrence. Thus, as we will see in the description of their term extraction procedure in subsection 2.2.2, the adduced property of limited variability or modifiability has no influence on their actual term extraction algorithm, which is merely based on the "repetitive" notion of frequency of co-occurrence counting of candidate items.

The other general property of terminological noun phrases postulated by Justeson & Katz (1995) states that terminological NPs also differ in structure from non-lexical NPs. For each domain corpus and dictionary resource examined in their study, samples of 200 technical terms were analyzed, of which 92.5% to 99% were found to be noun phrases. An interesting (because typically assumed) finding was that most terms were actually multi-word items with a length greater than one. Indeed, almost 80% of all terms across all domains examined were multi-word terms, with the average length of NP terms being 1.91 words. Justeson & Katz (1995) explain this by the fact that one-word terms are typically quite ambiguous or polysemous and thus multi-word terms are preferred. Between 50% and 63% of the multi-word-terms analyzed were two-word items,[56] between 6% and 20% were three-word items, and only up

[55]It should be noted that Justeson & Katz (1995) exclude determiners (articles and quantifiers) from the class of NP modifiers because, first, they are applicable to almost any NP and, second, because they tend to indicate discourse pragmatics rather than lexical semantics.

[56]Daille (1996), who conducted a (manual) corpus study for both English and French terms from

to 6% were four words long or more. Of these multi-word terms, the vast majority (97%) only contained *nouns* and *adjectives*, and hardly any other parts of speech, such as prepositions or adverbials.[57] These findings on term length are also in line with other NLP approaches to term extraction, e.g. (Jacquemin, 2001) and (Jacquemin & Bourigault, 2003) who state that multi-word items should be the focus of automatic procedures for the acquisition and recognition of terms from text whereas one-word terms, besides being far less frequent, should be rather subject to word sense disambiguation procedures and thus constitute a completely different field of computational approaches.

2.3 Assessment of Linguistic Definitions for Collocations and Terms

The previous subsection 2.2.7 has shown that computational linguistics or NLP researchers working on terminology extraction typically refrain from defining terms from a theoretical terminological perspective or from considering corresponding properties put forth by terminologists. This may have to do with the fact that the theory of terminology, in defining (and justifying) its own field, has sought to draw a clear demarcation line from linguistics. This might even have led some NLP researcher to take a completely utilistic stance on this issue by simply defining terms as the output of a term extraction procedure (Jacquemin, 2001). In any case, if NLP researchers define properties of terms or terminology at all, they do this so on linguistic grounds. The most prominent work on this, (Justeson & Katz, 1995), reveals interesting findings about the structural constitution of terms in text, pointing to a constrained variability which in turn leads to repetitiveness (i.e. frequency of co-occurrence). Interestingly, this (manually derived) empirical observation has been independently echoed by some current terminologists in the vein of Wüster's General Theory of Terminology (Cabré Castellví, 2003) which appear to have loosened the strict division to linguistics (see subsection 2.2.1, in particular Cabré Castellví (2003)'s postulated linguistic conditions 2 and 6). Justeson & Katz (1995)'s main insight centers around

the telecommunications domain, also found that most terms are bigrams, for both languages. However, their actual proportion was not quantified.

[57] In the physics domain, however, Justeson & Katz (1995) report that there are disproportionately high numbers of adverbials, as e.g. witnessed by the term "*almost periodic function*".

the observation that variation of modifying words within a terminological phrase, be it their insertion, deletion or substitution, either changes the referent of the term (i.e., a different "entity") or turns the expression into a non-term. Although, in order to put Justeson & Katz (1995)'s linguistic analysis into practice, sophisticated linguistic analysis may be necessary to identify the modifiers (and the head) of a noun phrase,[58] the basic insight into possible terminological NP modifications and their effect is intriguing and thus may be worthwhile to be incorporated into a linguistically enhanced statistical association measure as the backbone for automatic term identification (cf. subsection 4.2.3). This is even so much inciting, as the linguistic properties with respect to terminological NP variability, although being analyzed and elaborated on quite extensively, were not included into Justeson & Katz (1995)'s term identification algorithm, but rather only its presumed effects – repetition (i.e., frequency of co-occurrence).

Unlike theoretical research on terminology, research on collocations has been a vital part of linguistics research and, in the case of British contextualism, even the driving force (see subsection 2.1.2). Also, computational linguists working on collocation extraction (see e.g. Manning & Schütze (1999) in subsection 2.1.3) address various linguistic properties of collocations and, typically, refer to linguistic work done on collocations. Of the linguistic properties addressed, however, it is mainly the contextualist property of lexical co-occurrence (see subsubsection 2.1.4.1) which has a direct or indirect role, both in various collocation extraction algorithms (see section 3.1 ahead) or in respective term extraction algorithms, such as e.g. the one proposed by Justeson & Katz (1995) (see section 3.2 ahead).

Among the linguistic properties of collocations assembled and synthesized in subsection 2.1.4, however, it is the property on linguistic variability of collocations, *viz.* limited modifiability (see subsubsection 2.1.4.1) which appears to bear quite some resemblance with the aforementioned property of multi-word terms put forth by Justeson & Katz (1995), *viz.* limited or lack of variation. We have seen that for collocations this property surfaces as limiting the degree by which collocational components may be modified by *additional* lexical material while for terms this property, among others, rather restricts the modification by *substitutional* lexical material. Now, in subsubsection 2.1.2.2 we have actually introduced a linguistic frame of reference – Firth's

[58] This is something Justeson & Katz (1995) do not attempt. Rather, they employ shallow linguistic analysis (see subsection 3.2.1).

lexical-collocational layer of language description – which will help us to structure the notion of modifiability for both collocations and terms appropriately, *viz.* from a *syntagmatic* and a *paradigmatic* perspective, respectively (see subsection 4.2.1 below). Crucially, this in turn will help us to make these properties empirically quantifiable – in order to turn them into linguistically enhanced statistical association measures for the extraction of collocations and of terms from unrestricted text (see sections 4.3 and 4.4 below).

One additional aspect that the previous sections on defining collocations and terms have revealed is that these linguistic expressions tend to be located within different types of textual discourse. Thus, terminologists anchor terms, to various degrees though, as active ingredients to certain technical domains or subject fields. This, in turn, has the linguistic consequence that the issue of terms and terminology is very closely tied to the linguistic notion of sublanguage as a fundamental notion in describing domain-specificity, subject-specificity or a subject field from a linguistic perspective. Also, the concept of sublanguage has generated a lot of attention as a result of the increasing importance of specialized languages, both from a linguistic and language processing perspective (see subsection 2.2.6). In addition to that, it is interesting to note that also various sublanguage researchers have observed that there appear to be limitations and constraints on the grammatical and lexical structure of sublanguages. Thus, these observations also appear to fall in line with observations on the structure of terms made by some contemporary terminologists (Cabré Castellví, 2003) and NLP researchers (Justeson & Katz, 1995) which are extensively discussed above. Concerning collocations, there is the perception, at least from a linguistic perspective, that they are constructions which are part of what is typically referred to as *general language*. The notion of general language is itself vague and, like the notion of sublanguage, can not be pinpointed to a single monolithic block. In this respect, corpus linguists (e.g. Biber (1993)) have argued that general language may be viewed from different levels and perspectives – typically referred to as *registers* – which should be considered in assembling a well-balanced text corpus of a given language. But still, as far as the linguistic status of collocations is concerned, it is pervasive to all facets of general language, and may not be regarded as more or less prominent in one particular register or another.

Chapter 3

Approaches to the Extraction of Collocations and Terms

This chapter gives an extensive overview over the various approaches to collocation and term extraction which have been proposed in the computational linguistics research literature. Of course, the goal of such a chapter cannot be to discuss every single approach ever proposed but rather to focus on the most representative and influential ones. In this respect, it has to be noted that basically all approaches proposed for the extraction of collocations and terms from text make use of several standard statistical and information-theoretic association measures which compute some form of association score determining the collocativity or termhood of a given linguistic expression and decide whether or not it qualifies as a collocation or term. While these association measures were first devised and used in areas completely unrelated to computational linguistics, such as the statistical testing of differences for various experimental design set-ups in e.g. medicine or psychology, their underlying statistical capabilities became quite popular during the empirical turn of computational linguistics research in the 1990s.[1]

[1] Already as early as in 1964, the then precursors of today's computational linguistics and information retrieval communities, meeting at the *Symposium on Statistical Association Methods For Mechanized Documentation* documented by Stevens et al. (1965), took notice of the availability of statistical procedures which allow to compute some score determining the strength of association between two words (Dennis, 1965). It was only because of the ban of statistics from the computational linguistics community which occurred afterwards why their use did not gain wider prominence until the 1990s.

While the main focus of the first two sections 3.1 and 3.2 will be on discussing the most pertinent approaches to collocation and to term extraction, respectively, the aforementioned association measures – being the fundamental building block – will play a role in these discussions, although their underlying statistical properties will be described later in section 3.3. Also, an influential study on collocation extraction (Dunning, 1993) will be demoted to this section as it is tightly interlinked with one such particular association measure, *viz.* log-likelihood. Consequently, although we divide the computational approaches into those tackling collocation extraction, on the one hand, and term extraction, on the other hand, the boundaries between them are not always as clear-cut as the boundaries are in the linguistic literature between collocations and terms, as discussed in the previous chapter. This is certainly also due to the fact that the processing machinery applied to both kinds of linguistic expressions is similar if not even equal in both cases – a matter which was already hinted at in section 2.1.3. Still, by and large, such a division is corroborated by explicit or implicit statements either made by the authors of a certain study themselves or by references to it. Furthermore, what they all share is the filtering of text by some form linguistic processing in order to obtain a set of collocation or term candidates to which an association measure may then be applied. The fact that such linguistic filtering may also be a beneficial to meet the statistical assumptions of some association measures will be an issue in subsection 3.3.6.

3.1 Approaches to Collocation Extraction

Most of the procedures to collocation extraction may be distinguished in terms of either the kind of linguistic processing performed, which ranges from shallow part-of-speech assignments to full dependency-based syntactic parsing (see subsection 3.1.3 on Lin's approach), or the sort of association measure employed. Also, whether linguistic processing precedes or follows the application of an association measure to some collocation candidate set is where approaches may differ. In particular, the work described subsection 3.1.2 (Smadja, 1993) reverses the canonical order of first applying a linguistic filter and then an association measure. The approach outlined by Berry-Rogghe (subsection 3.1.1), on the other hand, suffers from the severe limitations of linguistic processing capacity that were prevailing at that time.

What distinguishes the approach by Lin (subsection 3.1.3) from the other ones is

that its main focus lies on fine-classifying an already extracted set of collocations. The work by (Evert & Krenn, 2001; Evert, 2005) described in subsection 3.1.4 is also distinctive in that it compares different association measures from a mathematical point of view and attempts to frame a sound evaluation setting to make them comparable in the first place. The insights from this study will in fact be guiding the way of how we will construct our evaluation setting in section 4.5 of this thesis.

3.1.1 Berry-Rogghe

Berry-Rogghe (1973) may be regarded as one of the earliest approaches to the automatic extraction of collocations from machine-readable text. Her approach to automatic collocation extraction was motivated by the goal to develop a general method to isolate "significant" collocations from machine-readable texts.[2] On the linguistic side, this study relied heavily on the contextualist lexical approach to collocations laid out by Halliday (1969) (see subsubsection 2.1.2.3 above) and thus attempted to isolate potential *nodal items* (or collocational bases, in the terminology of phraseologistic approaches) and their *collocates* from text. This, however, posed several challenges which were hard to overcome by the available linguistic processors at that time.[3] Even if a potential nodal item was determined, one particular problem was in estimating the appropriate range for the collocational span and find a potential collocate, which was approached by heuristically experimenting with different span sizes and without taking into account any sentence boundary information. Furthermore, due to the limitations of computing resources and processing power, the size of the text corpus (approximately 1000 sentences) was comparatively small. Moreover, no evaluation of the quality of the extraction procedure is reported, which again is due to the period in which the study was undertaken. Still, the interesting aspect of this study is that the notion of *significant collocation* was defined according of the established notion of statistical significance and thus, to a certain respect at least, anticipated what would become mainstream in computational linguistics 30 years later. In fact, Berry-Rogghe (1973)'s adopted (and heuristically adapted) statistical methods to identify collocations – the z-score – may be regarded as peculiarly "modern", in particular in

[2]The text corpus used in this study, however, consisted of literary and philosophical texts.

[3]For example, the automatic detection and disambiguation of parts of speech or the identification of phrasal units was almost impossible to achieve.

comparison to current standard references on statistical natural language processing (e.g. such as Manning & Schütze (1999)):[4]

> the aim is to compile a list of those syntagmatic items ('collocates') significantly co-occurring with a given lexical item ('node') within a specified linear distance ('span'). 'Significant collocation' can be defined in statistical terms as the probability of the item x co-occurring with the items a, b, c, ... being greater than might be expected from pure chance (Berry-Rogghe, 1973, p. 103)

Due to the aforementioned lack of linguistic preprocessing capacities (e.g. lack of sentence boundary recognition or part-of-speech disambiguation), Berry-Rogghe (1973) looks at all words within a predefined span and counts the number of co-occurrences of a potential nodal item (collocational base) and a potential collocate. Computing the z-score is then done in quite an idiosyncratic way in that the predefined size of the collocational span is factored into it.

Besides the severe limitations on linguistic processing capacity in this study, the applicability of the z-score to the task of collocation extraction has been questioned. For example, Dunning (1993) points out that the z-score substantially overestimates the significance of rare events. Hence, its application to statistical NLP problems in general may be considered inadequate and thus a close relative to it, the t-test, is typically preferred (see subsection 3.3.2 for a description of the statistical underpinnings of both the z-score and t-test).

3.1.2 Smadja

Smadja (1993), and its precursor Smadja & McKeown (1990), may be described as one of the classical works on collocation extraction from natural language text corpora. Its main focus is on the acquisition of collocational knowledge, in particular in addition to established grammatical and semantic rule inventories, for the task of language generation. This is motivated by the fact that language generation algorithms, which only rely on grammatical and semantic rules, fall short of preferentially

[4]Curiously, Berry-Rogghe (1973)'s work is not mentioned in Manning & Schütze (1999)'s chapter on collocations, which, besides this, is remarkably complete regarding previous approaches to collocation extraction.

selecting collocationally adequate word combinations, such as "*take a bath*", instead of syntactically and semantically similar (but incorrect) ones, such as "*have a bath*". According to Smadja, the former word combination is unpredictable (from the point of view of language generation) in the absence of knowledge about collocational rules.

Smadja (1993) distinguishes between three subtypes of collocations, *viz.* open compound collocations (e.g. "*stock market*", "*ice cream*"), phrasal collocations (e.g. "*take out*", "*pump up*"), and predicative collocations. Mainly the first and the third type are focused on although the prominence of compound collocations are certainly a result of both the language under investigation (i.e. English)[5] and the textual domain considered (i.e. stock reports).[6]

From the point of view of the linguistic grounding of Smadja's (1993) approach to collocation extraction, Smadja (1989) gives some information in that the most widely used subclass of predicative collocations is claimed to be tied to Mel'čuk (1995b) and Mel'čuk (1998)'s model of *lexical functions*, as described in subsubsection 2.1.1.1. Moreover, by referring to the notions of *collocational base* and *collocate*, Smadja's conception of collocations bears a certain resemblance to phraseological conceptions of collocations (see also subsubsection 2.1.1.1), however without explicitly mentioning it. It is, however, interesting to note that, although Smadja (1989) invokes Mel'čuk's linguistic work on collocations, there is no reflection of it in the extraction procedure proposed. For Smadja (1993), a collocation is considered as a syntagmatic word association which is based on the part of speech of its component words and which is characterized by a deviation to statistical standard values.

In concrete, Smadja's collocation extraction procedure XTRACT is composed of two subprograms. After a *collocational base* has been manually selected, the program XCONCORD creates its concordances (on a sentence level) from the text corpus. After tagging the concordances with their part of speech, the second subprogram XSTAT first removes all function words[7], although only preliminarily. Then collocational bigram

[5]For example, the types of compounds that fall in this class would be closed compounds in German (e.g. "*Aktienmarkt*") and thus would not be the target of a collocation extraction procedure.

[6]Here, it can also be seen again that, in particular in the area of computational extraction approaches, sometimes there is a fine, almost indistinguishable line between the extraction of collocations, on the one hand, and extraction of terms, on the other hand. Thus, Smadja (1993)'s approach and its focus on stock reports indiscriminatively tackles both the extraction of domain-specific terms and general-language collocations.

[7]Whether or not a token is a function word is determined by its respective part of speech tag.

candidates are generated by collecting directly and indirectly neighboring potential collocates. For these XSTAT computes an association strength score which to a certain degree resembles the z-score (see subsection 3.3.2 for a description of its shortcomings) and which determines whether or not the bigram candidate qualifies as a collocation. Given that w_1 is the potential collocational base and w_2 is the potential collocate co-occurring in the same sentence, this is done in the following way:

- The relative (signed) distances between w_1 and w_2 are computed and averaged to the mean distance
- To measure how the individual offsets of collocate occurrences differ, the sample deviation (i.e., the square root of the variance) of the mean distance is computed.
- An optional (and to be manually determined) smoothing factor may be applied.

The procedure retains those bigram candidates which lie above a manually defined threshold value and thus a fixed list of collocation candidates is obtained. Finally, a syntactic validation procedure, which relies on part of speech assignments, is applied to this list in order to filter out syntactically invalid collocations. In this way, e.g., noun-adverb or preposition-adjective combinations are discarded from the collocation set.

Contrary to most current approaches to both collocation and term extraction, Smadja applies a linguistic filter *after* statistically computing lexical association scores. In more recent approaches to collocation extraction linguistic preprocessing typically precedes the computation of statistical association scores as this allows for better control over linguistic structures to which the collocation or term identification procedure can be applied.[8] Concerning its underlying linguistic notion of collocations, Smadja's approach is probably considerably more permissive than that of other linguists or computational linguists. As already mentioned, although collocational concepts are cited in the form of Mel'čuk's lexical functions and phraseologistic terminology, many of the collocations found, if at all, may rather be classified as fixed phrases (see their definition in subsubsection 2.1.4.2). Still, for the purpose of

[8]In addition, applying linguistic filtering beforehand may also better meet the statistical assumptions made by many of the association measures applied to the tasks of collocation and term extraction – see subsection 3.3.6 for a detailed discussion.

Smadja (1993)'s application setting, *viz.* language generation, these types of word combinations[9] may certainly be useful to identify.

As for the quality of his extraction method, Smadja (1993) estimates the precision at about 80% by judging the collocational status of items in various collocation list output runs. Such an evaluation procedure, although common at that time, is problematic in various respects, as it only examines the top-ranked items in the output list and thus the accuracy score obtained just superficially reflects the algorithm's actual performance (see also the discussion below in subsection 3.1.4). Since adequate evaluation procedures for both term and collocation extraction should meet a well-defined array of evaluation criteria, they will be discussed in greater detail in section 4.5.

3.1.3 Lin

The approach to collocation extraction from text corpora taken by the the work of Dekang Lin is notably different from other approaches in that it scales up its linguistic processing step to full-fledged dependency parsing and attempts to analyze collocations from a semantic perspective. In particular, it attempts to sort out a particular subclass of collocations, *viz.* those to which the linguistic property of *non-compositionality* applies (for an extensive discussion of this property, see subsubsection 2.1.4.1). By parsing a 125-million word newspaper corpus (containing Wall Street Journal and San Jose Mercury articles) by the dependency parser MiniPar (Lin, 1993; 1994), Lin (1998a) and Lin (1998b) assemble a lexical dependency database consisting of dependency triples of the form (*head type modifier*) where *head* and *modifier* are words in the input sentence and *type* is the type of the dependency relation. The dependency types looked at were noun-verb and noun-adjective dependencies. This way, about 80 million dependency relationships were collected from the parsed corpus. The collocation database, then, was obtained by computing the log-likelihood ratio[10] of the respective frequency counts. All dependency triples above a manually defined threshold for the log-likelihood value were considered a collocation and thus a database of about 11 million "unique collocations" was obtained. Of each of these

[9]Examples given by Smadja include the following noun-adjective combinations: "*narrow escape*", "*powerful car*", "*strong protest*", etc.

[10]See subsection 3.3.3 below for a detailed description of this lexical association measure.

dependency triples, Lin (1999) then computes the mutual information (MI) value.[11] Here then, a second resource constructed and described by Lin (1997) comes into play, *viz.* an automatically constructed corpus-based thesaurus consisting of 11,800 nouns, 3,600 verbs and 5,600 adjectives/adverbs. In order to determine whether a collocation candidate is non-compositional or not, Lin (1999) makes use of the linguistic property of *non- or limited substitutability* with respect to collocations (see subsubsection 2.1.4.1 for its description). In concrete, he determines the compositionality status of a collocation candidate by comparing the mutual information values when substituting one of the words with a similar word from the thesaurus. Non-compositionality is assumed if such substitutions are not found in the collocation database or if their mutual information values are significantly different from that of the original phrase.

Although appealing both with respect to methodology and research objectives, Lin's (1999) approach is not so much a procedure to automatically extract collocations from natural language text (i.e., effectively separate them from the "non-collocations") but rather a method to fine-classify an already acquired and ranked set of collocations. As a matter of fact, the actual collocation extraction step is performed by applying the log-likelihood statistical lexical association measure to the parsed set of dependency triples. Not surprisingly then, the MI- and thesaurus-based method for fine-classification mostly yields those types of collocations that would be classified as *idioms*, in compliance to the collocation subtypes laid out in subsection 2.1.4.2, because these constitute the collocational subtype which is mostly non-compositional.[12] Hence, the procedure proposed by Lin virtually begins where actual collocation extraction methods, such as the ones described in subsections 3.1.1, 3.1.2, 3.1.4, and of course the methodology proposed in this thesis, leave off.[13]

There are also some principled problems with the approach laid out by Lin (1998a), Lin (1998b), and Lin (1999). First, concerning the evaluation of the set of non-compositional collocations identified, Lin (1999) compared these to two manually compiled idioms dictionaries, *viz.* the NTC's English Idioms Dictionary (Spears & Kirkpatrick, 1993) and the Longman Dictionary of English Idioms (Long & Summers,

[11] In order to be able to compute the mutual information (MI) value of a triple, Lin (1999) uses an extension to MI proposed by Alshawi & Carter (1994). See subsection 3.3.4 for a detailed description of this information-theoretic association measure.

[12] A look at the output lists given by Lin (1999) confirms this view.

[13] Consequently, Lin's procedure would not be suited for identifying the other two subtypes of collocations described in subsubsection 2.1.4.2, *viz.* support verb constructions and fixed phrases.

1979). Against the former dictionary, precision and recall values of 15.7% and 13.7%, respectively, were obtained, and against the latter one, these were 39.4% and 20.9%, respectively. As Lin (1999) notes, these results are clearly insufficient and, in the first place, not due to the identification methodology employed, but rather due to the notorious incompleteness of manually compiled dictionaries (see also the discussion in subsection 3.1.4 below). Moreover, the overlap in idioms between the two dictionaries is quite low, reflecting the fact that different lexicographers may have quite different opinions about which phrases are non-compositional idioms. Second, by exclusively building on parser-derived dependency triples, Lin (1999) is only able to examine collocation bigrams and, moreover, it is not clear how his method could be generalized to larger collocation n-grams. Third, by far the greatest sort of barrage, however, which distorted the evaluation of Lin's procedure were systematic parser errors. Almost 10% of the result set of presumably non-compositional idioms were in fact erroneous dependency triples, produced by MINIPAR, which easily passed the mutual information filter because of the procedure's inability to find similar substitutes for their component words in the collocation database. Besides this, Lin (1998b) already concedes that in constructing the collocation database from dependency triples, already various kinds of steps were taken to reduce the amount of parser errors. Firstly, only sentences with no more than 25 words were fed into the parser, and secondly, only complete parses were included, which reduced the amount of words in the parsed corpus to about 25% of the original corpus size. Moreover, Lin (1998b) also reports on poor local parse decisions (mainly due to an incomplete lexicon and/or ambiguous part of speech assignments) which had to be dealt with. Based on the assumption that the parser tends to generate correct dependency triples more often than incorrect ones, a set of 30 correction rules was manually devised in order to correct potential parsing errors.

The main points of the previous discussion lead to a clear caveat in using full syntactic parsers in order to arrive at a candidate set of collocations. Hence, this seems to suggest that such a collocation lexicon or database should rather be one of the inputs to full syntactic parsing – rather than being its output. This view is also widely supported in the literature on parsing. Interestingly, this does not only hold for rule- and lexicon-based parsers such as MINIPAR but also for statistical parsers. For example, already Collins (1997) showed that the performance of statistical parsers can be improved by using lexicalized probabilities which capture the collocational relation-

ships between words. One of the current cutting-edge hybrid (i.e. both constituent and dependency) parsers, *viz.* the STANFORD PARSER (Klein & Manning, 2003), uses an integral lookup component to collocations listed in the English WORDNET lexical database (Miller, 1995).[14] As a matter of fact, Lin's (1999) version of MINIPAR also uses a WORDNET collocation lookup component by treating entries found there as single words. An unwanted side effect of this, however, is that the parser-derived dependency triple collocation database is skewed in the first place because all non-compositional WORDNET collocations are not included in it at all.

3.1.4 Evert and Krenn

The previous three subsections have shown that there have been various statistical and information-theoretic measures employed for the task of automatic collocation extraction from natural language text. For example, we have seen the z-score being employed by Berry-Rogghe (1973) and Smadja (1993) as well as the log-likelihood and the MI values being employed by Lin (1999) and Lin (1998b). The natural question which arises from this is which one out of the wide array of association measures actually performs best for the task of collocation extraction.

There are several studies which attempt to compare two or more different association measures as for their collocation extraction performance. For example, Dunning (1993) directly compares the log-likelihood and χ^2 measures whereas Church & Hanks (1990) closely examine the MI measure and Church et al. (1991) are responsible for the popularity of the t-test, which they had compared (and found superior) to the MI measure. Being one of the first studies on collocation extraction for German, Breidt (1993) evaluates the MI and t-test measures for the extraction of German noun-verb collocations. Being only a preliminary study, it is based on a very small corpus and a list of 16 verbs which are typically found in support verb constructions. However, rather than comparing the different association measures, Breidt (1993) experiments with various parameters, such as the corpus size and the methods for extracting the word co-occurrences.

The criteria, however, according to which many collocation extraction studies pick

[14] This is noteworthy inasmuch that, besides the collocation component, both Collin's (1997) and (Klein & Manning, 2003)'s statistical parsers are absolutely lexicon-free, deriving all their parameters from treebank annotations.

out a particular measure (but not an alternative one) in order to arrive at their set of collocations, very often remain obscure. This certainly has to do with the fact that the settings in which various association measures have been evaluated tend to be rather subjective and superficial. The typical evaluation procedure is usually as follows: since most association measures output a ranked list of collocation candidates, the author of a paper or, rather seldom though, a linguist or lexicographer examines the top ranked candidates (which is typically referred to as an n-best list where n is the number of top ranked hits) as to whether they constitute a true collocation (i.e. a hit) or not. Since such an evaluation process is rather labor-intensive and cumbersome, n is usually very small, ranging from 50 to at most several hundreds.[15]

By far the most extensive and detailed analysis, which performs a comprehensive and comparative evaluation of different association measures on a common testing ground, has been carried out by Evert & Krenn (2001), Krenn & Evert (2001), and Evert (2005). One of the key findings in Evert & Krenn (2001) is that the widespread *modus vivendi* of evaluating various association measures for collocation extraction on heuristically determined n-best lists clearly leads to superficial judgments[16] about the measures to be examined and thus needs to be put on a more principled basis. In particular, it is suggested to increasingly examine n-best samples, which allows the plotting of standard precision and recall graphs for the whole candidate sets. For this reason, this thesis will also adopt such a principled evaluation procedure, which will be laid out in detail in section 4.5.[17]

Evert (2005) also works out in detail the mathematical and statistical properties of a selection of widely used standard association measures and, in addition, also provides theoretical support (i.e. from a mathematical and statistical perspective) for a widely used practice in statistical NLP, in general, and in employing statistical association

[15] In their various experiments comparing association measures to each other, Manning & Schütze (1999), Chapter 5, merely look at 20 candidates to arrive at conclusive statements about the presumed advantages or disadvantages of the respective measure.

[16] Up to the work of Evert & Krenn (2001), almost all studies on collocation extraction, and also term extraction, evaluated the goodness of their methods using the n-best approach.

[17] Although such an evaluation strategy may allow more objective and principled conclusions about the quality of various association measures, the downside of it is that it is quite labor-intensive as it needs a pre-selected candidate set of potential collocation candidates in which the actual collocations are identified. Still, section 4.5 will outline why and how such an evaluation strategy needs to be preferred and implemented.

measures in, particular, *viz.* the use of cut-off thresholds to exclude low frequency data (i.e. rare events) from statistical inference. With most researchers intuitively suspecting that statistical inference from small amounts of data is problematic (to say the least), Evert (2005) actually shows that *reliable* statistical inference is impossible *in principle* for low-frequency data because quantization effects and highly skewed distributions[18] dominate over the random variation that statistical inference normally takes into account.

Despite its mathematical and evaluative soundness, the work of Evert & Krenn (2001), Krenn & Evert (2001) and Evert (2005) reveals clear shortcomings, which may be exposed on several layers. The major shortcomings from a very general perspective are due to the extreme focus on mathematics and statistics, as a result of which the linguistics about collocations and their properties seems to have gotten lost. One of the fallouts of this is the exclusive focus on bigram word co-occurrences. From a mathematical and statistical perspective, such a procedure is entirely justified because many lexical association measure (e.g. χ^2 or log-likelihood) are only well-defined for word pairs (see also subsection 3.3.5 below). From a linguistic perspective, however, this is clearly insufficient since it is well-know that many collocations and also terms are larger n-gram units (i.e. at least trigrams, if not quadgrams). In Evert & Krenn (2001), the fact that several of the lexical association measures examined are not easily extensible beyond statistical events of bigram co-occurrences is completely ignored. Another aspect which curiously illustrates the over-emphasis on mathematics, on the one hand, and the lack of emphasis on linguistics, on the other hand, are the findings with respect to the question which lexical association measures actually perform best in the task of collocation identification from German adjective-noun and preposition-noun-verb combination. Thus, Evert & Krenn (2001) and Krenn & Evert (2001) report that the best-performing measure is t-test and the second best performing one mere co-occurrence frequency. According to Evert (2005), the former one is, from a theoretical perspective, not applicable to co-occurrence frequency data.[19] The applicability (and sufficiency) of the latter one, a simple counting of co-occurrences, calls into question why lexical co-occurrence data should be targeted with complex statistical machinery in the first place. Thus, given Evert's argumentation, one cannot but wonder whether

[18] In Evert (2005) this is shown for co-occurrence probabilities of pair types, which may be generalized to any statistical inference mechanism.

[19] See subsection 3.3.2 for Evert's (2005) arguments for this point.

there is not more to collocation and term extraction measures than theoretically ill-suited statistical tests or mere frequency of co-occurrence counting.

3.2 Approaches to Term Extraction

The clearest use case which distinguishes term extraction from collocation extraction is stated by Daille (1994) who emphasizes that automatically extracting terms from a domain-specific corpus is essential to reduce the time and labor needed to build a terminology database for a specific subject domain. From a methodological point of view, such a clarification might seem necessary as the vast majority of approaches to term extraction use the same processing techniques, in terms of linguistic filtering and association measures, as the approaches to collocation extraction described in the previous section. One notable exception described in subsection 3.2.3 below is Frantzi et al. (2000) which introduce an association measure that is geared at the extraction of domain-specific terms. The approach by Justeson & Katz (1995) (subsection 3.2.1), although seminal at establishing linguistic properties of terms, confines itself to applying frequency of co-occurrence as association measure. That this may not be such a bad choice is corroborated by Daille (1994) and Daille (1996) (subsection 3.2.2) who finds that there is no noticeable difference in extraction performance between mere frequency counting and applying the information-theoretic mutual information measure. Finally, the work by Jacquemin (2001) (subsection 3.2.4) places the task of term extraction in the wider context of knowledge acquisition.

The evaluation practice for most studies on term extraction is actually as problematic as for collocation extraction in that typically domain experts are consulted, who only inspect some top ranked number of a ranked output list returned by some measure. Therefore, the work on collocation extraction by (Evert & Krenn, 2001; Evert, 2005) which frames a sound evaluation setting is also pertinent to term extraction. Still, an exception to this form of insufficient evaluation is (Daille, 1994; 1996) who evaluates her terminology extraction procedure against the entries in a terminology database.

3.2.1 Justeson and Katz

We have already laid out Justeson & Katz (1995)'s definitional work on the linguistic properties of terms (see subsection 2.2.7), in which the property of recurrence (i.e. frequency) and limited variability were identified as the key characteristics which distinguish terminological NPs from non-terminological ones. The latter property, limited variability, although pertinent from a linguistic point of view and also acknowledged in the large body of research on sublanguage analysis (see subsection 2.2.6), is however not granted an independent linguistic status (which may thus be independently quantifiable) but is closely tied to the frequency of co-occurrence property. In fact, Justeson & Katz (1995) explain a term's higher frequency through its limited variability property and thus see no necessity to quantify this property separately. The basic association measure underlying their term extraction algorithm is thus mere frequency of co-occurrence counting, on which they impose a cut-off threshold of two ($f \geq 2$).[20]

On another dimension, Justeson & Katz (1995) have found that terminological phrases are multi-word units of which the vast majority are bi- and trigrams. For their experiments, they consequently restricted themselves to n-grams of this size. The input texts to their terminology identification algorithm were linguistically preprocessed by a part-of-speech filter.[21] In order to identify potential term candidates, the following part-of-speech regular expression pattern is applied:

- $(Adj|Noun)^{+}|((Adj|Noun)^{*}(Noun\ Prep)^{?}(Adj|Noun)^{*})Noun$

From this regular expression pattern, Justeson & Katz (1995) manually sort out what they call permissible patterns for bigrams and trigrams which reduces the set of allowable part-of-speech sequences to the following two for bigrams and the following five for trigrams:

- *Adj Noun*
  Noun Noun

[20] This approach to term extraction is also taken by other researchers, e.g. (Damerau, 1993).

[21] The expression *part-of-speech filter*, used by Justeson & Katz (1995) themselves, is somewhat misleading as to the actual linguistic processing because it is not a *part-of-speech tagger*. In fact, they perform a dictionary lookup for each word and retrieve all possible parts of speech. Then, the word is identified as a noun, adjective, or preposition, in that order of preference if any of these is retrieved as a part of speech for the word; otherwise the whole candidate string is rejected.

- *Adj Adj Noun*
  Adj Noun Noun
  Noun Adj Noun
  Noun Noun Noun
  Noun Prep Noun

Another restriction put on the permissible part-of-speech sequences concerns prepositions for which they recommend that they be excluded.[22] In particular, in a preliminary run on various texts it is found that if prepositions are allowed, relatively few of the candidates including them turn out to be valid terms, leading to a decline in precision. On the other hand, the recall gains through the inclusion of prepositions are so low that Justeson & Katz (1995) advise their exclusion from the set of allowable part-of-speech patterns.[23]

In order to prevent non-desirable expressions to slip through their part-of-speech and their frequency filter, Justeson & Katz (1995) cannot help themselves but advise another heuristic concerning the exclusion of specific words. These may be verbs interpretable as nouns (e.g. "*go*", "*see*", "*do*", "*can*") or general adjectives (e.g. "*following*", "*normal*"). It is admitted, however, that such a list of stop words may not be applied blindly and that every domain may require different ones.[24]

Concerning the evaluation of their term extraction procedure, Justeson & Katz (1995) only took three articles from three different domains (statistical pattern classification: 2300 words; lexical semantics: 6300 words; liquid chromatography: 14,900 words) and asked the authors of these texts to mark what they would consider the technical terms in the articles. Against this "gold standard", precision (called "quality") and recall (called "coverage") were evaluated. This evaluation procedure is mainly justified with the observation that terminological dictionaries are either insufficient or non-existent for many subject fields,[25] in particular for their domains under

[22] As a matter of fact, a good portion of this article reads as some kind of *best practices manual* for devising term extraction algorithms.

[23] It is noted that domain-specific terms containing prepositions are typically expressions which follow the *noun preposition noun* pattern, such as the statistical term "*degrees of freedom*" or the legal term "*freedom of speech*".

[24] Hence, it is admitted that the word "*can*" may not be removed when dealing with packaging or waste management texts. Similarly, the adjective "*normal*" must not be excluded when domain of interest is statistics (cf. the term "*normal distribution*").

[25] This observation is similar to what has been observed with respect of the coverage of general-

consideration. Although the method of evaluation pursued here would be regarded as clearly insufficient from the perspective of current evaluation standards, one curious finding was that with increasing text size, the precision dropped considerably, which Justeson & Katz (1995) explain as an inherent property of their frequency-based algorithm: being no longer stylistically obtrusive or inappropriate, longer texts would again allow the repetition of non-terminological NPs.

3.2.2 Daille

Concerning both the linguistic preprocessing of a domain-specific corpus in order to isolate potential term candidates and the subsequent deployment of lexical association measures, Daille (1996) and Daille (1994), in a study on terminology extraction for French terms from the telecommunications domain, proceed in a more sophisticated manner than Justeson & Katz (1995).

For linguistic preprocessing, a statistical part-of-speech tagger (not filter) is used although no indication is made as to the type of statistics employed (e.g. a Hidden Markov Model). Other than Justeson & Katz (1995), Daille (1996) only focuses on bigrams, which is justified by two reasons. First, as already pointed out in subsection 2.2.7, the majority of multi-word terms are actually bigrams and thus an effective term extraction procedure is already bound to find a substantial amount of terms among bigram candidates. Secondly, and probably more importantly, one of the lexical association measures employed in her study, *viz.* log-likelihood, is not well-defined for n-grams of a size larger than two (see subsection 3.3.5 below) and thus the linguistic scope of her approach contains an inherent limitation in the first place. The other association measure employed, mutual information, is extensible to larger sized n-grams.[26] Ignoring words void of semantic contents (such as determiners and adverbials), Daille (1996) only examines adjective-noun and noun-noun combinations, which in French surface as noun-adjective and noun-preposition-noun patterns. The candidate pairs (2,200 pairs) are obtained from two French corpora from the telecommunications domain which amount to about 800,000 words all together.

The research question which association methods to use in order to compute the

language collocation dictionaries (cf. (Lin, 1999) in subsection 3.1.3)

[26]In subsection 3.3.5, the trigram extension to the mutual information measure based on Lin (1998b) and Alshawi & Carter (1994) is presented.

degree of termhood is tackled by applying and comparing three measures to the set of bigram term candidates. Besides the "base statistics" of raw frequency counting, mutual information (MI) (Church & Hanks, 1990) and the log-likelihood measure, as it was first proposed by Dunning (1993), are also examined. In order to arrive at a meaningful comparison of these measures, Daille (1996) attempts to put the evaluation on a sounder basis than other studies, such as (Bourigault, 1995; 1992) but also Frantzi et al. (2000), which only have domain experts look at the top outputs of their procedures. For this purpose, the entries of an expert terminology database from the telecommunications domain are taken and matched against the set of 2,200 candidate bigrams. The problem with the evaluation approach, however, is that only bigram surface structures of the form *noun-(preposition)-noun*[27] were contained within the database term set. Hence, Daille (1996) only considered the respective surface structure of her candidate set, from which 1,200 candidate terms intersected with the database set. This means that 55% of the candidate bigrams are actual terms, which is a comparatively high proportion for a candidate set, and thus any conclusion derived about the quality of an association measure may have to be handled with care. Although only precision was examined, the results obtained were surprising, in particular for the author of the study, in that raw frequency counting actually performed equally well as the best "genuine" statistical association score, log-likelihood, which leads Daille (1996, p. 64) to the conclusion that frequency of co-occurrence "undoubtedly characterizes terms".[28] On the other hand, the poor performance of mutual information is explained with the linguistic preprocessing applied. This, however, seems to be unreasonable because it is not clear why and how an association measure like mutual information deteriorates when candidates are passed through a linguistic filter, while, at the same time, no such effect is observed for another association measure, *viz.* log-likelihood.

3.2.3 Frantzi and Ananiadou

A widely used measure to identify terms from domain-specific texts, C-value, has been presented by Frantzi et al. (2000) and Nenadić et al. (2004). Like other methods

[27]The prepositions are ignored for computing the association scores because they only serve a functional role in this construction, in particular for the French language.

[28]A similar conclusion is drawn by Dagan & Church (1995).

proposed, the C-value measure proceeds in a two-staged manner in that, first, a set of potential term candidates is obtained through linguistic filtering and, second, that set is ranked according to the measure.

Linguistic processing is performed in three steps. First, the domain corpus is part-of-speech tagged. As a second step, similar to Justeson & Katz (1995), a regular expression filter is applied only allowing certain part-of-speech sequences and excluding potential function words, such as determiners or pronouns. In particular, the following three patterns are applied:

1. $Noun^{+}Noun$
2. $(Adj|Noun)^{+}Noun$
3. $((Adj|Noun)^{+}|((Adj|Noun)^{*}(Noun|Prep)^{?})(Adj|Noun)^{*})Noun$

The third pattern allows the inclusion of prepositions, which may lead to a higher number of potential false positives, as is already noted by Justeson & Katz (1995) and also mentioned by Frantzi et al. (2000). Another more idiosyncratic step in the linguistic filtering process is the exclusion of words from a stoplist. For its compilation, a sample (i.e. one tenth) of the corpus was examined and words with high frequencies were included, in particular function words and general content words that are not likely to appear in terms (e.g. adjectives such as "*numerous*", "*several*", "*important*"). What Frantzi et al. (2000) themselves admit, however, that, apart from actual function words, the inclusion of so-called "general content words" may be dangerous because some of them may actually appear in terms (cf. the physics term example "*almost periodic function*" from Justeson & Katz (1995) in subsection 3.2.1). For this reason, it is suggested to adapt stop lists in domain-dependent manner, which, however, is clearly a suboptimal solution.

The statistical measure for term extraction, C-value, which Frantzi et al. (2000) introduce, is basically a frequency-based method and incorporates several types of frequencies, which are then taken to compute a termhood score for a certain term candidate:[29]

- The total frequency of occurrence of the candidate term in the corpus.

[29]The way this measure is formally defined will be presented in subsection 3.3.8 below.

- The frequency of the candidate term as part of other longer candidate terms.
- The number of these longer candidate terms.
- The length of the candidate term (in number of words).

As a clear advantage, because it is mainly a frequency-defined measure, the C-value is able to handle term candidates of length greater than two, unlike many of the statistical association scores which are only well-defined for bigrams (see Daille (1994) in subsection 3.2.2). A parameter considered for this purpose is the length of the candidate string in terms of the number of words. Since longer n-grams are less likely to appear a certain number of times in a corpus than shorter n-grams, the candidate length parameter attempts to normalize this difference.

One major reason for not only using mere frequency of co-occurrence counting, which has turned out to be successful as a term extraction measure in Daille (1994), and what makes C-value different from it is the incorporation of nested terms, i.e. the frequency of candidate terms as part of longer ones. Frantzi et al. (2000) motivate this with the example term "*soft contact lens*", which is a term in the domain of ophthalmology. A method that just uses frequency would extract it given it appears frequently enough in the corpus. Its substrings "*soft contact*", which is not a term on its own, and "*contact lens*", which is a term on its own, however, would be also extracted, so the argument, since they would have frequencies at least as high as "*soft contact lens*". The necessity for such a nested term approach, however, lies in the linguistic filters employed. As is correctly noted, both "*soft contact*" and "*soft contact lens*" would be identified by their linguistic filter 2 above, which then of course requires some way of ruling out the former expression as a possible term candidate. A different kind of linguistic preprocessing (e.g. by noun phrase chunking) may not have yielded a non-term expression like "*soft contact*" in the first place, and thus the necessity to incorporate the presence or absence of nested terms into a term extraction measure may be decrepit.

In a further, but independent step, Frantzi et al. (2000) compute context information by means of a measure (NC-value), which is obtained in two steps. From the ranked term list generated by the C-value, context words (verbs, nouns, and adjectives) within a specified window are extracted for the top n multi-word term candidates (where the value of n and the size of the context window have to be manually determined). For each of the context words, a weight is computed by taking

the ratio of the number of terms the context words appears with and total number of terms looked at. In this way the NC context value can be obtained for each candidate in the term list. What is noted by Frantzi et al. (2000), however, is that the context information factor is not a term extraction measure per se but may be rather applied in addition to (and thus must be viewed independently of) any such measure. Hence, it may also be applied to term lists produced by association measures such as frequency of co-occurrence, log-likelihood, mutual information, etc.[30] The context information factor is then added to the score produced by the C-value to calculate the final association value. This, however, is done in a rather arbitrary way by assigning the weights of 0.8 and 0.2 to the C-value and the context factor, respectively, which have been chosen manually after several experiments and comparisons of results. Hence, from the descriptions of Frantzi et al. (2000) it is not clear whether these weights are e.g. general weights or whether they have to be determined for every new domain separately. In computing the C-value, another arbitrarily set value concerns the generation of the term list, into which only those term candidates are included whose C-value lies above a predefined threshold. The incorporation of such an additional threshold is rather obscure, given the fact that according to standard practice in term extraction, Frantzi et al. (2000) already filter candidates by only examining those above a certain frequency threshold (which amounts to four, in their case).

The results obtained by applying both the C-value and the NC-value to a 1-million word corpus of eye pathology medical records seem to indicate that they turn out to be better compared to frequency of co-occurrence. The evaluation carried out, however, exhibits several weaknesses. Lacking a reference terminology for the subject domain examined (ophthalmology), Frantzi et al. (2000) report that they evaluated the quality of their approach by having a domain expert scan the output list produced. One problem with this is that only one expert seems to have been consulted and thus the judgments as to what constitutes a term are not checked against the judgments of a second domain expert – in short: some sort of inter-rater consistency (see also subsections 4.5.1.3 and 4.5.2.3) is completely missing.[31] As a second problem with the

[30]Moreover, several other suggestions have been made how to determine term context information, e.g. by Grefenstette (1994) or Sager (1990).

[31]Frantzi et al. (2000) admit themselves that domain experts – being neither linguists nor terminologists – may disagree on the notion of termhood. This, however, would make some sort inter-rater

evaluation, it is not clear how the domain expert determines the true terms. Frantzi et al. (2000, p.116) seem to indicate that this is done by the domain expert scanning the list from the top to the bottom. This, however, is a clearly biased procedure because the top portion of any (useful) term extraction output list contains a much higher proportion of actual terms than e.g. the lower portion. Thus, scanning from top to bottom will probably bias an evaluator's judgment because it does not reflect the random distribution of terms and non-terms in such a list if it were not ordered. Hence, any judgement of a domain expert as for the true terms in a candidate set would have to be done *a priori* any application of a term extraction method to that candidate set (see also chapter 5 in Evert (2005) for why this is essential.) Furthermore, it is not reported what proportion of the top of the output is examined (i.e. the size of n of these top n candidates is not known). As a result of this approach to evaluating and determining the true terms, it is not known what the proportion of actual terms in the candidate set is, thus making it impossible to determine any exact recall values for the term extraction procedures examined.

3.2.4 Jacquemin

The research by Christian Jacquemin on term extraction (Jacquemin et al., 1997; Jacquemin, 1998; Jacquemin & Tzoukermann, 1999; Jacquemin, 2001) is actually far more comprehensive than any of the other approaches to automatic term extraction presented in this section. This is particularly the case as Jacquemin's work is not only focused on term extraction in the *extraction* sense of the word, but also encompasses the whole NLP and knowledge acquisition framework in which term extraction is to be located. In fact, Jacquemin (2001) puts the issue of computational terminology in a wider context in that a distinction is made between term discovery, on the one hand, and term deployment, on the other hand. On the term discovery side, then, a distinction is drawn between term extraction (or acquisition) and term enrichment. Term extraction is the task at hand when there is insufficient (or even no) terminological data available for a particular technical domain. The input to this task are subject-specific (sublanguage) domain text corpora and the output ranked lists of terms ordered by decreasing degrees of termhood. It is also this kind of task in computational terminology that this thesis focuses on. What is crucial for this endeavor,

consistency checking even more essential.

as also pointed out by Jacquemin & Tzoukermann (1999) and Jacquemin (2001), is the availability of a high-performance lexical association measure in order to arrive at ranked output lists as optimal as possible.

The second issue in term discovery, term enrichment, as described by Jacquemin (2001) and one of his work's major foci, is the endowment of terminological data with additional lexical material in the form of term variants. Term variants may be conceived of as linguistic expressions which basically denote the same concept than the *preferred term* but are expressed in a linguistically different way. One example of this may be the widespread linguistic phenomena of acronyms of a lexical full form.[32] Another, more complex but less widespread form of term variation is typically a different morpho-syntactic construction denoting the same term concept, such as number agreement (*e.g. language generator – language generators*) or prepositional phrase post modification (e.g. *generator of languages*).[33] In Jacquemin & Tzoukermann (1999)'s approach, recognition of term variation is performed by FASTR, a highly complex unification-based grammar formalism inspired by Lexicalized Tree Adjoining Grammar. The backbone of FASTR is a large set of hand-built (French) term grammar meta-rules which are designed to generate term variants (from the original terminological data) and attempt to find them in FASTR-processed text by approximate rule-structure matching. In fact, later versions of FASTR (Jacquemin, 2001) even extend their variant recognition to the semantic layer (such as marking "*context-free language generation*" as a semantic variant of "*language generation*").

Although Jacquemin's approach to term discovery may be described as very ambitious and comprehensive, the downside of it is that it exclusively relies on large hand-built sets of grammar and term meta-rules which, in this case, are even only confined to the French language. Moreover, as also Jacquemin (2001) himself admits, an approach in this vein tends to over-generate potential term variants and thus also includes many false positives in the variant result sets. Thus, this may typically also necessitate a lot of manual post-editing, in addition to the manual effort already involved in grammar rule construction. Furthermore, Jacquemin's (2001) notion of semantic term variant appears to be very promiscuous as it basically allows any ad-

[32]For example, respective full form of the acronym "*CFG*" is "*context-free grammar*".

[33]An additional reason why morpho-syntactic term variation takes such a prominent role in Jacquemin's work may also be due to the fact that it is primarily centered around the French language, which is known to be morpho-syntactically much more productive than English.

ditional lexical material to be included in a variant set, without, however, being able to define the exact semantic relation between term and variant.[34]

The application setting of such a comprehensive approach to computational terminology, however error-prone, may be sought in the domain of knowledge acquisition, according to Jacquemin (2001). This is also the area where the aforementioned issue of term deployment may be located. Thus, given a high-quality list of terms as well as a set of their respective lexical, syntactic or even semantic variants, it is not only possible to construct a terminological database, but also to semi-automatically upgrade it with thesaurus-like relations between terms, such as taxonomic or even further kinds of relations. On the one hand, this may serve as a more comprehensive model of the subject domain under investigation and, on the other hand, it may also help in the controlled indexing of document collections with index terms in order to facilitate applications such as information retrieval.

3.3 Lexical Association Measures and their Application

The previous two subsections have already anticipated that a large array of statistical algorithms has been applied to the modelling and identification of co-occurrence or collocational behavior of words. Long time dismayed by mainstream linguistics (and hence also by computational linguistics), statistical approaches to NLP have experienced a surge starting from the mid-1990s, which is lasting up to today. Still, as outlined in section 2.1.2, investigations into the probabilistic nature of language were well-known via the British contextualist linguistic tradition.

A rather trivial example of the probabilistic properties of language is that some words occur more frequently in language than others. An immediate consequence of this is a correlation between the frequency of a word and its function. In almost all word frequency lists across various language corpora, including those from different genres and subdomains, the top ten to 30 words are quite similar. The majority of the the top ten words will consist of so-called function words (such as determiners, conjunctions, prepositions, etc.). These words share a common property in that they

[34]For example, Jacquemin (2001) denotes the expression "*malignancy in orbital tumours*" as a semantic variant of "*malignant tumour*" without defining the relationship more exactly.

form closed sets which are not as readily expansible as their meaning- and content-bearing lexical counterparts, *viz.* content words. Still, in spite of their functional character, function words form a constitutive part of collocations but not of terms, as will be shown later on.

In this section, we will examine the statistical foundations of the most widely used standard lexical association measures. These measures basically fall into three camps for each of which we will look at the main representatives. The first kind of association measures are the statistical ones, of which t-test (subsection 3.3.2) and log-likelihood (subsection 3.3.3) are the most successful representatives. The second class derives its theoretical foundations from information theory, with mutual information (MI) and heuristic variants thereof being the popular representatives (subsection 3.3.4). The last category is characterized by a property which the other types of association measures also employ to various degrees and which plays a crucial role in characterizing both collocations and terms (as already hinted at in the previous two sections), i.e. frequency of co-occurrence (subsection 3.3.7). In its most basic form, this may already be taken as an association measure itself quite successfully, as it has been shown in various studies (see subsection 3.2.2 above). C-value (which has already been introduced in subsection 3.2.3 above), a heuristically motivated variant of frequency, will also be defined in subsection 3.3.8 below as it has become one of the standard measures for the extraction of terms. In fact, while C-value is basically only employed for the task of term extraction, the other kinds of association measures have been employed for measuring both collocativity and termhood. At first, however, we will review the basic statistical assumptions on which the vast majority of these association measures rely (subsection 3.3.1). From this, it will become clear that most of them suffer from considerable shortcomings as they rely on statistical assumptions which are typically not borne out by natural language, although applying linguistic filters beforehand alleviates some of these deficiencies (subsection 3.3.6). One association measure – in fact log-likelihood as the statistically most sound one (Evert, 2005) – even suffers from an additional "handicap" in that it is not well defined for n-grams larger than size two (subsection 3.3.5), which is a decisive requirement for any association measure for the extraction of collocations and terms.

3.3.1 Statistical Foundations

In this subsection, we will introduce the statistics terminology relevant for describing the lexical association measures employed in this thesis. A more fine-grained and detailed description, in particular of the mathematical underpinnings, may be found in Evert (2005), but also in Agresti (1990), Agresti (1992), Lehmann (1997), Wasserman (2005) and Manning & Schütze (1999).

In terms of a *statistical model*, the goal of statistical analysis is to make inferences about the model *parameters* from the observed data. The core of any statistical model rests on the definition of a *sampling distribution*, which specifies the probability of a particular *observation* (or group of observations) given some hypothesis about the parameter values. Applied e.g. to the NLP task of collocation identification, one such model parameter is the statistical association between two words whereas an observation may be a contingency table (see below) which, in turn, may be derived from the data in a natural language corpus representing the sampling distribution. One important aspect is that, by the nature of statistical reasoning, the sampling distribution must contain some element of randomness, which may however differ from case to case. For example, one such element may be the arbitrary choice of a language corpus (or a certain linguistic construction from it) from a (admittedly all too often hypothetical) set of alternatives. Of course, the form and variability of such a kind of sampling distribution for linguistic data depends on various factors, such as text genres and types, subject matters, author styles etc. Another influence is of course the amount of noise introduced (or deleted) e.g. by automatic linguistic preprocessing. The influence of such linguistic factors is hard to account for by statistical means. Thus, the sampling distribution is usually constructed in such way that a language corpus can be interpreted as a *random sample* of a large hypothetical body of language data, typically referred to as the *population*. Then, the model parameters describe properties of the population and the random sample model enables inferences about these properties from the observed data.

As a sort of pars pro toto for the numerously existing probability sampling distributions, two distributions recurrently used for statistical NLP applications are the binomial distribution and the normal distribution. Being a discrete distribution (whose variables can take on only discrete values), the binomial distribution is a discrete probability distribution with two parameters: the number of successes in a sequence

of n independent yes/no experiments, each of which yields success with probability p. Manning & Schütze (1999, p.51) mention, as a prototypical example assuming such a distribution, the task of finding out how commonly a verb is used transitively by looking through a language corpus for instances of this verb and noting whether each use is transitive or not. As a sort of continuous counterpart (i.e. a distribution whose variables take on continuous values), the normal distribution (also called Gaussian distribution, or more informally – the "bell curve") is considered to be adequate for modeling data in many domains. The parameters for this distribution are given by the mean μ and the standard deviation σ.

In particular with respect to lexical association measures for collocation and term extraction procedures, an accepted way to frame observations is by applying them to a two-by-two contingency table representing the co-occurrence frequencies of word pairs. From such a table, then, model parameters, such as the statistical association between words, can be derived under the specifications of the sampling distribution. That is, such a table is typically used to collect the *observed frequencies* of word pair types thus yielding a four-way classification. By cross-summing the four cell frequencies, the *marginal frequencies* can be computed.

	$V = v$	$V \neq v$	
$U = u$	O_{11}	O_{12}	$= R_1$
$U \neq u$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

Table 3.1: Observed and marginal frequencies

More formally, the observed and marginal frequency data for a word pair (u, v) may be represented as follows in table 3.1 (adapted from Evert (2005)). The cell counts of a contingency table are called the observed frequencies O_{11}, O_{12}, O_{21} and O_{22}. The sum of all four observed frequencies (the sample size N) is equal to the total number of token pairs extracted from a corpus. The row totals of the observed contingency table are R_1 and R_2, while C_1 and C_2 are the corresponding column

totals. Sometimes. the row and column totals are denoted as marginal frequencies (as they are written in the margins of the table), and O_{11} is sometimes called the joint or observed co-occurrence frequency.

At the heart of determining statistical association lies the concept of testing for the *null hypothesis of statistical independence* which indicates that there is no statistical association (e.g. between the components of a word pair type). In particular, the marginal frequencies are used to compute the *expected frequencies* (E_{11}, E_{12}, E_{21} and E_{22}) which indicate what the frequencies of the four cells would be under the null hypothesis, i.e. if there would be no association between the components of a word pair and thus the words would co-occur completely by chance. More formally, the expected frequency data for a word pair (u, v) may be represented as follows in table 3.2.

	$V = v$	$V \neq v$
$U = u$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$U \neq u$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

Table 3.2: Expected frequencies and their computation from marginal frequencies

Of course, the computation of lexical association scores for the task of identifying collocations and terms from natural language text data is motivated by the assumption that the scores provide extensive counter-evidence against the null hypothesis for actual collocations and terms, i.e. that for them there is a higher than chance occurrence. To illustrate this, actually observed, marginal and expected frequencies in one such contingency table are given below for the German preposition-noun-verb (PNV) collocation "*zu Ende gehen*" (to come to an end).[35] These frequencies were computed from a ten-million word corpus of German newspaper texts (see subsection 4.5.2 for a description of this resource and how it is used for the experiments in this thesis). Just comparing the observed and expected frequencies in tables 3.3 and 3.4

[35]As can be seen, the notion of *pair type* does not necessarily imply a word bigram because components of larger n-grams may be collapsed.

shows that, in this case, there actually seems to be a higher than chance occurrence for this particular collocation, since $O_{11} >> E_{11}$

	$V = gehen$	$V \neq gehen$	
$U = zuEnde$	100	150	250
$U \neq zuEnde$	1,877	130,009	131,886
	1,977	130,159	132,136

Table 3.3: Observed and marginal frequencies for a German PNV collocation.

	$V = gehen$	$V \neq gehen$
$U = zuEnde$	3.7	246.3
$U \neq zuEnde$	1,973.3	129,912.7

Table 3.4: Expected frequencies for the same German PNV collocation.

Almost all standard association measures compare the observed frequencies with the expected frequencies under the null hypothesis in some manner and thus compute a test statistic, which is typically referred to as the *association score*.[36] The way this is done is different from case to case. What they typically all have in common is that the

[36]Computing the test statistic (i.e., the association score) is typically enough for the purposes of collocation or term identification. In actual statistical hypothesis testing, in particular with respect to exact hypothesis tests, the purpose is to compute the significance or p-value of the observed data, which can be interpreted as the amount of evidence provided by the observed data against the null hypothesis. This may be done e.g. by summing over all contingency tables that provide at least as much evidence against the null hypothesis as the observed table (see Agresti (1990)). It is needless to say that computing exact p-values is computationally very expensive.

score assigned to collocation and term candidates is used to rank them (typically in descending order) and thus an explicit ordering according to the degree of (computed) collocativity or termhood is yielded. In general there is a distinction between one-sided and two-sided measures. This depends on whether a measure distinguishes between positive and negative associations (which is the case for one-sided measures) or not (which is the case for two-sided measures).[37] Positive association denotes that parts of a word pair co-occur more often than by chance (i.e. if they were independent), and negative association indicates that they co-occur less often. From this, there follows a correlation with the sidedness of a measure. In the case of one-sided measures, high scores indicate a strong positive association whereas low scores (including negative ones) denote that there is no indication for a positive association (which, however, may mean that components are either independent or negatively associated). On the other hand, for two-sided measures high scores are an indication of any kind of strong association, be it positive or negative, whereas low scores (regardless of the sign) denote (near-)independence. A two-sided measure whose scores are always positive (such as the log-likelihood measure – see below) can be (and should be, for the purpose of computing an association score to generate a ranked list) easily converted into a one-sided measure by changing the sign of the association score. Evert (2005) demonstrates that this may be done in cases when the observed frequency O_{11} is smaller than its expected counterpart E_{11}.[38]

In the following we will outline the most relevant (because most successful) association measures used in various studies for the task of collocation and term extraction. In the case in which it is suitable, also alternative notations and formulas with different parameters (as e.g. used by Manning & Schütze (1999)) will be described, in particular when they are necessary to motivate and derive an extension to n-grams of size larger than two. Two of these measures, t-test and log-likelihood, belong to the class of so-called asymptotic statistical hypothesis tests. The other association

[37] These notions are taken from the area of statistical hypothesis testing where they are also labeled *one-tailed* or *two-tailed*. In general a test is called two-sided or two-tailed if the null hypothesis is rejected for values of the test statistic falling into either tail of its sampling distribution curve, and it is called one-sided or one-tailed if the null hypothesis is rejected only for values of the test statistic falling into one specified tail of its sampling distribution curve – see Agresti (1990) and Evert (2005) for detailed mathematical accounts.

[38] In such a case, $O_{11} < E_{11}$ indicates that there is a negative association between the component parts of word pair.

measure widely used is mutual information (MI) which has to be counted to the class of information-theoretic measures.

3.3.2 T-test

Before actually characterizing the t-test, a close relative of it has to be described, *viz.* the z-score, which also has been used in collocation extraction studies (cf. (Berry-Rogghe, 1973) in subsection 3.1.1). Although based on a discrete binomial distribution, the test statistic (see equation 3.1) converges to a standard normal one for large sample sizes (i.e., large N).[39]

$$z\text{-}score := \frac{O_{11} - E_{11}}{\sqrt{E_{11}}} \tag{3.1}$$

A practical problem with the z-score is that its values may become very large for low expected frequency E_{11} (due to its status as approximate variance in the denominator), which yields highly overestimated scores for low frequency data. For this purpose, Church et al. (1991) suggest using the t-test (also referred to as Student's t-test or t-score) which obtains the variance from the observed frequencies rather than from the expected frequencies under the null hypothesis. The t-test, which is a one-sided hypothesis test based on a normal distribution,[40] may be formalized as follows in terms of observed and expected frequencies (see also Evert (2005)):

$$t\text{-}test := \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \tag{3.2}$$

Evert (2005) argues at length that, from a theoretical perspective, the t-test is not applicable to co-occurrence frequency data because, on the one hand, the null hypothesis states that the sample is drawn from a normal distribution with mean E_{11} whereas, on the other hand, the variance is estimated directly from the sample (i.e.

[39]This approximation, however, is theoretically problematic (see Evert (2005)) but may be dealt with by applying Yates's continuity correction (Yates, 1934) which improves the approximation by adapting the observed frequencies.

[40]In the strict sense, the t-test has a so-called t distribution which, however, approximates to a normal distribution for large enough samples (i.e. for large N).

from O_{11}). So much the more surprising, however, is the fact that it performs quite well in collocation and term extraction tasks (cf. (Evert & Krenn, 2001), (Krenn & Evert, 2001), (Church et al., 1991) – and also in this thesis) and actually better than theoretically more well-founded association measures such as log-likelihood or mutual information.

By focusing on the notion of observed and expected means, Manning & Schütze (1999) (adopting it from Church et al. (1991)) offer a different take on the t-test statistic, which in practice is however numerically fully equivalent to the observed and expected frequency notation.

$$t\text{-}test := \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{3.3}$$

Here, $\bar{x}$ denotes the observed mean and μ denotes the expected mean whereas s^2 and N denote the sample variance and the sample size, respectively. According to Manning & Schütze (1999), $\bar{x}$ and μ may be computed in a straightforward way, *viz.* by scaling the observed frequency and by scaling the expected frequency (under the null hypothesis of independence) by the sample size N.[41] For our previous example (the German PNV collocation "*zu Ende gehen*") outlined in tables 3.3 and 3.4, this would yield the following:

$$\bar{x} = P(zu\ Ende\ gehen) = \frac{freq(zu\ Ende\ gehen)}{N} = \frac{100}{132136} \approx 0.0008 \tag{3.4}$$

In Manning & Schütze (1999), the expected mean μ under the null hypothesis of independence is computed scaling the raw frequencies of each word by the sample size and multiplying them. According to this, the expected mean value for the co-occurrence of "*zu Ende gehen*" would be the following:

$$\mu = P(zu\,Ende) * P(gehen) = \frac{freq(zu\ Ende)}{N} * \frac{freq(gehen)}{N} = \frac{250}{132136} * \frac{1977}{132136} \approx 0.00003 \tag{3.5}$$

[41] This may also be interpreted as using maximum likelihood estimates to obtain probabilities from a probability function.

Because the standard deviation s^2 is quite difficult to determine in practice, most researchers using Church et al. (1991)'s take on the t-test (including Manning & Schütze (1999)) approximate s^2 by the sample mean $\bar{x}$ (which is the observed frequency scaled by the sample size). Plugging in the values into equation 3.3 yields a t-score of approximately of 9.517, which in effect is the same as if we compute it according to equation 3.2.

3.3.3 Log-Likelihood

A rather different kind of test statistic is given by the so-called log-likelihood (also often referred to with the symbol G^2), which is based on the asymptotic χ^2 distribution. It is actually the fact that the underlying sampling distribution is not normal but asymptotic that made Dunning (1993) vehemently promote the log-likelihood test as the *accurate* test statistic for natural language data which may show highly skewed distributions (in opposition to other test statistics which assume a normal sampling distribution). Although the actual χ^2 test (see Manning & Schütze (1999) for an account) may be the better test for independence in mathematical statistics, Dunning (1993) pointed out that the situation is different for natural language data (which exhibits highly skewed contingency tables) and thus the log-likelihood test should be preferred. In fact, Evert (2005) shows at great length through numerical simulation that the log-likelihood statistic turns out to be the most accurate and convenient measure for the significance of association because it best approximates the exact p-values of Fisher's exact test, which is considered to be the prototype of a truly exact hypothesis test (Fisher, 1922).[42]

The actual log-likelihood test statistic derived by Dunning (1993) (and presented in Manning & Schütze (1999)) is both awkward and unintuitive and thus we will formalize it along the way suggested by (Evert, 2005), *viz.* by means of the observed and expected frequencies of a 2 x 2 contingency table:[43]

$$2 \sum_{ij} O_{ij} \; log \frac{O_{ij}}{E_{ij}} \tag{3.6}$$

[42]Other test statistics only compute approximations of their p-values, which may only valid for large enough samples.

[43]Below, the natural logarithm of the fraction is taken.

It can be seen that, apart from the underlying sampling distribution, what makes the log-likelihood different from the other test statistics considered here (and what makes it similar to its χ^2 relative) is that all cells of the contingency table are taken into its computation. For the t-test, z-score and mutual information (see below) only the observed and expected co-occurrence frequencies O_{11} and E_{11}. are considered. Another point of difference is the two-sidedness of the test statistic which means that high scores may indicate either kind of strong association, be it positive or negative. Fortunately, since the log-likelihood test only yields positive scores, this unpleasant effect (i.e. at least for the task of collocation and term extraction resulting in a ranked output list) may be reversed by converting it to a one-sided test. This is done by changing the sign of the scores for candidates that exhibit a negative association, which in the logic of 2 x 2 contingency tables are those for which the observed co-occurrence frequency is smaller than the expected one, i.e. $O_{11} < E_{11}$.

3.3.4 Mutual Information

An association measure motivated by information theory (Shannon, 1948; 1951; Fano, 1961; Cover & Thomas, 1991) is mutual information (MI) which is standardly (i.e., information-theoretically) defined as holding between two random variables. The way, however, MI is used for the task of collocation and term extraction (Church & Hanks, 1989; 1990; Church et al., 1991; Church, 1995; Daille, 1996; Manning & Schütze, 1999) is rather different as mutual information is typically taken to hold between two random variables instead of their *values*, as it is applied in NLP. In concrete, this is referred to as *pointwise* mutual information (PMI) which measures the overlap between two particular events x and y such that the ratio between their observed joint probability $P(X \cap Y)$ and their independent (i.e. expected) probability $P(X)P(Y)$ is simply taken (and the binary logarithm is applied to make it conform to information-theoretic requirements).

$$I(x, y) = log_2 \frac{P(xy)}{P(x)P(y)} \tag{3.7}$$

In the notational language of observed and expected frequencies (Evert, 2005), MI may be then formalized along the following lines:

$$MI = log_2 \frac{O_{11}}{E_{11}} \tag{3.8}$$

One of the major problems observed with MI is, like in the case of the z-score, an overestimation bias for low-frequency events, i.e. bigrams composed of low-frequency words will receive a higher score than bigrams composed of high-frequency items (Manning & Schütze, 1999), which is of course contrary to what a good association measure should accomplish. For this reason, various more or less well-motivated heuristic extensions have been proposed and used (e.g. Hodges et al. (1996)), most of which attempt to increase the impact of the co-occurrence frequency, typically on its numerator. For example, in order to increase the impact of co-occurrence for MI, Daille (1994) experiments with various exponents in the numerator (i.e. MI^k with $k = 2...10$) and heuristically finds (and determines) $k = 3$ to yield the best result for the task of term extraction (cf. also subsection 3.2.2).

$$MI_{Daille} = log_2 \frac{(O_{11})^3}{E_{11}} \tag{3.9}$$

3.3.5 Extensions to Larger-Size N-Grams

Basing test statistics on a 2 x 2 contingency table, although theoretically the soundest as well as the most intuitive and elegant way, may quickly come to the limits as soon as the linguistic structure of collocations and terms goes beyond the well-defined scope of word bigrams. Although Justeson & Katz (1995) note that roughly two thirds of terms are two-word combinations, the other one third also needs to be accounted for. In the case of collocations, the picture may look similar. Although no comparable study in the vein of Justeson & Katz (1995) has been undertaken, a look at any (English or German) collocation dictionary, e.g. (Dudenredaktion, 2002) or (Benson et al., 1997) – however incomplete they may be (as noted e.g. by Lin (1999)) – reveals that there are larger collocational units that go beyond word bigrams. Admittedly, many multi-word (i.e. larger size n-gram) collocations may be collapsed to bigrams (which is indeed a common practice exactly because of the necessity to work with bigrams), such as in the case of German preposition-noun-verb collocations in which

the preposition and the noun are collapsed into one unit (e.g. in Krenn & Evert (2001) and also in this thesis). Still however, also NLP researchers working on the task of collocation and term extraction are aware of the fact that the (linguistic) world (of collocations and terms) does not only consist of bigrams.[44] Therefore, there have been extensions to association measures proposed, which however are mostly heuristically motivated rather than theoretically well-founded.

There is one decisive criterion that an association measure must fulfill in order to qualify for such a potential extensibility to larger-size n-grams, i.e., it must be possible to define its test statistic alternatively to and independently of a 2 x 2 contingency table. As a matter of fact, such an independent definition is only possible for those measures which only consider the observed and the expected co-occurrence frequencies, i.e. O_{11} and E_{11}, because these parameters may also be computed from maximum likelihood estimates yielding sample means and expected (distribution) means, i.e. $\bar{x}$ and μ.

In this vein, computing $\bar{x}$ for the t-test (see subsection 3.3.2 above) may be easily extended to a trigram with the particular events a,b, and c (N again denotes the sample size):

$$\bar{x} = P(abc) = \frac{freq(abc)}{N} \tag{3.10}$$

Analogously, the same may be done for μ:

$$\mu = P(a) * P(b) * P(c) = \frac{freq(a)}{N} * \frac{freq(b)}{N} * \frac{freq(c)}{N} \tag{3.11}$$

Then, all that is left to do is to plug in these computations into the equation given for the t-test (i.e. into equation 3.3). In a parallel vein, a trigram extension to the MI association measure may proceed along the following lines:[45]

[44] Consider, for example, the more complex structural types of collocations such as preposition-noun-noun-verb or noun-noun-verb, to name just a few.

[45] This trigram extension to the MI measure has actually been proposed by Alshawi & Carter (1994) and Lin (1999) whereby Lin uses conditional instead of joint probabilities because the trigram MI measure is run on dependency triple outputs (see subsection 3.1.3) for which of course the independence assumption may not be motivated at all.

$$MI_3 = log_2 \frac{P(abc)}{P(a)P(b)P(c)} \tag{3.12}$$

As can be also seen from the above two equations, these two association measures may still be further extended to larger-size n-grams (e.g. to quad- or pentagrams).

The picture on extensibility looks quite different with respect to the log-likelihood measure presented in subsection 3.3.3. As can be seen from equation 3.6, the computation of the log-likelihood test statistic is *inherently* tied to all four cells of a contingency table. As a consequence, neither an extension based solely on sample and expected means is possible nor is there any other well-defined way to compute the other cells. Admittedly, one could attempt to collapse a trigram into a bigram but then the immediate question arises which two of the three component parts should undergo this procedure. Whereas it may be clear for the case of collocational preposition-noun-verb combinations, it is completely obscure in the case of trigram (or even higher-order n-gram) terms within noun phrases. Hence, it has to be concluded that the theoretically most well-founded statistical association measure is not extensible beyond the bigram scope in a well-defined way.[46]

3.3.6 Shortcomings and Linguistic Filtering

Besides the issue of (non-)extensibility of an association measure, the previous three subsections have shown that, at least for the NLP tasks of collocation and term extraction, there is no one-to-one correspondence between the statistical soundness of an association measure, on the one hand, and a corresponding superior extraction performance, on the other hand. This may be evidenced by the fact that, for example, Evert & Krenn (2001) and Krenn & Evert (2001) (see subsection 3.1.4 above) report that it is actually the t-test, next to co-occurrence frequency, which performs best for the task of collocation extraction. In a similar vein, Daille (1996) reports that, for the task of term extraction, co-occurrence frequency performs equally well to log-likelihood, the theoretically most well-founded association measure (according to Evert (2005)), and even better than the information-theoretic mutual information measure (see subsection 3.2.2 above).

[46]The same holds e.g. for the χ^2 measure (see Manning & Schütze (1999)) whose computation is also tied to all four cells of a contingency table.

These observations seem to point to a general problem with natural language text as sample data both for statistical hypothesis testing and for information-content measures. Going back and re-examining the statistical considerations outlined in subsection 3.3.1 does indeed reveal some of the discrepancies between the assumptions made by statistical and information-theoretic models, on the one hand, and their correspondence in natural language text data, on the other hand. One fundamental premise made is the independence of word combinations (or more formally: random variables) as a default assumption, either with respect to a null hypothesis for the case of statistical hypothesis tests or with respect to the mutual information content for information-theoretic measures. This assumption is of course highly unrealistic for natural language data, and at best a necessary idealization in lack of a better model. Still, this property (or better: non-property) of natural language is of course known to NLP researchers working on collocation and term extraction and hence there is one major heuristic to at least approximate the independence assumption, *viz.* linguistic filtering or (pre-)processing. Strictly speaking, a true violation of the independence assumption for natural language data only occurs if unrestricted word sequences (in a text) are considered (and assumed to be independent), i.e. word sequences where no a priori linguistic structure is assumed. Of course, the actual probability of any such word sequence is strongly affected by the fundamental structure of natural language (be it e.g. grammatical or semantic or both) and thus is diametral to the notion that any word may be associated with any other word in an unrestricted manner. Applying a linguistic filter on natural language text data, such as a part-of-speech (POS) tagger, a phrase chunker or even a syntactic parser (see e.g. the approaches described in sections 3.1 and 3.2), creates a subset of collocation or term candidates to which association measures may be applied. One major effect of creating such a subset is that the independence assumption may be taken to be much more valid. This is because if the universe of statistical possibilities is reduced to sequences where a preposition and noun co-occur together with a verb, the null hypothesis that the co-occurrence of this sequence is due to chance turns out to be a much more accurate assumption. Hence, linguistic preprocessing is not only a mere structure-adding operation but also helps to make natural language data more "statistics-ready" for lexical association measures.

Another discrepancy between the assumptions made by statistical and information-theoretic models and their correspondence in natural language text data

is that most of the test statistics assume a normal distribution, or at least a distribution which may not be assumed for natural language (e.g. the χ^2 distribution for the log-likelihood measure – see subsection 3.3.3). In this sense, the test statistics introduced here so far may all be described as *parametric*.[47]

One of the main observations made about frequency distributions for natural language is, however, that they tend to be highly skewed. The most prominent illustration of this may be given by the famous frequency distribution known as *Zipf's law* (Zipf, 1935; 1949). By counting how often each word type occurs in a text corpus and then listing them in the order of their occurrence frequency, the relationship between between the frequency of a word f and its its position in the ranked order, i.e. its rank r, may be determined. This "law" may be stated in the following way (adopted from Manning & Schütze (1999)):

$$f \propto \frac{1}{r} \tag{3.13}$$

What this means is that there is a constant k such that $f * r = k$. Hence, e.g., the 50th most common word in a corpus sample will occur with three times the frequency of the 150th most common word. Still, despite its appearance, what Zipf (1949) states as a "mathematical law" may be rather described as a roughly accurate characterization of certain empirical facts about words. It is actually Mandelbrot (1954) who achieves a closer fit to the empirical distribution of words by deriving a more general (but similar) relationship between frequency and rank.

3.3.7 Frequency of Co-Occurrence

Given that there appear to be substantial discrepancies between the assumptions made by test statistics-inspired lexical association measures and the actual properties of natural language text, the question arises whether the most simple and obvious lexical association measure – frequency of co-occurrence – may not provide a clear-cut and performant manner to extract collocations and terms. Indeed, in particular for many linguistic definitions of collocations (cf. the contextualist tradition outlined in

[47] To be more accurate, all test statistics that estimate population parameters are parametric in that they assume that the distributions of their variables belong to established parameterized classes of probability distributions.

subsection 2.1.2,[48] but also e.g. van der Wouden (1997)) and also linguistic definitions for terms (cf. subsection 2.2.7 above), their descriptors often contain such terms as "frequent co-occurrence", "recurrent co-occurrence", "habitual co-occurrence", or "typical co-occurrence".

As a matter of fact, several studies on both collocation extraction (e.g. Evert & Krenn (2001), Krenn & Evert (2001)) and on term extraction (e.g. Daille (1996)) have shown that the performance of frequency of co-occurrence is at least on par with more complex statistical association measures, such as the t-test or log-likelihood. Other ones, such as Justeson & Katz (1995) even solely rely on frequency counting to extract terms from domain-specific texts.

Formally, by means of the notations used for 2 x 2 contingency tables, frequency of co-occurrence (freq) may be rendered by the joint observed frequency.

$$freq = O_{11} \tag{3.14}$$

Alternatively, with the help of the sample size N we may use maximum likelihood estimates to obtain probabilities from a probability function, viz. the joint probability for two events x and y.[49]

$$freq = P(xy) = \frac{freq(xy)}{N} \tag{3.15}$$

In this way, it is of course also clear that frequency of co-occurrence is not restricted to n-grams of a certain size (i.e. to bigrams).

3.3.8 C-value

The C-value measure, as described in subsection 3.2.3, is virtually a heuristically motivated modification of the frequency-based measure, which incorporates the presence or absence of nested candidate terms as well as the length of the candidate string as

[48] The notion of frequency of co-occurrence with respect to collocations is pervasive in different forms in the frequentist and empiricist tradition of British contextualism, e.g. in Firth (1957)'s recurrence criterion and its extension in Halliday et al. (1965).

[49] This notation is also referred to as relative frequency (Manning & Schütze, 1999).

additional parameters into its computation. Frantzi et al. (2000) formalize it for a term candidate a in the following way:

$$\text{C-value}(a) = \begin{cases} log_2|a| * freq(a) & \text{if } a \text{ is not nested} \\ log_2|a| * (freq(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} freq(b)) & \text{otherwise} \end{cases}$$

Here, T_a is the set of extracted candidate terms that contain a, $P(T_a)$ is the number of these candidate terms. It is clear that the C-value is a measure on the frequency of co-occurrence of candidate term a. The negative effect of a being a substring of other longer candidate terms is caused by the negative sign in front of $\sum_{b \in T_a} freq(b)$. The independence of a from these longer terms is yielded by $P(T_a)$. Having $P(T_a)$ as the denominator of the negatively signed fraction reflects the fact that the greater this number is, the bigger is its independence (and vice versa). In addition, the candidate term length $|a|$ is also factored into the computation of the C-value. The positive effect of this candidate term length is restrained by applying the binary logarithm on it.

Chapter 4

Linguistically Enhanced Statistics To Measure Lexical Association

The last section 3.3 in the previous chapter has shown that standard statistical and information-theoretic association measures possess certain properties in their underlying statistical assumptions which may turn out to be diametral to the properties of natural language text data. Among these is the fact that many test statistics either assume a normal distribution or distributions which do not reflect the highly skewed distributional properties of natural language text. Another unrealistic assumption made by virtually all test statistics, in order to be able to compute their association scores, is the assumption that the co-occurrence (or combination) of one word with another one, as a default at least, tends to be independent, and hence any statistical evidence to the contrary of this independence assumption is taken to increase the association strength between such words. This assumption may be at least corroborated by employing some degree of linguistic filtering which creates a subset of collocation or term candidates for which the adequacy of the independence assumption is more appropriate.

With respect to this, it also has to be mentioned that there have been two association measures presented which fall outside the class of parametric test statistics or information-theoretic measures which dominated section 3.3, *viz.* frequency of co-occurrence (see subsection 3.3.7) and the C-value (see subsection 3.3.8), which may be described as a heuristically modified version of frequency of co-occurrence. There are actually two interesting observations which have to be pointed out for

these two measures. First, unlike some test statistics, both association measures are not confined to bigrams but may be easily applied to n-grams of any size, giving them a degree of extensibility which some other test statistics lack (e.g. log-likelihood). Another finding, reported by several studies, is that the extraction performance of frequency of co-occurrence, both for collocations (Evert & Krenn, 2001; Krenn & Evert, 2001) and for terms (Daille, 1996), appears to be on par with statistical and information-theoretic measures. Such a kind of finding is interesting insomuch as, if it may be empirically confirmed further, it could have the potential to call into question the necessity to employ statistical or information-theoretic association measures in the first place. The reason for this is that frequency of co-occurrence counting is of course computationally less expensive than applying numerically much more elaborated association measures, to which single types of various frequency counts and estimations (e.g. observed and expected frequencies – see subsection 3.3.1) are only the input to complex association score computations.

Considering the fact that statistical and information-theoretic lexical association measures make assumptions which fall outside the properties of natural language and considering the fact that there is some empirical evidence that they do not appear to outperform mere frequency of co-occurrence counting in a substantial way, the question arises whether there are procedures which are more suitable to the properties of collocations and terms in order to measure their lexical association and which, in this way, are able to deliver more substantial results in extracting them from text. A way to phrase this question slightly differently would be: if standard statistical assumptions and properties are not sufficient to measure lexical association for collocations and terms, are there any *linguistic* properties which may be more suitable for these tasks? After all, in the linguistics and terminology literature, there have been many accounts on various linguistic properties of collocations and terms proposed (see the discussions in chapter 2 – in particular, sections 2.1 and 2.2 and the the assessment in section 2.3). Hence, in this chapter we shall present two statistical procedures which take into account linguistic properties of collocations and of terms in order to measure their lexical association. In order to be able to soundly derive these two procedures, we have to formulate both their statistical and their linguistic requirements. On the statistical side, we have to make sure that we do not make any assumptions which run contrary to the properties of natural language in general as well as collocations and terms in particular. These requirements will be presented

in section 4.1. On the linguistic side, we have to ensure that we utilize observable properties which are suitable to formalization and quantification in a such manner that they may be used as input parameters to statistical computations. We will see that there are such properties and that their linguistic underpinnings may be traced back to the lexical-collocational layer of Firth's (1957) model of language presented in section 2.1.2, in particular to its notion of syntagmatic and paradigmatic context. These linguistic requirements will be presented in section 4.2.

From these two kinds of requirements we will present two new linguistically motivated approaches to statistically measure lexical association for collocations and for terms. For the case of collocation extraction, we propose a lexical association measure based on the linguistic property of *limited syntagmatic modifiability* (section 4.3) and for the case of term extraction, we propose a lexical association measure based on the linguistic property of *limited paradigmatic modifiability* (section 4.4). Lastly, in section 4.5 we will also extensively lay out the requirements for constructing an extensive testing ground in order to thoroughly evaluate both measures, in particular in comparison to the frequency-based, statistical and information-theoretic approaches which have been proposed in the computational linguistic literature on collocation and term extraction from natural language text.

4.1 Statistical Requirements

The statistical requirements which have to be put forth in order to formulate linguistically motivated statistical measures for lexical association have to take several aspects into account. As we have already elaborated on previously, it is essential that such an association does not make any assumptions that run contrary to the properties of natural language text data (subsection 4.1.1). Furthermore, extensibility to n-grams of size larger than two has to be granted (subsection 4.1.2). Co-occurrence frequency as a factor needs to be included as it has surfaced prominently in the discussions on the linguistic properties of collocations and terms throughout this thesis (subsection 4.1.3). Finally, a lexical association measure is inherently bound to computing some sort of association score which in turn yields a ranked output on (collocation or term) candidate sets. In subsection 4.1.4, we will explain both the prequisites and the effects of such a ranking property.

4.1.1 Avoidance of Non-Linguistic Assumptions

One major flaw of the standard statistical and information-theoretic association measures is that they make certain assumptions about the distributional properties of their natural language sample data (e.g. that it is normally distributed) which may not be warranted in the light of the highly skewed nature of natural language distributions. Thus, one crucial requirement is that a linguistically motivated association measure be *non-parametric*. In a strict statistical sense, non-parametric statistical models differ from parametric ones in that the structure of the model is not specified in advance but is instead determined from the data. The term "non-parametric" is not intended to indicate that such models completely lack parameters but that the number and nature of the parameters are flexible and not fixed in advance. Here, we do not interpret non-parametricity in a strict statistical sense such that we would formulate a non-parametric (or distribution-free) inferential statistical method as a mathematical procedure for statistical hypothesis testing. Rather, we define it in a broader and more procedural manner such that a linguistically motivated statistical association measure needs to refrain from making any assumptions on the distributional properties of the (language) sample. In addition, such a measure also needs to avoid any sort of statistical hypothesis testing because this, as a default, always computes lexical association scores with respect to some linguistically unrealistic assumption about the null hypothesis, which, in this case, is the independence of word combinations.

There is one qualification which has to be made with respect to the desired exclusion of a linguistically unrealistic assumption about the independence of word combinations. As already laid out in subsection 3.3.6, this assumption may be at least approximated by applying linguistic filters (e.g., in the form of part-of-speech taggers and/or phrase chunkers) which, in turn, generate a subset of candidates for which the independence assumption may be taken to be much more valid. Still, the necessity of linguistic preprocessing may also clearly be motivated by the mere fact that both collocations and terms, by default, do (unquestionably) possess linguistic structure. Thus, collocations may be manifested, e.g., as preposition-noun-verb, noun-verb combinations etc. whereas terms are typically manifested within noun phrases (see subsection 2.2.7). Hence, it is already for this syntactic reason – and thus independent of any statistical considerations about independence assumptions – that some form of

linguistic preprocessing of collocation and term candidates needs to be applied.

4.1.2 Extensibility of Size

Another essential requirement for a linguistically motivated statistical association measure is that it be able to extract n-grams of sizes larger than two. We have already seen that only statistical association measures which exclusively take into account the observed and the expected co-occurrence frequencies are capable of being extended to larger-size n-grams (such as the t-test – see subsection 3.3.5). On the other hand, the mathematically most well-founded statistical association measure, log-likelihood, is not extensible beyond the bigram scope in a well-defined way and is thus – besides the problematic statistical assumptions being made – even less suitable with respect to the requirements for a linguistically adequate association measure. It should be mentioned here that extensibility beyond the bigram scope is particularly important with respect to the extraction of domain-specific terms. As examined by Justeson & Katz (1995) (subsection 2.2.7), approximately one third of multi-word terms are larger than bigrams (i.e. mostly trigrams and (a smaller amount of) quadgrams). A term extraction measure which is not capable of recognizing such larger-sized units would certainly miss a substantial proportion of terms in a text collection. With respect to general-language collocations, although it is in principle equally desirable to have an extensible measure, their syntactic surface manifestation is of course not merely confined to noun phrases (like it is almost exclusively the case for terms) but a wider variety of syntactic patterns. Related to this is another particular difference between domain-specific terms and general-language collocations: whereas it is safe to leave out stop words or stop POS tags (such as determiners, quantifiers, pronouns) for the recognition of terminological noun phrases (and thus from considerations about the length or size of terms), such function words may be integral parts of collocational expressions.[1] Therefore, for these reasons the concept of "size of a collocation" is much harder to define in linguistic theory and hence of course even harder to determine in NLP practice.

[1]For example, consider the English collocational support verb construction "*to come to* ***an*** *end*", which contains an indefinite article.

4.1.3 The Frequency Of Co-Occurrence Factor

Several subsections in the last chapter have introduced statistical and information-theoretic lexical association measures of salient computational complexity, such as the t-test (see subsection 3.3.2), log-likelihood (see subsection 3.3.3) and mutual information (see subsection 3.3.4). At the same time, however, we have also noticed that frequency of co-occurrence (see subsection 3.3.7) may turn out as a viable and computationally less expensive alternative. This is corroborated by the fact that several studies (Evert & Krenn, 2001; Krenn & Evert, 2001; Daille, 1994) reported that the performance of frequency co-occurrence is at least on a par with other statistical association measures, for the task of collocation and term extraction (see subsections 3.2.2 and 3.1.4). Moreover, in addition to some of its statistical counterparts, co-occurrence frequency counting has no restrictions as for the size or length of the collocation or term candidates considered. This is one of the major reasons, why Frantzi et al. (2000)'s C-value introduced in subsection 3.3.8 is mainly defined as a heuristic modification of frequency of co-occurrence counting, *viz.* to provide a length-independent measure for the extraction of terms.

Hence, for these reasons, a linguistically motivated statistical measure of lexical association (and in fact any measure of lexical association) needs to factor in (observed) frequency of co-occurrence, at least to some degree. And in fact, already all statistical and information-theoretic measures fulfill this requirement in various ways through incorporation of the observed joint frequency O_{11}, as can be witnessed in their formal representations in the notational language of 2 x 2 contingency tables presented in section 3.3. The potential of co-occurrence frequency is also corroborated by linguistic research on collocations, in particular in the vein of British (Neo-)Firthian contextualism (also referred to as the *frequentist or empiricist tradition*[2] – see subsection 2.1.2), in which it is most prominently expressed in Firth's (1957) recurrence criterion.

4.1.4 Output Ranking

In subsection 3.3.1, we extensively discussed the statistical foundations of statistical, information-theoretic and frequency-based lexical association measures. One characteristic which they all have in common is the computation of some sort of association *score* from their input parameters. From a statistical perspective, such a score – at

[2]Notice the term "frequentist" in the descriptor of this linguistic research tradition.

least for the statistical association measures – is an indication to which degree the null hypothesis of independence may be rejected or not.[3] Since its premise is a set of collocation and term candidates which are derived by some form of linguistic filtering, the computation of such an association score for each candidate has a major effect on the output of collocation or term extraction procedures in that an explicit ranking of the candidates may be carried out, resulting in a ranked output list.

From a linguistic perspective, such a ranking is not as far off as it might appear at first sight. On the contrary, it may well be interpreted as the assignment of different degrees of collocativity (to collocation candidates) or termhood (to term candidates). This, in turn, makes sense in the light of the fact how the linguistic status of an expression being a collocation or a term is perceived by humans. As a matter of fact, terminologists, for example, do not always agree on whether a given expression constitutes a term or not. This observation has been stated both by (theoretical) terminologists (Wüster, 1979; Cabré Castellví, 2003) and by researchers on automatic term extraction (Frantzi et al., 2000; Daille, 1994) independently. From the terminological side, such dissense is reflected by the fact that typically a large body (committee) of domain experts (see also subsection 2.2.1) convenes periodically to decide on inclusions of new entries into major terminological resources, such as e.g. the Unified Medical Language System (UMLS)[4] (see subsection 4.5.3.3 for a description of it). In a similar vein, for a human to judge whether a given linguistic expression (e.g. a preposition-noun-verb combination) constitutes a collocation or not may be not as straightforward to decide as it may appear at first glance, and hence, there are various degrees of inconsistencies regarding such a judgement – depending on the type of linguistic classification asked for. We will return to this issue in subsection 4.5.2.3 where we introduce and explain our experimental test collection for collocation extraction.

Coming full circle again to the issue of assigning association scores to collocation and term candidates, their resulting ranking thus indicates the *confidence* with which an extraction procedure (in particular, the underlying lexical association measure)

[3] In exact statistical hypothesis testing, it would rather be the resulting p-value which provides the actual counter-evidence against the null hypothesis. For the purpose of collocation and term identification in natural language text, this additional (cost-intensive) computational step is not essential (see subsection 3.3.1).

[4] `http://www.umlsinfo.nlm.nih.gov`

determines whether or not and to what degree a given candidate actually constitutes a collocation or term.

In an ideal system, then, the output of an association score-based ranking procedure would naturally be such that the following two conditions are met:

- The true collocations or terms (i.e., the targets) are ranked in the upper portion of the output list.
- The non-collocations or non-terms (i.e., the non-targets) are ranked in the lower part of the output list.

From such a ranked output list, then, the performance quality of an association measure may be determined in different ways – depending on the size and completeness of the output list – ranging from merely counting the targets among the top n ranked candidates (with n ranging from 50 to several hundreds) to applying sophisticated performance evaluation metrics which are well established in the information retrieval community, *viz.* precision and recall. One major advantage of the fact that the output of lexical association measures typically transforms a set of collocation and term candidates into a ranked output list is that it makes the performance quality of these measures easily comparable to each other. Thus, the requirement for a linguistically enhanced association measure to produce such a ranked output may not only be motivated from a linguistic perspective (i.e., different degrees of termhood or collocativity) but also from a comparative evaluation perspective. The issues and concerns with respect to a suitable evaluation platform for lexical association measures for the tasks of collocation and term extraction will be discussed extensively in section 4.5.

4.2 Linguistic Requirements

If we want to formulate the requirements for a linguistically motivated statistical association measure, we have to step back and recapitulate those linguistic properties of collocations and terms which have the potential to serve as observable properties suitable to formalization and quantification in such a manner that they may be used as input parameters to statistical computations. For both collocations and terms, linguistic properties have been identified on the syntactic and on the semantic level

which distinguish them from other linguistic expressions (see section 2.3 above). Because of several reasons to be laid out below, in this thesis, we will focus on the linguistic property of *limited modifiability* found on the syntactic layer, which holds for both collocations and terms from a different perspective for each. What makes this property especially suitable is the fact that it can be aggregated within an explicit linguistic frame of reference, *viz.* the collocational (or lexical) layer of Firth's (1957) model of language description, which will be be recapitulated in the next subsection 4.2.1. Subsection 4.2.2 then will lay out the linguistic requirements which the property of limited modifiability has to meet, in particular why it is within the syntagmatic context (in Firth's model) in which it has to isolated, in order to be suitable for the task of collocation extraction. In addition, we also discuss potential alternative linguistic properties and give reasons why they are not as well suited for a linguistically motivated association measure for extracting collocations. In a similar vein, subsection 4.2.3 will establish the corresponding requirements which the property of limited modifiability has to meet to be incorporated into a linguistically enhanced statistical association measure for the task of extracting terms. Again, we will substantiate in detail why for terms it is the paradigmatic context (in Firth's model) in which this property has to be located.

4.2.1 Firth as Linguistic Frame of Reference

The model of language description laid out by Firth (1957), and in particular its lexical-collocational layer, – see subsubsection 2.1.2.2 above – must of course not be confused with a formal or even mathematical model of language (as e.g. Harris (1968) attempts to formulate). It is rather an attempt to formulate a linguistic context as a frame of reference for isolated words (or sentences), and from a current linguistic perspective, it may certainly be seen as a simplification of linguistic context structure.[5] Still, the lexical-collocational layer of Firth's model (repeated for convenience in figure 4.1) may be taken as an appropriate linguistic frame of reference in which the linguistic property of limited modifiability may be neatly configured, both for collocations and for terms.

As can be seen, the main feature of Firth's lexical-collocational layer is the di-

[5]It should be noted, however, that simplification, in particular when combined with abstraction, is a legitimate step in building a model.

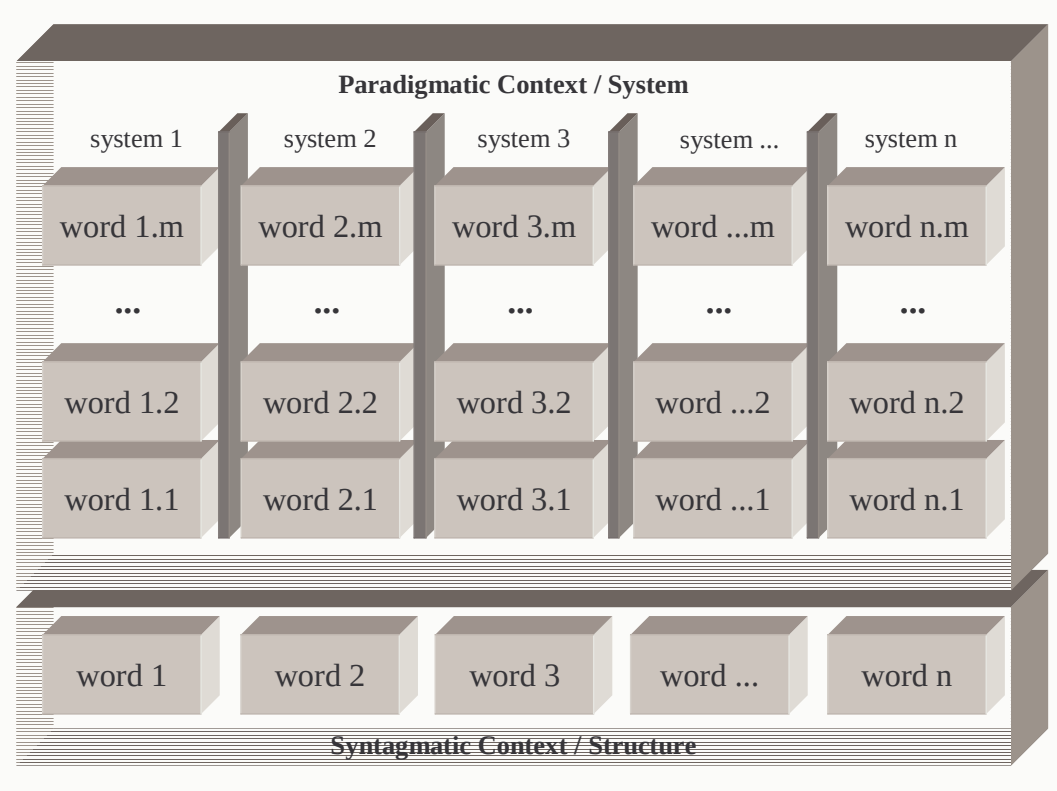

Figure 4.1: The lexical-collocational layer of Firth's model of language description.

vision into a *syntagmatic* and a *paradigmatic* context of text words. In particular, the syntagmatic structure of a text results from a sequence of (subsequent) words whereas the paradigmatic structure is derived by their empirically determined possible substitutions. Although Firth points out that collocations are word occurrences in the syntagmatic context constituted of two or more words, it remains unclear how their boundaries are determined within a text. This problem, however, can be easily overcome by allowing for some syntactic preprocessing which provides the linguistic structure (e.g. part of speech and/or phrasal elements) from which the boundaries for collocations and terms may be determined.[6] This approach is not only in line

[6]Not filtering collocation or term candidates by some form of linguistic preprocessing is not so unusual as it might appear at first sight. As already described in subsection 3.1.2, Smadja (1993)

with common linguistic understanding in that every form of a linguistic expression does have some form of underlying syntactic structure, but it is also compliant with Firth's model of language description which actually consists of four descriptive layers (see subsection 2.1.2.1) and in which the lexical-collocational layer is *on top of* the syntactic one.

As we will see in the following subsections, Firth's lexical-collocational layer provides an appropriate linguistc frame of reference to formalize the notion of limited modifiability for both collocations and terms in the syntagmatic and the paradigmatic context, respectively. This, in turn, will also enable us to emphasize the fact there are linguistic differences between collocations and terms because, after all, collocations may be better defined as general-language constructs surfacing in a variety of syntactic constructions whereas terms rather fall into the class of domain-specific sublanguage constructs and are confined to noun phrases (see also the discussion in subsection 2.3 above). Such differences are inherently ignored by the statistical and information-theoretic association measures which have been most widely used for the extraction of collocations and texts from natural texts (see section 3.3).

4.2.2 Linguistic Requirements for Collocations

For collocations, the syntactic property of non- or limited modifiability has been orginally framed within the lexicographic approach to collocations (Benson, 1989), and is picked up by Manning & Schütze (1999) in describing the linguistic characteristics of collocations for a computational linguistics audience. On a coarse-grained level, this property states that many collocations cannot be freely modified by additional lexical material (see subsection 2.1.4 above). On such a level, this remains a very blurry definition which is not further laid out (or even formalized) by Manning & Schütze (1999) or by Benson (1989). In order to arrive at a more precise formulation (which will be introduced in section 4.3 below), we have to first identify where to locate modifiability on the syntactic level for collocations. Typically the notion of "additional lexical material" may be best placed on the phrasal level on which we define a phrase consisting of a head and a set of potential modifiers (i.e., the additional lexical material). Hence, the head of a noun phrase (NP) is typically a noun, the head

first attempts to identify a set of collocation candidates from (linguistically) unfiltered text and only then submits them to a linguistic filter (i.e. a POS tagger) as a sort of syntactic validation procedure.

of a verb phrase (VP) a verb, etc. Modifiers may fall under a wide range of part of speech patterns, ranging from adjectives and adverbs to determiners and numbers, to name just the most canonical ones. Because collocations may syntactically surface as a possible combination of all of these phrases, a first necessary step is to identify phrasal patterns for collocations. In most studies (see section 3.1), this is done by filtering out certain POS patterns (e.g. preposition-noun-verb). At the phrasal level, however, such patterns should be defined in a slightly more coarse-grained way, e.g. as prepositional-phrase (PP)-verb (or preposition-NP-verb) patterns.[7] Then, modification with additional lexical material may be defined as the addition of such material in front of the head of a phrase, e.g. the kinds of determiners and/or adjectives placed in front of a noun. Crucially, an *addition operation* (of lexical material) may also be described as a *syntagmatic operation* in the syntagmatic context, with help of the linguistic model laid out by Firth (1957). Then, if the linguistic property states that modifiability of collocations is not given or rather limited, we may term this property as *Limited Syntagmatic Modifiability* – or in short: LSM. We will derive this linguistic property formally in subsection 4.3.1 below.

At this point, the question may be raised whether the two other linguistic properties of collocations previously outlined in subsection 2.1.4,[8] *non- or limited compositionality* and *non- or limited substitutability*, may not be equally (or even more) suitable to be included into a linguistically motivated statistical association measure. The main difference between LSM and non- or limited compositionality is that, whereas LSM is clearly a syntactic property, non- or limited compositionality is clearly a semantic property. The problem with the latter one is that it does not equally hold for collocations in general but is rather prominent in one specific subtype, *viz.* idiomatic phrases (see the discussion in subsubsection 2.1.4.2). The other two subtypes of collocations, i.e. support verb constructions and fixed phrases, are characterized by their varying degrees of and contributions to semantic compositionality between their lexical constituent parts. Hence, deriving an observable quantification of such a property would actually first require to generate a high-quality set of collocation candidates which then may be further classified into various subtypes. This, however,

[7] In a similar vein, NP-verb patterns or adjective-phrase (AdjP)-noun patterns may be filtered out.

[8] Besides frequency of co-occurrence, of course, which is already taken for granted to play an important role.

is exactly the approach taken by Lin (1999) and Lin (1998b) (as described in detail in subsection 3.1.3 above) who first compute such a set of candidates by applying the log-likelihood association measure to them (see subsection 3.3.3) and, then, in order to arrive at a more fine-grained classification, attempt to identify the non-compositional phrases among them. It is also Lin who actually incorporates the other linguistic property, i.e. non- or limited substitutability, into the semantic classification of collocation candidates, namely by testing whether their component parts are substitutable with (near-)synonymous words from a thesaurus. Thus, it can be seen that non- or limited compositionality and substitutability are more or less two sides of the same (semantic) coin. Another reason in favor of a syntactic property (instead of a semantic one) as an integral part of a linguistically motivated association measure to generate collocation candidates is that, from a canonical perspective on the different linguistic layers, syntactic processing typically feeds into semantic interpretation, and thus including a syntactic property appears to be the more natural option.

4.2.3 Linguistic Requirements for Terms

In formulating a linguistically motivated statistical association measure for the task of term extraction, the question at what linguistic level – e.g. syntactic or semantic – such a suitable quantifiable property should be determined may follow along the same lines as in the case of collocation. Although they may exhibit a fair amount of semantic compositionality, in a certain respect, terms in a terminological system denote semantically distinct and atomic entities.[9] Such a semantic observation, however, is a difficult property to formalize, in particular for an association measure whose task is to distinguish terms from non-terms. Hence, such an endeavor may be rather again pursued on the syntactic level, even so much the more as we have already discussed the linguistic properties of terms which may be capitalized on for formulating such an association measure.

A good starting point to isolate such a property is given by theoretical terminologists who have loosened the strict division to linguistics. In particular Cabré Castellví (2003) (see subsection 2.2.2) postulates several linguistic properties of terminological units and also states that such terminological units are more constrained with re-

[9]For example, in such terminological resources as the biomedical UMLS (see subsection 4.5.3.3) each term has its own unique identifier.

spect to their syntactic structure. Unfortunately, no further explanations, let alone linguistic examples, are given. This line of reasoning, however, is further refined from the sublanguage perspective (as described in detail in subsection 2.2.7). In particular Harris (1988), in collecting evidence for his assumption that the correlation between differences in structure and differences in information is stronger in sublanguages, reported that there are less varied patterns of substring combinability in sublanguage.[10] If Harris' observations on restricted combinability is meant to hold for sublanguage in general, it must of course also be assumed as a linguistic property of terms (as sublanguage constructs) in particular.

Justeson & Katz (1995), (one of the few) NLP researchers working on term extraction (see subsection 2.2.7) who are also concerned with the linguistic properties of terms, find that the property of repetition (i.e. frequency of co-occurrence) of terms is a quite pervasive phenomenon in text. They attribute this to yet another linguistic property of terms, *viz.* lack of variation among the component parts of terms, especially (adjectival) modifiers.[11] In their corpus and dictionary study, they look at two variation operations, namely deletion and substitution of modifiers, and basically conclude that either operation leads either to a reference to a different term or to a non-term (i.e. a common non-specific noun phrase). For this reason, terms in general refrain from such operations. Hence, a closer look at these two variation (or modification) operations may be helpful to isolate a formalizable and quantifiable linguistic property to derive a linguistically motivated statistical association measure for the task of distinguishing terms from non-terms. If we consider the deletion operation on modifiers, the first issue to notice is that this is an operation whose result may yield another term which may exhibit some sort of taxonomic (e.g. *is-a*) relationship to the original one. For example, taking the already mentioned term "*hydraulic oil filter*" from the mechanical engineering domain, an omission of "*hydraulic*" yields the term "*oil filter*" which may be seen taxonomically as a more general class term. Devising and applying a procedure for finding such *semantically* interesting taxonomic (or other) relationships, however, is something that should rather be applied to an already

[10] It should be recalled that Harris worked on string grammars consisting of symbols to better be able to derive mathematical properties of (sub)language use.

[11] As already mentioned in 2.2.7, Justeson & Katz (1995) exclude determiners (articles and quantifiers) from the class of NP modifiers because, first, they are applicable to almost any NP and, second, because they tend to indicate discourse pragmatics rather than lexical semantics.

existing set of terms. Another aspect which should be considered is that such an deletion operation may also yield a non-specific common noun phrase (i.e. a non-term). For example, the noun phrase "*side effect*" is a term entry in the 2004 edition of the UMLS biomedical terminology resource (UMLS, 2004). Deleting the nominal modifier "*side*" yields the highly general, semantically ambiguous and ubiquitous (non-term) noun "*effect*".[12] Hence, for our task of devising an association measure which is concerned with distinguishing terms from non-terms among a set of linguistic expressions (i.e. noun phrases), the (modifier) deletion operation may entail too many semantic ramifications, which are difficult to control as potential parameters and thus also difficult to quantify statistically.

Therefore, it may be worthwhile to see whether the other modification operation adduced by Justeson & Katz (1995), substitution, would not offer a more elegant solution to our task of deriving a linguistically motivated statistical association measure for term extraction. An important aspect about the substitution operation is that it may again be well motivated within Firth's model of language description (see subsection 4.2.1 above) in which empirically determined possible substitutions of words define the *paradigmatic* structure of the lexical-collocational layer. Thus, because terms in general are not prone to such modifications, we may name such a property *limited paradigmatic modifiability* or in short: LPM. In order to arrive at a more precise formulation (which will be introduced in section 4.4 below), we have to first identify where to locate LPM on the syntactic level for terms. In contrast to collocations which are syntactically much more diverse, we have already previously elaborated that the most natural (because pervasive) syntactic structure in which terms surface is the noun phrase (NP). This also has the pleasant side effect that, because we focus on NPs from the outset (which may be the output of linguistic preprocessing by means of a phrase chunker), we do not have to concern ourselves with (manually) finding and generating possible part of speech patterns (typically nouns and adjectives) in which terms may be manifested, as many other studies have done before (such as Justeson & Katz (1995), Daille (1996) or Frantzi et al. (2000)).[13]

[12] In particular, of course, if such deletion operations are performed on *bigram* modifiers, they yield highly ambiguous and ubiquitous nouns. Bigrams, however, constitute the structural type of terms with the highest proportion in any domain (see subsection 2.2.7).

[13] This is also relevant because there are subject fields, such as the biomedical domain, in which not only the typical parts of speech noun and adjective may components of terms but also various other ones such as numbers and symbols, which may lead to an inflation of possible POS patterns

Another aspect with respect to empirically determined possible substitutions in the paradigmatic context is that this should not be restricted to the set of modifiers of the head noun of an NP (as e.g. Justeson & Katz (1995) do in their manual corpus and dictionary analysis) – for a number of reasons. First, determining the head of an NP is by no means an easy task but requires more elaborate linguistic analysis, such as syntactic parsing, which is inherently more error-prone and thus less suitable to be ported to different domains as e.g. more shallow linguistic processing such as phrase chunking.[14] Second, also heads of NPs, be they terminological NPs or not, may undergo substitution. For example, in our example term from the engineering domain, "*hydraulic oil filter*", it is possible to substitute the nominal head "*filter*" with the noun "*pump*", yielding the valid terminological expression "*hydraulic oil pump*".

4.3 Limited Syntagmatic Modifiability for Collocation Extraction

In this section, we will define the linguistically enhanced statistical association measure for collocation extraction (or to put it differently: for distinguishing collocations from non-collocations). Given the analysis in subsection 4.2.2, we have found that the linguistic property of *Limited Syntagmatic Modifiability* (LSM) suits best to be both formalized and statistically implemented. In line with this analysis, we will proceed by locating LSM on the phrasal level, tied to specific part of speech (POS) patterns which may serve as the syntactic surface manifestation of potential collocations. In this thesis, our target syntactic structure are preposition-noun-verb (PNV) triples in the German language. In particular, it is the noun phrase (NP), of which the noun N of the PNV triple is the head, which serves as the phrasal unit in which to locate potential syntagmatic attachments in the form of additional lexical material. However, we will define the association measure based on LSM in such a way that it will be generalizable across a variety of syntactic target structures for collocations (subsection 4.3.1). In addition, we will also illustrate the effect that an association measure based on LSM has both on collocations and non-collocations, compared to frequency

to be specified.

[14]For example, consider the biomedical term "*diabetes mellitus type 1*" in which simply stating a head rule which takes the rightmost noun (i.e. "*type*") would yield a wrong analysis.

of co-occurrence counting (subsection 4.3.2).

4.3.1 Defining Limited Syntagmatic Modifiability

We define LSM as the linguistically motivated statistical association measure for a generic collocational syntactic target structure POS. Of this, we take a particular phrasal head $pHead$ and the associated additional lexical material, termed the syntagmatic attachment. Let n be the number of distinct syntagmatic attachments of a particular POS tuple with a selected phrasal head ($POS_{tuple,pHead}$). The probability $\mathcal{P}$ of a particular syntagmatic attachment $attach_k$, $k = [1, n]$, is described by its frequency $freq$ scaled by the sum of all syntagmatic attachment frequencies.

$$\mathcal{P}(POS_{tuple,pHead,attach_k}) := \frac{freq(POS_{tuple,pHead,attach_k})}{\sum_{i=1}^{n} freq(POS_{tuple,pHead,attach_i})} \tag{4.1}$$

with $\sum_{i=1}^{n} freq(POS_{tuple,pHead,attach_i}) = freq(POS_{tuple,pHead})$. It should be noted that the zero attachment of the POS tuple, i.e., the one for which no syntagmatic attachments occur is also included in this set.

With this, we describe *Limited Syntagmatic Modifiability* $\mathcal{LSM}$ of a POS tuple by its most probable syntagmatic attachment:

$$\mathcal{LSM}(POS_{tuple,pHead}) := arg\ max\ \mathcal{P}(POS_{tuple,pHead,attach_k}),\ k = [1, n] \tag{4.2}$$

At this point, it should be mentioned that, if necessary, more than one phrasal head $pHead$ may be selected and used to determine its LSM. This may be the case when dealing with a syntactic target structure POS which contains more than one phrasal unit. For example, given a target structure such as noun-preposition-noun-verb (NPNV, i.e. a verb associated with a noun phrase and a prepositional phrase), two phrasal heads (i.e. the two nouns) may be selected. We will incorporate this aspect into our final measure $\mathcal{COLL}$-$\mathcal{LSM}$ below.

To adhere to the requirements postulated above with respect to a linguistically motivated statistical association measure, some factor regarding frequency of co-occurrence has to be taken into account (see subsection 4.1.3) Thus, besides $\mathcal{LSM}$,

we take the relative co-occurrence frequency for a specific POS tuple $\mathcal{P}(POS_{tuple})$, with m being the number of POS tuple candidate types:

$$\mathcal{P}(POS_{tuple}) := \frac{freq(POS_{tuple})}{\sum_{j=1}^{m} freq(POS_{tuple_j})} \tag{4.3}$$

The final linguistically motivated statistical association measure for a generic collocational syntactic target structure is defined in such a way that it takes into account that more than one phrasal head may be selected to determine LSM, by computing the product. Then, with co-occurrence frequency, we have everything to define $\mathcal{COLL}$-$\mathcal{LSM}$:

$$\mathcal{COLL}\text{-}\mathcal{LSM}(POS_{tuple}) := \prod_{i=1}^{|pHead|} \mathcal{LSM}(POS_{tuple,pHead_i}) \times \mathcal{P}(POS_{tuple}) \tag{4.4}$$

In this thesis, our collocational syntactic target structure are preposition-verb-noun (PNV) triples in the German language. As the syntactically most natural phrasal head candidate, we take the (rightmost)[15] noun N of the noun phrase (NP) to determine its LSM. Thus, with n being the number of distinct syntagmatic attachments of a PNV triple with the phrasal head N ($PNV_{triple,N}$), the probability of a particular syntagmatic attachment $attach_k$, $k = [1, n]$ is as follows:

$$\mathcal{P}(PNV_{triple,N,attach_k}) = \frac{freq(PNV_{triple,N,attach_k})}{\sum_{i=1}^{n} freq(PNV_{triple,N,attach_i})} \tag{4.5}$$

Analogous to equation 4.2, the $\mathcal{LSM}$ of a PNV triple is defined by its most probable syntagmatic attachment:

$$\mathcal{LSM}(PNV_{triple,N}) = arg\ max\ \mathcal{P}(PNV_{triple,N,attach_k}),\ k = [1, n] \tag{4.6}$$

[15] Here, in line with subsection 4.2.2, we assume the output of a phrase chunker to determine noun phrases.

Because we only consider one phrasal head for its LSM, the final measure – with frequency as the second factor – is defined as:

$$\mathcal{COLL}\text{-}\mathcal{LSM}_{PNV}(PNV_{triple}) = \mathcal{LSM}(PNV_{triple,N}) \times \mathcal{P}(PNV_{triple}) \tag{4.7}$$

For expository purposes, we will label the final linguistically motivated statistical association measure for collocations as LSM for the remainder of this thesis.

4.3.2 Illustrating Limited Syntagmatic Modifiability

In this subsection, we will illustrate the effects that the linguistically motivated statistical association measure LSM has both on collocations and non-collocations. Furthermore, also the effect on a particular type of collocation, support verb constructions (see subsubsection 2.1.4.2), is examined. We will exemplify this with the help of our 114-million-word German language newspaper corpus (which we will describe in detail in subsection 4.5.2.1). For this purpose, we will take a closer look at two PNV triples which both occur with a frequency of 84 in the corpus, the collocation "*an Land ziehen*" (to reel in) and the non-collocation "*in Aktien investieren*" (to invest in stocks). To be more exact, according to the three major subtypes of collocations distinguished in subsubsection 2.1.4.2, the expression "*an Land ziehen*" may be classified as an *idiomatic phrase*.[16] Table 4.1 below lists the syntagmatic attachments for both PNV triples together with their occurrence frequencies.

As can be seen, the idiomatic expression only contains one syntagmatic attachment, *viz.* the zero attachment (which on the surface is identical to the PNV triple), which is also its most frequently occurring syntagmatic attachement (84 times). The non-collocational PNV triple, on the other hand, possesses numerous syntagmatic attachments (35 to be exact – not all of them are listed in Table 4.1 due to space limitations), with its most frequent syntagmatic attachment occurring 38 times. If we computed the LSM for both PNV triples according to equation 4.6, it would be obvious that the collocation would receive a much higher score than the non-collocation.

[16] See subsection 4.5.2.3 on how the gold-standard classification of the PNV triples into collocations and non-collocations is achieved.

PNV Triple	{Syntagmatic Attachment} + Phrasal Head	Frequency
'an Land ziehen' *'to reel in'*	{ } + Land	84
'in Aktien investieren' *'to invest in stocks'*	{ } + Aktien	38
	{ deutsche } Aktien	6
	{ europäische } + Aktien	5
	{ amerikanische } + Aktien	4
	{ polnische } + Aktien	1
	{ malaysische } + Aktien	1
	{ viele } + Aktien	1
	{ vielversprechende } + Aktien	1
	{ türkische } + Aktien	1
	{ russische } + Aktien	1
	{ israelische } + Aktien	1
	{ wertorientierte } + Aktien	1
	{ zukunftsträchtige } + Aktien	1
	{ unterbewertete } + Aktien	1
	{ überbewertete } + Aktien	1
	{ entsprechende } + Aktien	1
	{ erfolgreiche } + Aktien	1
	{ entsprechend erfolgreiche } + Aktien	1
	{ die entspechenden } + Aktien	1
	{ mehrere } + Aktien	1
	{ einzelne } + Aktien	1
	{ die } + Aktien	1
	{ diese } + Aktien	1
	...	...

Table 4.1: Collocational and non-collocational PNV Triples with Associated Syntagmatic Attachments

The previous example about the rather strict limited (in fact non-) syntagmatic modifiability for the idiomatic phrase may raise the question to what extent collocations may actually be syntagmatically modified at all. At least linguistic intuition would tell that some degree of syntagmatic modifiability should be possible. Hence,

we examine another collocation, "*unter Druck geraten*" (to get under pressure), which occurs 443 times in our corpus. The difference with the previous collocational expression is that this one may typically be classified as a support verb construction, according to the collocational subtypes laid out in subsubsetion 2.1.4.2. Table 4.2 below lists the syntagmatic attachments for this PNV triple together with their occurrence frequencies.

PNV Triple	{Syntagmatic Attachment} + Phrasal Head	Frequency
'unter Druck geraten'	{ } + Druck	395
'to get under pressure'	{ starken } + Druck	5
	{ erheblichen } + Druck	5
	{ massiven } + Druck	4
	{ starken politischen } + Druck	2
	{ zunehmenden } + Druck	2
	{ verstärkten } + Druck	2
	{ schweren } + Druck	2
	{ erheblichen } + Druck	2
	{ schweren politischen } + Druck	1
	...	...

Table 4.2: Support Verb Construction PNV Triple with Associated Syntagmatic Attachments

As can be seen, the support verb construction PNV triple does indeed possess some degree of syntagmatic modifiability. This is interesting insomuch as collocation extraction, of course, may not only be viewed as a goal by itself, but may also be utilized to create collocation lexicons for both language processing and generation (Smadja & McKeown, 1990). From this perspective, the LSM measure may actually yield quite a valuable by-product for the development of lexicons or collocational knowledge bases, *viz.* a list of possible lexical modifications associated with a particular collocational entry candidate. From a lexical-semantic viewpoint, then, a collocation may be described by the lexical semantic word classes used for modification.[17] As can be seen in Table 4.2, for the PNV triple "*unter Druck geraten*" (to get under pressure), the

[17]Of course, lexical material is always at least partially dependent on the text genre and register in question.

phrasal head noun "*Druck*" (pressure) appears to be modified by a certain semantic class of adjectives, such as "*stark*" (strong), "*massiv*" (massive), "*schwer*" (heavy), "*erheblich*" (considerable).

From the previous exemplary illustrations on the LSM property of German PNV triple collocations and non-collocations, one may wonder whether LSM, besides being a linguistically motivated statistical association measure to distinguish collocations from non-collocations, may also serve as a predictor on the subtype of a collocational expression (i.e. whether the expression is an idiomatic phrase or not). We will address and evaluate both questions in subsection 5.1.3 below.

4.4 Limited Paradigmatic Modifiability for Term Extraction

In this section, we will define the linguistically motivated statistical association measure for term extraction (i.e., for distinguishing terms from non-terms) and, given the analysis in subsection 4.2.3, we have found that the linguistic property of *Limited Paradigmatic Modifiability* (LPM) suits best to be both formalized and statistically implemented. In line with the linguistic requirements derived in subsection 4.2.3, we will proceed by defining LPM on the syntactic level of the noun phrase (NP) because it has been shown to be the most natural and pervasive syntactic structure for the surface manifestation of terms (subsection 4.4.1). In doing so, we will ensure that LPM will be generalizable and applicable to NP n-grams of all sizes although in practice only n-grams up to size four are relevant (see Justeson & Katz (1995)). In addition, we will also illustrate the effect that an association measure based on LPM has both on terms and non-terms, compared to frequency of co-occurrence counting (subsection 4.4.2).

4.4.1 Defining Limited Paradigmatic Modifiability

The linguistic property that distinguishes terms from non-terms, the *limited paradigmatic modifiability* (LPM) of multi-word terminological units, serves as the basis for our measure of termhood. In line with the our linguistic Firthian frame of reference (see subsection 4.2.1) LPM considers the modifiability of the paradigmatic context for a particular word token within a potential terminological expression. In this way,

an n-gram multi-word expression $word_1...word_n$ within a noun phrase may be viewed as containing n word token slots in which each of the slots is filled by a particular word token. For example, in our trigram term example from the mechanics domain, "*hydraulic oil filter*", slot 1 is filled by the word "*hydraulic*", slot 2 by "*oil*" and slot 3 by "*filter*". Since the *paradigmatic modifiability* of such an n-gram is supposed to be *limited* for terms, the most natural way to define LPM is by the probability with which one or more such slots *cannot* be filled by other word tokens, i.e., the tendency not to let other words appear in particular slots. Since with higher-order n-grams one has to take into account the various combinatory possibilities to fill such slots, a procedure needs to adopted which yields such combinatory constellations. The linguistically most natural way is to employ the standard combinatory formula (Rosen, 1999) – without repetitions to avoid duplicate patterns. Thus, for an n-gram (of size n) to select k slots (i.e., in an unordered selection) may be defined in the following way.

$$C(n,k) = \frac{n!}{k!(n-k)!} \tag{4.8}$$

As depicted in table 4.3 below, for $n = 3$ (a word trigram) and $k = 1$ and $k = 2$ slots, there are three possible selections for each k, and for $k = 3$ slots, there is one possible selection. In this way, k can be thought of as a placeholder for any possible word token and its frequency which fills this position.

k slots	possible selections for trigram
$k = 1$	k_1 word_2 word_3 word_1 k_2 word_3 word_1 word_2 k_3
$k = 2$	k_1 k_2 word_3 k_1 word_2 k_3 word_1 k_2 k_3
$k = 3$	k_1 k_2 k_3

Table 4.3: Possible selections for $k = 1$, $k = 2$ and $k = 3$ for a trigram noun phrase

In order to arrive at the computation of the LPM for a particular n-gram noun phrase, one intermediate computation needs to be done, i.e. for a particular k ($1 \leq$

$k \leq n$; n = length of n-gram), the frequency of each possible selection, *sel*, needs to be determined. Because this frequency is at least as high as the n-gram frequency, the paradigmatic modifiability for a particular selection *sel* may be defined by the n-gram's frequency scaled against the frequency of *sel* which, in turn, results in a well-defined probability value (i.e. between 0 and 1). Thus, with $|sel|$ being the number of distinct possible selections for a particular k, the limited paradigmatic modifiability of possible *k-selections*, $lpm_{k\text{-}sel}$, of an n-gram can be derived as the product of all the k-selection modifiabilities:

$$lpm_{k\text{-}sel}(n\text{-}gram) := \prod_{i=1}^{|sel|} \frac{f(n\text{-}gram)}{f(sel_i, n\text{-}gram)} \tag{4.9}$$

As a last step, in order to derive the limited paradigmatic modifiability, LPM, of an n-gram, the product over all its k-selection modifiabilities needs to be computed.

$$LPM(n\text{-}gram) := \prod_{k=1}^{n} lpm_{k\text{-}sel}(n\text{-}gram) \tag{4.10}$$

It is important to note with respect to LPM that setting the upper limit of k to n (which is the size of an n-gram and thus $n = 3$ for trigrams) actually has the pleasant side effect of including frequency of co-occurrence in our termhood measure. In this case, the only possible selection $k_1k_2k_3$ (see also table 4.3 above) as the denominator of equation (4.9) is equivalent to summing up the frequencies of all trigram term candidates. Hence, it is ensured that we adhere to the requirements postulated above with respect to a linguistically motivated statistical association measure, *viz.* that some factor regarding frequency of co-occurrence has to be taken into account (see subsection 4.1.3).

An additional point to be noticed about LPM is that it is a combinatorics-based algorithm. Although it is well-known that such algorithms in general do possess a high degree of time-consuming computational complexity (Rosen, 1999), this is of no practical relevance for our task of term extraction. As already pointed out at various occasions in this thesis, Justeson & Katz (1995)'s analysis revealed that only a small minority (i.e. less than 6%) of terms are actually contained within n-grams equal or larger than size four (see subsection 2.2.7)[18] and thus complexity considerations do not turn out to be a point of practical concern.

[18]Furthermore, of course, the larger an n-gram becomes, the sparser its occurrence in natural language text is.

4.4.2 Illustrating Limited Paradigmatic Modifiability

In this section, we will illustrate the effects that our linguistically motivated statistical term association measure LPM has on both terms and non-terms. For this purpose, we will look at the corpus data that we use for our term extraction experiments, i.e. a 100-million word English corpus from the biomedical subdomain of Hematopoietic Stem Cell Transplantation and Immunology, which was downloaded from the world's largest bibliographic database for biomedicine, MEDLINE. The assembly and linguistic processing of this corpus will be described in detail in subsection 4.5.3.1.

In tables 4.4 and 4.5, we exemplarily computed the LPM scores both for the trigram term "*open reading frame*"[19] and for the trigram non-term "*t cell response*". For our illustrative purposes to show the effects of LPM, we computed separate scores for the lower ks, $k = 1, 2$ (i.e. lpm_1 and lpm_2), as well as for the complete LPM, i.e. for $k = 1, 2, 3$, which thus incorporates the frequency of co-occurrence of a term candidate. As can be seen in the two tables, a lower frequency for a particular k-selection induces a more limited paradigmatic modifiability for that k-selection which in turn is expressed as a higher probability for $lpm_{k\text{-}sel}$ (i.e. the product of all particular k-selections under consideration, as shown in equation 4.9 above), and vice versa of course. As table 4.4 shows, the term "*open reading frame*" has a much higher LPM value for $k = 1, 2$ (0.11) than the non-term "*t cell response*" (0.000008) shown in table 4.5. Evidently, this is due to the fact that the non-term allows a much higher number of different paradigmatic substitutions at its k-slot positions than the term does. Also, incorporating and computing the complete LPM value for $k = 1, 2, 3$, which includes frequency of co-occurrence, does not change the fact that the term gets a much higer LPM score (0.00002) than the non-term (0.000000002).

These LPM scores have the desirable effect that in the respective output list rank (of 28,000 ranked term candidates here), the term "*open reading frame*" is placed on a high rank (rank 56) whereas the non-term "*t cell response*" gets assigned to a much lower rank, i.e. rank 1000. If, on the other hand, only the frequencies of both expressions would be considered, the two tables reveal that the non-term occurs over 17 times more often in the biomedical text corpus than the term. Accordingly, a

[19]This term denotes the portion of an organism's genome which contains a sequence of bases that could potentially encode a protein. Subsubection 4.5.3.3 describes in detail how the actual terms are distinguished from non-terms in our experiments.

n-gram	freq	*LPM* (k=1,2)	*LPM* (k=1,2,3)
"*open reading frame*"	153	0.11	0.00002

k slots	possible selections *sel*	freq	$lpm_{k\text{-}sel}$
$k = 1$	k_1 reading frame	213	0.72
	open k_2 frame	153	1.0
	open reading k_3	155	0.99
		lpm_1 =0.71	
$k = 2$	k_1 k_2 frame	257	0.6
	k_1 reading k_3	221	0.69
	open k_2 k_3	429	0.36
		lpm_2 =0.15	
$k = 3$	k_1 k_2 k_3	960,538	0.0002
		lpm_3 =0.0002	

Table 4.4: *LPM* and k-selection modifiabilities for $k = 1$ and $k = 2$ for the trigram term "*open reading frame*"

n-gram	freq	*LPM* (k=1,2)	*LPM* (k=1,2,3)
"*t cell response*"	2,335	0.000008	0.00000002

k slots	possible selections *sel*	freq	$lpm_{k\text{-}sel}$
$k = 1$	k_1 cell response	2,993	0.8
	t k_2 response	2,490	0.94
	t cell k_3	21,960	0.11
		lpm_1 =0.08	
$k = 2$	k_1 k_2 response	38,215	0.06
	k_1 cell k_3	110,718	0.02
	t k_2 k_3	26,703	0.09
		lpm_2 =0.0001	
$k = 3$	k_1 k_2 k_3	960,538	0.002
		lpm_3 =0.002	

Table 4.5: *LPM* and k-selection modifiabilities for $k = 1$ and $k = 2$ for the trigram non-term "*t cell response*"

ranking based on mere frequency of co-occurrence would rank the non-term "*t cell response*" very high on rank 19 whereas it would place the term "*open reading frame*" quite low on rank 787 on the output list. In fact, even lower-frequency trigrams gain a prominent ranking if they exhibit a more limited paradigmatic modifiability. For example, the trigram term "*porphyria cutanea tarda*" is ranked on position 28 by LPM, although its co-occurrence frequency is only 48 (which results in rank 3291 on the frequency-based output list). Despite its lower frequency, this term is judged as being relevant for the domain under consideration.[20] Of course, as is evidenced in tables 4.4 and 4.5, it should be noted that the termhood scores (and the corresponding list ranks) computed by LPM also include the $k = 3$ selection and, hence, take into account a reasonable amount of frequency load. As can be seen from the previous ranking examples, however, this factor does not override the limited paradigmatic modifiability factors of the lower-order ks (i.e. $k = 1, 2$).

On the other hand, LPM will also demote true terms in their ranking, if their paradigmatic modifiability is less limited. This is particularly the case if one or more of the word tokens of a particular term often occur in the same k-slot of other equal-length n-grams. For example, the trigram term "*bone marrow cell*" occurs 1757 times in our corpus and is thus ranked quite high (position 18) by frequency. LPM, however, ranks this term on position 583 because the word token "*cell*" also occurs in many other trigrams at this position and thus leads to a less limited paradigmatic modifiability. Still, as laid out extensively in subsection 4.2.3, the underlying assumption of the LPM approach is that such a case is more an exception than the rule and that terms are linguistically more frozen than non-terms and are thus not as prone to such modifications as non-terms are, which is exactly the intuition behind our association measure of limited paradigmatic modifiability.

4.5 Evaluation Setting

In the previous two sections, we have motivated and defined two linguistically inspired statistical association measures both for collocation extraction (limited syntagmatic modifiability – LSM) and for term extraction (limited paradigmatic modifiability – LPM). While any methodological innovation needs to be put on a theoretically and

[20] It denotes a group of related disorders, all of which arise from a deficient activity of the heme synthetic enzyme uroporphyrinogen decarboxylase (URO-D) in the liver.

definitionally sound basis, this alone would not be worth much if we were not able to evaluate its performance against other comparable standard methods and actually find that the method under consideration outperforms its standard competitors to a substantial degree – thus making the whole enterprise worth while in the first place.

Therefore, in this section we will construct a comprehensive evaluation setting in which we are able to compare both our linguistic association measures with the standard statistical and information-theoretic measures described in section 3.3. For this purpose, we will first establish the general requirements for such an evaluation setting in subsection 4.5.1 where we focus on several necessary prerequisites, *viz.* the assembly of an appropriate text corpus, the linguistic processing to obtain valid extraction candidates, the procedures to classify these candidates as to whether they are (non-)collocations or (non-)terms and – at the heart – appropriate evaluation metrics which allow for a comprehensive and multi-faceted view on the performance of the different association measures under scrutiny. After having laid out these general requirements, we explain in detail how we have implemented the respective evaluation setting for the task of collocation extraction (subsection 4.5.2) and for the task of term extraction (subsection 4.5.3).

4.5.1 General Requirements for Evaluation

In this subsection, we will outline the general steps which are necessary in order to arrive at an appropriate evaluation setting, independent of the fact whether we deal with collocation extraction or with term extraction. First, we will highlight the considerations which need to be taken into account in assembling a suitable text corpus for the respective extraction task (subsubsection 4.5.1.1). Then, we focus on how to obtain extraction candidates from syntactic target structures by performing some form of linguistic preprocessing on the text corpus (subsubsection 4.5.1.2). Subsequently, we need to worry about how to classify the target candidates as (non-)collocations or (non-)terms and thus arrive at an appropriate gold standard which makes performance evaluation possible in the first place (subsubsection 4.5.1.3). Subsubsections 4.5.1.4 and 4.5.1.6 will focus on the heart of any performance evaluation enterprise – the performance evaluation metrics – which we will motivate both from a quantitative and a qualitative perspective. Good evaluation practice also requires the construction of appropriate baselines as well as testing whether the results obtained are statistically significant (subsubsection 4.5.1.5).

4.5.1.1 Assembling the Text Corpus

The first step that needs to be done in creating an appropriate testing and evaluation setting both for collocation and term extraction measures is to assemble a suitable text corpus on which the extraction procedures may be run. Besides addressing the question in which language the texts of the corpus should be written, a key consideration has to be centered around the domain(s) and/or genres of the text collection. Because collocations may most appropriately be considered as linguistic expressions which are most pervasive in general language (see section 2.1 and 2.3), a text corpus needs to be assembled which may be considered representative for general language. This, of course, is by no means a straightforward task because "general language" as such is rather a fuzzy notion and may be almost regarded as broad-faceted as the notion of language itself consisting of numerous genres and registers. Hence, a necessary step is to find a kind of *text corpus provider* of text sources which may at least partially fulfill the claim of covering general language (see subsubsection 4.5.2.1 below).

The task of assembling an appropriate text corpus for evaluating different term extraction procedures may be considered as a comparatively easier enterprise. Because terms are considered as linguistic expressions which are typically confined to a particular subject domain and sublanguage (cf. section 2.2), the key consideration here is to choose a subject domain with a suitable corpus repository. One requirement that a corpus repository should fulfill, both for general-language collocations and domain-specific terms, is that it is large enough so that reliable statistics may be computed from it (see subsection 4.5.2.1). At the same time, however, it should also be possible to vary (i.e. increase or decrease) the corpus size substantially and still have the lexical association measures under scrutiny produce the same result patterns. This is to ensure that their observed empirical properties and respective differences are not mainly due to corpus size. For this reason, we will perform all experiments and evaluations, both for collocation and for term extraction, on two substantially different corpus sizes (see subsections 4.5.2.2 and 4.5.3.2 below).

4.5.1.2 Extracting and Counting the Candidates

Once an appropriate text corpus has been assembled, the next step is to acquire a set of collocation and term candidates on which the respective extraction procedures may

be run. We have laid out in detail (cf. section 4.2) that both collocation and term candidates may be best obtained by performing some form of linguistic preprocessing on the text corpus. Through this, an appropriate syntactic target structure may be selected which is in line with the linguistic observation that both collocations and terms are typically manifested in certain syntactic surface structures. In addition, linguistic filtering also considerably eases the violations which natural language as sample data typically causes on the statistical distribution and independence assumptions on word combinations (see subsections 3.3.6 and 4.1.1). Although this is not so much a concern for both our linguistically motivated statistical association measures, it is essential for the standard parametric association measures which crucially rely on these assumptions. Because one of the objectives of this thesis is to extensively compare standard and linguistic association measures, applying linguistic filters to acquire collocation and term candidates makes such a comparison possible in the first place because it grants the standard measures "equal opportunity" with respect to their non-parametric competitors LSM and LPM.

Another question which has to be raised in this respect deals with the sort of linguistic preprocessing which may be deemed appropriate to obtain a set of candidates, in particular considering the fact that collocations may surface in a variety of syntactic constructions (cf. subsection 4.2.2). Although the most elaborate form of syntactic analysis, full syntactic parsing, has been attempted for the task of acquiring a set of collocation candidates (Lin, 1998a; 1999), it has rather proven to be a persistent source of errors which had to be corrected in a time-consuming manner. Thus, the cost-benefit ratio of parser deployment may not be regarded as justified (see the discussion in subsection 3.1.3), in particular considering the fact that collocation dictionaries should rather be the input than the output of full syntactic parsing (Collins, 1997; Klein & Manning, 2003). For these reasons, most studies employ some form of shallow syntactic analysis to isolate their collocation candidate target structures, either by POS tagging (e.g. (Dunning, 1993)) or phrase chunking (e.g. (Evert & Krenn, 2001; Krenn & Evert, 2001)). Similar arguments may be put forth in the case of acquiring a set of term candidates from a text corpus. Here, however, the use of a shallow preprocessor is additionally corroborated by the linguistic fact that domain-specific terms are typically manifested within noun phrases (cf. the discussion in subsections 2.2.7 and 4.2.3). Hence, all that is at most needed is the deployment of a (noun phrase) phrase chunker, or even only the specification of appropriate NP-specific POS

patterns to filter the output of a POS tagger.

Because the identification of collocations and terms from a set of candidates relies on association measures employing some form of statistical computations, the candidates (and possibly their component parts – depending on the association measure) need to be counted after their linguistic isolation. While this may rather be considered as a straightforward procedure, a question immediately to be raised is whether all candidates identified and counted should be included or only those above a certain frequency threshold. Virtually all studies both on collocation and on term extraction (e.g. (Smadja, 1993; Dunning, 1993; Daille, 1996; Lin, 1999; Frantzi et al., 2000; Evert & Krenn, 2001)) specify some frequency cut-off threshold. Such thresholds are all based on the widely held assumption that low-frequency data may be considered as noise whose inclusion in the sample data may distort sound statistical estimations. The frequency cut-offs typically employed are $freq \geq 5$ or more and it is Evert (2005) who also provides mathematical evidence that, indeed, probability estimates and p-values for frequency data below this threshold are distorted in unpredictable ways due to quantization effects and erratic population shapes (see subsection 3.1.4 above). Hence, a necessary condition is that any approach to the definition and implementation of lexical association measures, either linguistically motivated or standard, needs to specify a frequency cut-off threshold. Evert (2005) also points out that the cut-off frequency of five is just the minimum threshold below which no candidates should be considered. Higher frequency thresholds are of course possible, in particular if the overall corpus size and thus the size of the candidate set is very large. For example, in our case of setting up an appropriate evaluation framework for collocation and term extraction below, the specification of such thresholds also needs to be guided by practical concerns in order to avoid unrealistically large candidate sets.

4.5.1.3 Classification of the Targets

A crucial question which any approach to collocation and term extraction needs to address is the evaluation of the performance quality of the extraction methods employed. Because virtually all statistical and information-theoretic lexical association measures output an ordered ranked list of candidates (see subsection 4.1.4), many studies, either on collocation extraction (Manning & Schütze, 1999; Dunning, 1993) or on term extraction (Frantzi et al., 2000; Collier et al., 2002),

evaluate the goodness of their algorithms by having the ranked output examined by a domain expert, in the case of term extraction, or by a lexicographer, in the case of collocation extraction (Smadja, 1993), and thus identify the actual collocations or actual terms (i.e., the targets) among the ranked candidates.[21] There are several problems with such an approach. Because such an evaluation procedure is rather labor-intensive and time-costly, the actual number of ranked candidates examined, let us call it n, is typically very small, ranging from 50 up to several hundreds (e.g. in (Manning & Schütze, 1999; Frantzi et al., 2000; Dunning, 1993)) although the size of the output list is much larger. This, in turn, leads to very superficial judgments about the quality of the association measures examined, in particular because it only allows the calculation of a cursory accuracy score. Another problem besides the paucity of the output candidates examined is that, because the evaluation is performed on the ranked candidates with a substantial proportion of targets already placed in the upper portion of the list,[22] the raters are already biased by the high density of actual collocations or terms that they encounter. For these reasons, Evert & Krenn (2001) and Evert (2005) recommend that human judges identify the actual collocations or terms in the *complete* unordered (i.e. unranked) candidate set obtained by linguistic filtering (see the previous subsubsection 4.5.1.2) *before* any lexical association measure is applied. In doing so, it is possible to calculate and plot the whole array of standard quantitative performance evaluation metrics (see the next subsubsection 4.5.1.4), thus allowing for a much more principled and reliable evaluation on the complete candidate set and not just a cursory portion of it.

At this point one may wonder whether using a collocation or term dictionary as a gold standard against which the actual target collocations or terms in a candidate set may be identified would not be a more feasible (and less costly) procedure for evaluating association measures than relying on human judgments. With respect the availability of such resources for general-language collocations, we have already seen that they are notoriously incomplete and deficient (Lin, 1999; 1998b), even for well-documented languages such as English. For this reason, if the goal is to carry out a reliable and thorough evaluation of extraction procedures for the task of collocation

[21] Sometimes, it is even the authors of a study themselves who examine the output of their extraction methods, e.g. Manning & Schütze (1999).

[22] That is, if the association measures examined are any good this is most probably the case.

extraction, it is virtually impossible to avoid consulting human judges in order to have a set of collocation candidates classified according to the targets in it. This, however, also implies another time-costly but necessary task, *viz.* assuring some form of quality control that the classification judgments performed by humans contain a sufficient amount of stability and reliability. This is typically determined by measuring the agreement between different human judges (or termed differently: annotators) in some way, which is typically referred to as inter-annotator agreement. In subsubsection 4.5.2.3, we will describe extensively how we implemented this sort of quality control for our collocational candidate set of German PNV triples.

In the case of evaluating term extraction procedures, the prospect on the (time-saving) availability of comprehensive domain-specific terminological resources does not look as bleak as in the case of such resources for general-language collocations. For example, the work by Daille (1994) and Daille (1996) bases the evaluation of the association measures examined (see subsection 3.2.2 above) on the entries of a terminology database from the telecommunications domain – with the major caveat issued, however, that the resource lacks completeness. Fortunately, the picture looks different in the biomedical domain which this thesis focuses on. As a matter of fact, it hosts one of the most extensive and carefully curated terminological resources which has evolved over the years, is constantly updated and reflects community-wide consensus achieved by expert committees, *viz.* the UMLS Metathesaurus (Bodenreider, 2004).[23] We will describe extensively in subsubsection 4.5.3.3 how we will exploit this resource as a gold standard to construct an appropriate evaluation setting for term extraction.

4.5.1.4 Quantitative Performance Evaluation

Given the evaluation setting described in the previous subsubsection – in which the actual collocations or terms are identified in the *complete* unordered (i.e. unranked) candidate set obtained by linguistic filtering before any lexical association measure is applied – we may use the whole array of standard quantitative performance evaluation metrics, such as precision, recall, F-score, and fallout, which have been proposed and are standard practice in information retrieval (Baeza-Yates & Ribeiro-Neto, 1999; Rijsbergen, 1979; Manning & Schütze, 1999). For this purpose, however, we first

[23] `http://umlsinfo.nlm.nih.gov`

need to clarify the context in which these evaluation metrics must be placed in order to be able to evaluate lexical association measures (LAM). Thus, a (collocation or term) candidate set may be thought of as containing a set of *target* (i.e. actual) collocations or terms. An association measures then *selects* a set of candidates that it considers to be collocations or terms. This situation may be depicted in a 2 x 2 contingency table, such as table 4.6.

LAM	*target*	*not target*
selected	true positive	false positive
not selected	false negative	true negative

Table 4.6: Context of quantitative performance evaluation.

The instances labeled as *true positive* (TP) and *true negative* (TN) are those that an association measure correctly selects. The wrongly selected instances are labeled as *false positive* (FP) whereas the cases that are failed to be selected are considered *false negative* (FN). Hence, we are now in position to define *precision* as an evaluation metric for the proportion of selected items that an association measure correctly identified as (target) collocation or terms:

$$precision := \frac{TP}{TP + FP} \tag{4.11}$$

Conversely, *recall* may be defined proportion of target collocation or terms that an association measure selected:

$$recall := \frac{TP}{TP + FN} \tag{4.12}$$

If overall performance is to be evaluated, a typical metric to use is the weighted harmonic mean of precision and recall, the balanced *F-score*:

$$F\text{-}score := \frac{2 * (precision * recall)}{precision + recall} \tag{4.13}$$

Typically, this performance evaluation metric is only considered useful if a distinctive comparison between precision and recall is not considered insightful for the evaluation task at hand. Thus, F-score does not need to be applied if, for example, single precision and recall evaluations are more telling, which is typically the case for extraction tasks like the ones focused on here (Evert, 2005).

An evaluation metric less frequently used (Manning & Schütze, 1999) is *fallout*, which is defined as the proportion of non-targets that are erroneously selected:

$$fallout := \frac{FP}{FP + TN} \tag{4.14}$$

Fallout is mainly used as an evaluation metric of how difficult it is to construct a system (or in our case: to devise an association measure) that produces few false positives. In our case of selecting target collocations or terms from a set of candidates, such an evaluation metric may actually be informative because the number of non-targets is generally quite large (see subsubsections 4.5.2.3 and 4.5.3.3 below), thus making it almost inevitable that there will be misclassifications. Typically, however, fallout is not used by itself but in connection with recall to set up so-called *receiver operating characteristic* (ROC) curves, as pointed by Manning & Schütze (1999). These show how different proportions of false positives (i.e. the fallout) influence the true positive rate (i.e. the recall).

Because collocation or term candidate sets obtained by means of linguistic filtering may typically turn out to be quite large (see also our own experiments in subsubsections 4.5.2.2 and 4.5.3.2 below), it may be helpful to consider association measure performance after a certain proportion of the ranked output list has been scanned. In this manner, for example, a selective recall value may indicate what proportion of the target collocations or terms a lexicographer or domain expert would already have considered if a collocation or term extraction system presented such a ranked list to him or her. For this reason, Evert & Krenn (2001) suggest to incrementally calculate performance metrics by dynamically considering n highest ranked samples. i.e. incremental portions of the ranked output list, until the complete candidate output list

has been examined. Thus, in putting the evaluation of lexical association measures into such a well-founded framework, a sound comparison from different perspectives may be carried out between our linguistically motivated association measures, on the one hand, and the standard statistical and information-theoretic ones, on the other hand.

4.5.1.5 Baselines and Significance

In a comprehensive evaluation setting, it is essential to establish a *baseline* which indicates the lower bound of system performance. For lexical association measures, such a lower bound baseline with respect to the precision evaluation metric may be best defined as the percentage of (collocation or term) targets in the candidate set. Such a procedure may then be interpreted as the likelihood of finding a target by randomly picking from the candidate set. Establishing such a lower baseline as the only one, however, entails that every lexical association measure which outperforms it must be regarded as a potentially useful measure of collocativity or termhood. For this reason, we also postulate a more challenging (but still easy to implement) baseline which has shown to be quite competitive with respect to statistical and information-theoretic association measures, *viz.* frequency of co-occurrence (see subsection 3.3.7). This is contrary to other studies which regard it as an association measure among others (Evert & Krenn, 2001; Krenn & Evert, 2001; Daille, 1996). Conversely to establishing a lower bound baseline, it is also possible to define an upper limit for lexical association measure performance (Evert, 2005). Such an optimal association measure would rank all targets at the top of the output list and thus entirely fulfill the requirements postulated in subsection 4.1.4 for an ideal system.

It may be the case that in comparing association measures (or systems in general) by means of precision and recall values and graphs, the observed differences are rather small thus raising the question whether they are existent at all or merely due to chance.[24] Therefore, it may be necessary to test whether the differences observed in evaluation results are statistically significant. For testing such differences between association measures, it may be best to do so on the basis of incremental portions of the ranked output list. A more difficult question is which one out of the wide range

[24]The reasons for such chance differences may be manifold, e.g., choice of the text corpus, noise due to linguistic filtering, unreliability of human judgments or incompleteness of (terminological) resources, to name just a few.

of significance tests may be best applied. For example, Krenn & Evert (2001) use Pearson's chi-squared test but note at the same time that this test assumes independent samples which is strictly speaking not admissible for the comparison of different (association-measure derived) rankings performed on the same (collocation or term) candidate set. Hence, it is more advisable to use the McNemar test (Sachs, 1984; McNemar, 1947) which is a special *non-parametric* version of the chi-squared test and may thus also be applied to dependent samples. In concrete, employing McNemar as a significance test of differences between two lexical association measures (LAM) may be depicted by means of the 2 x 2 contingency table in Table 4.7. The values of the cells may then be computed by taking a certain portion of the ranked output list, which of course may be done incrementally.

	LAM_1 target	LAM_1 non-target
LAM_2 target	a	b
LAM_2 non-target	c	d

Table 4.7: The McNemar significance test of differences for comparing two lexical association measures (LAMs).

The actual test statistic is then given by the following equation:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \tag{4.15}$$

The factor -1 marks Yates' discontinuity correction (Yates, 1934) which, however, only needs to be applied in the case of small samples, i.e. approximately $b + c < 30$. When taking a closer look at the McNemar test, a striking observation is that the cases when the two association measures agree contribute no information to the decision whether the differences are significant or not. In fact, it is the only cases when the two association measures *disagree*, i.e. b and c, which contribute to this significance test. In all our experiments in the next chapter, we will apply the McNemar test for a (very strict) confidence interval of 99%.

4.5.1.6 Qualitative Performance Evaluation

The methods of performance evaluation outlined in the previous two subsections (i.e. precision, recall, F-score, fallout and ROC in subsubsection 4.5.1.4 and the McNemar significance test in subsubsection 4.5.1.5) may be roughly characterized as examining the differences between different association measures from a mainly *quantitative* perspective. For the sake of a truly comprehensive evaluation and comparison of our linguistically motivated association measures LSM and LPM with respect to the standard measures, it may be also informative to provide some sort of *qualitative* performance evaluation. For devising such a qualitative performance evaluation for lexical association measures, it may be helpful to recall that in subsection 3.3.7, we have outlined how frequency of co-occurrence may be used as an easily implementable measure. As previously pointed out, several studies (e.g., Daille (1996), Evert & Krenn (2001)) have already reported on quite competitive evaluation results of this association measure with respect to statistical and information-theoretic ones. It is also for this reason that we have defined frequency of co-occurrence as a second more challenging baseline (see the previous subsection 4.5.1.5).

It is therefore reasonable to postulate that a qualitative performance evaluation should target the differences between frequency of co-occurrence, on the one hand, and the lexical association measure under consideration, on the other hand. In devising an appropriate set of qualitative criteria we take up the two conditions laid out in subsection 4.1.4 which state that an ideal association measure should rank the targets from the candidate set in the upper portion of the output list and, conversely, the non-targets in the lower portion. Then, the following four qualitative achievement objectives for a lexical association measure may be formulated with respect to frequency of co-occurrence:

1. Keep the targets[25] in the upper portion.
2. Keep the non-targets in the lower portion.
3. *Demote* the non-targets from the upper portion.
4. *Promote* the targets from the lower portion.

[25] Which were ranked by frequency of co-occurrence.

In subsection 5.1.2 and in subsection 5.2.2 in the next chapter, we will examine for collocation extraction and for term extraction, respectively, how these four criteria may be taken to compare the different rankings assigned by a certain association measure and by frequency of co-occurrence. It should be noted that the first two qualitative criteria are more static whereas the last two may be more described as dynamic. For this purpose, it is best to choose the middle rank as a mark to divide a ranked output list into an upper portion and a lower portion. Then the targets and non-targets assigned to these portions by frequency may be examined and quantified, according to the four criteria, to what degree the other association measures changed these rankings (or not).

4.5.2 Evaluation Setting for Collocation Extraction

In this subsection, we will describe the evaluation setting for the task of collocation extraction with respect to German-language preposition-noun-verb (PNV) combinations. First, we will describe how we assemble a balanced general-language text corpus (subsubsection 4.5.2.1) as well as the morphological and syntactic analyzers with which we preprocess this corpus linguistically in order to obtain the syntactic target structure (subsubsection 4.5.2.2). Finally, we explain in detail how we classified the target PNV candidates into collocations and non-collocations by relying on human linguistic judgments. In particular, we also focus on how quality control was carried out with respect to the reliability of these judgments – in order to ensure a sound and valid gold standard for our performance evaluations (subsubsection 4.5.2.3).

4.5.2.1 Text Corpus and Linguistic Filtering

As laid out in subsection 4.5.1.1 above, collocations are typically to be considered as general language expressions and, consequently, an ideal text corpus should be as representative as possible for general language. For the (British) English language and as a pars pro toto of such a text corpus, the *British National Corpus* (BNC)[26] (Leech, 1992; 1993) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The written part of the BNC (90%) includes, for example, extracts from regional

[26] `http://www.natcorp.ox.ac.uk`

and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) includes a large amount of unscripted informal conversation, recorded by volunteers selected from different age, region and social classes in a demographically balanced way, together with spoken language collected in all kinds of different contexts, ranging from formal business or government meetings to radio shows and phone-ins. Unfortunately, such a representative (and freely available) corpus does not exist for the German language, which is the focus for collocation extraction in this thesis . Although the COSMAS corpus[27] (Stickel, 1994) may be considered as a large electronic text collection for the German language, it is mainly composed of newspaper text material[28] and, in addition, only accessible via a web service or a desktop client program.

For these reasons, we decided to assemble our own German text corpus by downloading 114 million words of text material from the online archive of the German newspaper Die Welt[29] ranging in years from 1996 to 2001. In order to ensure as much genre diversity and textual representativeness as possible, we included articles from all textual categories offered, i.e., politics, business and economy, finance, sports, culture and society, science and research, information technology, travel, lifestyle and cars. Then, this text collection was linguistically processed by applying a POS tagger and a phrase chunker to it. The POS tagger used was the Hidden-Markov-Model (HMM) based TNT tagger (Brants, 2000) which was trained on the German NEGRA corpus (Skut et al., 1997), which consists of 355,096 tokens (20,602 sentences) of syntactically annotated newspaper text material. The part of speech tagset used for this annotation is the STTS tagset (Thielen & Schiller, 1996) which may be considered as the standard tagset for the German language. Phrase chunking, in particular identifying noun phrases (NPs) and prepositional phrases (PPs), was then performed by implementing a set of cascaded finite state (regular expression) chunk rules based on (STTS) part of speech sequences, along the lines suggested by Abney (1991) and Abney (1996). Finally, in order to normalize morphological variation of the component parts of the the PNV triple collocational target structure (and thus eliminate poten-

[27] `http://www.ids-mannheim.de/cosmas2`

[28] See `http://www.ids-mannheim.de/cosmas2/referenz/korpora.html`

[29] `http://www.welt.de/archiv`

tial noise for counting), all main verbs and common nouns were lemmatized to their base form by a morphological analyzer for German (Lezius et al., 1998).

4.5.2.2 Target Structure and Candidate Sets

From the linguistically preprocessed text output (see the previous subsection 4.5.2.1), preposition-NP-verb patterns were automatically selected in the following way: taking a particular preposition as a fixed point, the immediately following NP[30] was selected together with either the preceding or the following main verb. From such preposition-NP-verb combinations, we extracted and counted both the various *heads*, in terms of *P*reposition-*N*oun-*V*erb (PNV) triples as our collocational syntactic target structure, and all the associated *syntagmatic attachments*, i.e., here any additional lexical material which also occurs in the noun phrase, such as articles, adjectives, adverbs, cardinals, etc. The extraction (and counting) of the associated syntagmatic attachments is of course essential to our linguistically motivated statistical association measure LSM described in section 4.3 above.

As we have already pointed out in subsubsection 4.5.1.2 above, it is necessary to specify a frequency cut-off threshold thus limiting the number of candidates to be included in the candidate set. In the case of collocation extraction, the setting of such a threshold also needs to be guided by practical concerns because the targets (i.e. the actual collocations) need to be manually identified (i.e. by human annotators) in the complete collocational candidate set in order to ensure a sound and reliable evaluation (see the discussions in subsubsections 4.5.1.3 and 4.5.1.4). Therefore, in order to obtain a candidate set whose human classification is practically feasible at all in time and effort, we set the frequency threshold to $f > 9$ and only included PNV triples above this cut-off threshold from our 114-million-word German newspaper corpus in our collocational candidate set. Table 4.8 contains the frequency distribution of both PNV triple tokens (i.e., all single-instance linguistic expressions) and types (i.e., distinct linguistic expressions), both with and without the frequency threshold applied.

As can be seen, there is a huge decrease in numbers of the PNV triple tokens and types if a frequency threshold is applied. In particular, the distinct PNV triple types

[30]Thus, the NP is of course taken to be the phrasal unit from which we isolate our phrasal head N for our PNV triples, as established in the definition of the LSM association measure in subsection 4.3.1.

frequency	PNV triples	
	candidate tokens	candidate types
all	1,663,296	1,159,133
$f > 9$	279,350	8,644

Table 4.8: Frequency distribution of PNV triple tokens and types for our 100-million-word German newspaper corpus

which in effect constitute the collocational candidate set amount to 8,644, which is a feasible size in terms of human annotation time and effort.

As we have laid out in subsection 4.5.1.1 above, in order to ensure that the observed empirical properties and respective differences of lexical association measures are not mainly due to corpus size, we will also run our experiments and evaluations on a substantially different corpus size. For this purpose, we reduce the size of our German newspaper corpus to about 10% of its original size, thus yielding 10 million word tokens. Table 4.9 shows the respective frequency distribution in terms of PNV triple tokens and types.

frequency	PNV triples	
	candidate tokens	candidate types
all	132,136	117,062
$f > 4$	12,529	1,035

Table 4.9: Frequency distribution of PNV triple tokens and types for 10 million words of German newspaper corpus

As can be seen, we have set the frequency cut-off threshold to $f > 4$ which is in line with the requirements for a minimum threshold advocated by Evert (2005) (see

subsection 4.5.1.2). From this, a collocational candidate set amounting to 1,035 PNV triples was obtained. These PNV triples are of course a proper subset of the 8,644 which were obtained from the 114-million-word corpus.

4.5.2.3 Classification of Candidate Set and Quality Control

In order to manually identify the actual target collocations for our gold standard, we took the collocational candidate set derived from the 100-million-word corpus (i.e., the 8,644 PNV triples) and divided it into three roughly equal-sized portions. Each of them was then given to a human annotator whose task it was to mark the true collocations in the set. All annotators were native speakers of German and graduate students of linguistics. They were given an annotation manual, in which the guidelines included the linguistic properties described in subsubsection 2.1.4.1 and a description of the three collocational classes and how they may be distinguished from free word combinations, as outlined in subsubsection 2.1.4.2. The manual[31] is given in Appendix A at the end of this thesis. Besides the coarse-grained classification of whether a PNV triple candidate was a true collocation or not, the annotators also had to do a three-category fine-grained classification of the collocational targets they identified, i.e. they had to decide whether the collocation was an idiomatic phrase (category 1), a support verb construction or a narrow collocation (category 2), or a fixed phrase (category 3), according to the collocational subtypes established in subsubsection 2.1.4.1. Table 4.10 gives an overview of the proportion of actual PNV triple collocations identified and the subproportions of the three collocational categories in both our large-sized (114 million words) and our small-sized (10 million words) German newspaper corpus.

As can be seen, the proportion of actual collocations in the small-sized corpus amounts to over one third and is thus substantially higher than in the large-sized corpus in which it only reaches a little bit less than 14%. In terms of absolute numbers, increasing the corpus over ten times (i.e. from 10 million words to 114 million words) increases the number of candidates over eight times (from 1,035 to 8,644), but only triples the number of actual collocations (from 335 to 1,180). One effect that this is due to is certainly the frequency cut-off threshold of four, which however is already set as low as possible (see subsection 4.5.1.2).

We have substantiated in subsection 4.5.1.3 that in the case of human annotation

[31] The guidelines, of course, had to be written in German.

	100 million words	10 million words
PNV triple candidates	8,644	1,035
actual collocations	1,180 (13.7%)	355 (34.3%)
idiomatic phrases	700 (59.3%)	185 (52.1%)
support verb constructions / narrow collocations	355 (30.1%)	141 (39.7%)
fixed phrases	125 (10.6%)	29 (8.2%)

Table 4.10: Proportion of actual PNV triple collocations and sub proportions of the three collocational categories.

there needs to be some form of quality control which ensures the reliability of the judgments. This way of measuring the agreement between different human judges is referred to as *inter-annotator* agreement. For this purpose, we randomly selected 800 out of the 8,644 collocation candidates and gave them to each annotator for both coarse-grained and fine-grained classification. Agreement, then, may be calculated simply by the absolute agreement rate or, statistically more sophisticated, by Cohen's Kappa coefficient (Cohen, 1960; Carletta, 1996; Kim & Tsujii, 2006). The absolute agreement rate $P(a)$ is simply the number of times the annotators agree scaled by the number of items to annotate.

$$P(a) = \frac{\text{\# items annotator agree on}}{\text{\# items to annotate}} \tag{4.16}$$

In addition to the absolute agreement $P(a)$, the Kappa coefficient κ also takes into account the expected chance agreement $P(e)$, as given in equation 4.17.

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{4.17}$$

The κ value obtained from this computation may range between -1 and 1. Negative

κ indicates that the absolute agreement is less than chance agreement and positive κ indicates a higher than chance agreement. Having established a kind of benchmark, Landis & Koch (1977) discriminate the ranges of the κ values with designated strengths of agreement, as outlined in table 4.11

κ value	Strength of Agreement
< 0	Poor
0.0 - 0.2	Slight
0.21 - 0.4	Fair
0.41 - 0.6	Moderate
0.61 - 0.8	Substantial
0.81 - 1.0	Almost Perfect

Table 4.11: Ranges of the Kappa coefficient and designated strengths of agreement

For the calculation of the expected chance agreement $P(e)$ on a binary classification task, it is most convenient to establish 2 x 2 contingency tables of observed and expected frequencies along the lines proposed in subsection 3.3.1. We illustrate the observed coarse-grained classifications (i.e., whether a given candidate is a collocation or not) by means of the judgments of two of our linguistic annotators, designated *Annotator 1* and *Annotator 2*, on the 800 randomly selected collocation candidates, given in table 4.12.

		Annotator 1		
		collocation	not collocation	Total
Annotator 2	collocation	79	16	95
	not collocation	30	675	705
	Total	109	691	800

Table 4.12: Observed coarse-grained collocation classifications by two annotators

We can get the observed absolute agreement rate $P(a)$ by adding up the first cell (where both classify the candidates as collocations) and the fourth cell (where both classify the candidates as non-collocations) of the 2 x 2 table (i.e., $79 + 675 = 754$) and scaling it by the number of items to classify (i.e., 800). This yields an absolute agreement rate of 0.94. Now, the expected chance agreement $P(e)$ may be obtained by computing the expected frequencies of these two cells (as described in table 3.2 in subsection 3.3.1) and again scaling the value by 800. Observed and expected agreement values may now be plugged into equation 4.17, resulting in a κ value of 0.74 which indicates *substantial agreement* according to table 4.11 above.

	# Candidates	$P(a)$	κ value
Annotator 1 and 2	800	0.94	0.74
Annotator 2 and 3	800	0.97	0.86
Annotator 1 and 3	800	0.93	0.68
Average	800	0.95	0.76

Table 4.13: Overview of agreement rates and Kappa values for coarse-grained classification.

	# Candidates	$P(a)$	κ value
Annotator 1 and 2	79	0.82	0.65
Annotator 2 and 3	78	0.90	0.79
Annotator 1 and 3	69	0.83	0.66
Average	75.3	0.85	0.7

Table 4.14: Overview of agreement rates and Kappa values for fine-grained classification.

In the case of calculating the κ agreement for the fine-grained classification of collocational candidates (i.e. deciding to which of the three categories an actual collocation belongs to), we may of course only consider those candidates which are actual collocations and on which two annotators agree on their status of being collocations. Because the κ statistics is set up to depend on binary decisions, we calculated the fine-

grained agreements as to whether the annotators decided whether or not an actual collocation was an *idiomatic phrase* or not. Tables 4.13 and 4.14 give an overview of the absolute agreement rates and κ values thus obtained, both for the coarse-grained and the fine-grained classification, respectively

As can be seen from the two tables, with respect to the coarse-grained classification of collocational candidates, all inter-annotator agreements show a high degree of absolute agreement, averaging to 0.95, and a substantial degree of κ agreement, averaging to 0.76. As for the fine-grained classification of actual collocations, the absolute agreement rates are less (with an average of 0.85), reflecting the fact that this task is obviously linguistically much more difficult and intricate. Still, the κ agreement, averaging to 0.7, is still indicative of a substantial degree of agreement, even for this challenging task.

4.5.3 Evaluation Setting for Term Extraction

Because terms may be considered as linguistic expressions which are typically confined to a particular subject domain and sublanguage (cf. section 2.2), in general, the task evaluating different term extraction procedures boils down to choosing such a domain. A key consideration here is to select a subject domain with a large enough text corpus repository in order to compute reliable statistics from it and, ideally, with an already existing and sufficiently comprehensive terminological resource which may serve to automatically classify the term candidate set. Fortunately, the biomedical domain fulfills both of these requirements and, for this reason, it was chosen as the subject domain of choice for this thesis. Thus, in this subsection, we will describe the assembly and linguistic processing of the particular biomedical subdomain text corpus (subsubsection 4.5.3.1), the syntactic target structures from which we obtained our term candidate sets (subsubsection 4.5.3.2), as well as the corresponding biomedical terminological resource for the automatic gold-standard classification of these candidate sets (subsubsection 4.5.3.3).

4.5.3.1 Text Corpus and Linguistic Filtering

The biomedical literature database from which we collect our domain-specific text collection is MEDLINE. This bibliographic repository, which as of February 2007 contains 16 million abstracts from approximately 5,000 selected publications covering

biomedicine and health from 1950 to the present, is hosted by the U.S. National Library of Medicine (NLM)[32] and searchable via the PubMed Entrez system.[33] The field of biomedicine, of course, encompasses a large amount of subdomains and special topics, all of which focus on biological, medical, clinical, pharmaceutical aspects, to name just a few. Therefore, for this thesis, we focus on the subdomain of Hematopoietic Stem Cell Transplantation (HSCT) and Immunology, which lies at the interface between genomic/proteomic research, on the one hand, and medical/clinical application, on the other hand.[34] In order to isolate this subdomain from the Medline text collection, we make use of NLM's controlled indexing vocabulary, the Medical Subject Headings (MeSH)[35] which contains approximately 20,000 terms used to manually add metadata descriptors to Medline abstracts. In order to target the right texts, we selected 35 MeSH terms describing HSCT and Immunology.[36] Then we queried Medline with these indexing terms, setting the publication date range from 1990 to 2006.[37] Because manual indexing consistency for Medline has been reported to be very poor (Funk & Reid, 1983), the MeSH terms were OR-ed in order to ensure a subdomain coverage of abstracts as complete as possible.[38] Then, by running this query, we downloaded approximately 400,000 abstracts form Medline amounting to 100 million words of text material.

This text collection was then linguistically processed by applying a POS tagger and a phrase chunker to it. In particular, we employed the Genia tagger (version 3.0) for this purpose, which performs part-of-speech tagging and phrase chunking for English biomedical text at state-of-the-art performance levels (Tsuruoka & Tsujii, 2005). Because the syntactic target structure for terms are noun phrases, we performed two additional linguistic processing operations on them. First, we filtered out a number

[32] `http://www.nlm.nih.gov/`

[33] `http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed`

[34] HSCT is used for a variety of malignant and nonmalignant disorders to replace a defective host marrow or immune system with a normal donor marrow and immune system.

[35] `http://www.nlm.nih.gov/mesh/`

[36] The MeSH index terms were selected in consultation with a domain expert and are listed in Appendix B.

[37] Typically, publications prior to 1990 are considered outdated for fast-changing subdomains, such as molecular biology.

[38] Also, analogous to our experimental setting for collocations (see subsection 4.5.2.2), the text corpus needs to be large enough to be reduceable to 10% of its original size in order to run our experiments and evaluations on different corpus sizes.

of stop words from the noun phrases in order to reduce the amount of noise. In order to ensure that no potential content words be filtered out, we determined these stop words by their part of speech tag (such as determiners, pronouns, measure symbols etc.)[39] instead of using a stop word list, as is traditionally done. It should be noted that stop words, unlike in the case of collocations, do not function as integral parts of terminological expressions and thus filtering them out is actually a preprocessing step widely employed by other term extraction studies as well (Frantzi et al., 2000; Daille, 1996; Jacquemin, 2001). The second additional processing step concerns the morphological normalization of term candidates, which has shown to be beneficial for term extraction (Nenadić et al., 2004). For this purpose, we normalized the nominal head of each noun phrase (typically the rightmost noun in English) to its base form via the full-form UMLS SPECIALIST LEXICON (Browne et al., 1998), a large repository of both general-language and domain-specific biomedical vocabulary.

4.5.3.2 Target Structures and Candidate Sets

From our linguistically processed corpus, we extracted the noun phrases and counted their occurrence frequencies. In order to obtain term candidate sets which are ample to have LPM scores computed from (see subsubsection 4.4.1 above), we categorized the noun phrases according to their length. In line with the observations put forth by Justeson & Katz (1995), we restricted ourselves to NPs of length 2 (bigrams), length 3 (trigrams) and length 4 (quadgrams) because these are the constructions where the vast majority of terms are typically manifested (see subsection 2.2.7 above).

As we have already pointed out in subsubsection 4.5.1.2 and done for collocations in subsubsection 4.5.2.2, a frequency cut-off threshold needs to be specified thus limiting the number of candidates to be included in the candidate set. Although in the case of classifying the term candidate sets (i.e. identifying the actual terms in them), we do not have to rely on manual classification (as will be explained in subsubsection 4.5.3.3 below), setting the frequency cut-off to the lowest possible value $f > 4$ may not be a good idea either, in particular not for our large 100-million word text corpus (see table 4.15). Because term candidate sets which have been ranked by a lexical association measure are in practice still the input for manual post-inspection by a domain expert,

[39] The same effect is achieved if, instead of filtering out stop POS tags from a noun phrase, only NP-specific POS patterns are specified as a linguistic filter (Justeson & Katz, 1995; Frantzi et al., 2000).

n-gram length	cut-off	NP term candidates	
		tokens	types
bigrams	no cut-off	5,795,447	1,111,248
	$f > 9$	3,991,566	66,669
trigrams	no cut-off	2,963,186	1,620,696
	$f > 7$	960,538	28,499
quadgrams	no cut-off	1,590,591	1,284,759
	$f > 5$	207,661	9,859

Table 4.15: Frequency distribution for n-gram noun phrase term candidate tokens and types for the 100-million-word MEDLINE text corpus

it is advisable to have the frequency cut-off threshold set higher in order to avoid too large ranked output lists. For these reasons, we set the thresholds for the bi-, tri-, and quadgram term candidates to $f > 9$, $f > 7$ and $f > 5$, respectively. As can be seen from table 4.15, even setting the thresholds to these higher levels still yields very large candidate sets, in particular for the bigram candidates. Hence, for such a large corpus, the thresholds may be even set higher in practice.

On the other hand, it is well advisable to set the frequency cut-off to the lowest possible value with respect to our smaller 10-million word text corpus, as can be seen from table 4.16. In particular in the case of quadgrams, the number of candidate types only amounts to 912. This data sparsity, as we will see in section 5.2 below, will have some effect on the performance evaluation.

The fact that the number of observation types drops sharply with increasing n-gram size is, of course, a well-known property observed in the area of language modeling, e.g. for speech recognition (Jurafsky & Martin, 2000), and thus may yield data sparsity which may turn out to be problematic especially for smaller-sized language corpora.

n-gram length	cut-off	NP term candidates	
		tokens	types
bigrams	no cut-off	615,415	206,009
	$f > 4$	357,174	19,001
trigrams	no cut-off	315,071	215,203
	$f > 4$	73,320	4,721
quadgrams	no cut-off	167,396	146,803
	$f > 4$	13,009	912

Table 4.16: Frequency distribution for n-gram noun phrase term candidate tokens and types for the 10-million-word MEDLINE text corpus

4.5.3.3 Classification of Candidate Sets

The vast majority of term extraction studies evaluates the goodness of their extraction procedures by having their ranked output examined by domain experts who identify the true positives among the ranked candidates. Similar to various studies on collocation extraction, typically only the top n candidates on the ranked output list are considered in such an evaluation procedure, with n being rather small ranging from 50 to several hundreds (see subsubsection 4.5.1.3 above). There are also several problems with such an approach for the term extraction evaluation. First, very often only one such expert is consulted and, hence, inter-annotator agreement cannot be determined (as, e.g., in the studies of Frantzi et al. (2000) or Collier et al. (2002)). Furthermore, what constitutes a relevant term for a particular domain may be rather difficult to decide – even for domain experts – when judges are just exposed to a list of candidates without any further context information. Thus, rather than relying on *ad hoc* human judgments in identifying the target terms in a candidate set, as an alternative we may take already existing terminological resources into account, in particular if they have evolved over many years and usually reflect community-wide consensus achieved by expert committees. With these considerations in mind, the biomedical domain is an ideal test bed for evaluating the goodness of term extraction methods because it

hosts one of the most extensive and most carefully curated terminological resources, viz. the UMLS METATHESAURUS (Bodenreider, 2004). Therefore it is possible to take the mere existence of a term in the UMLS as the decision criterion whether or not a candidate term is also recognized as a biomedical term relevant for our subdomain of HSCT and Immunology.

Accordingly, for the purpose of evaluating the quantitative and qualitative performance of our LPM measure against the standard association measures in recognizing multi-word terms from the biomedical literature, we assign every word bigram, trigram, and quadgram in our candidate sets (see tables 4.15 and 4.16 above) the status of being an actual term, if it is found in the 2006 edition of the UMLS METATHESAURUS (UMLS, 2006). In this respect, it is essential that we exclude UMLS vocabularies which are not deemed relevant for HSCT and Immunology. Such vocabularies may include, among others, nursing and health care billing terms and codes. Appendix B of this thesis lists the complete list of UMLS source vocabularies included in our experiments. Thus, in this respect, the word trigram "*open reading frame*" (from our illustration of LPM in subsection 4.4.2) above is listed as a term in one of the UMLS vocabularies considered,[40] whereas "*t cell response*" is not listed anywhere.

	100 million words	10 million words
bigram candidates	66,669	19,001
actual terms	14,054 (21.1%)	5,487 (28.9%)
trigram candidates	28,499	4,721
actual terms	3,459 (12.1%)	1,108 (23.5%)
quadgram candidates	9,859	912
actual terms	890 (9.0%)	204 (22.4%)

Table 4.17: Proportion of actual terms among the bigram, trigram and quadgram term candidates in the large and small corpus.

[40] Actually, it is listed both in MESH and the Gene Ontology (GO) (Gene Ontology Consortium, 2006).

Table 4.17 gives an overview of the different n-gram candidates and their proportions of actual terms determined in this way, both for the large 100-million-word and the small 10-million-word MEDLINE corpus. Similar to what has been determined for the proportion of actual collocations (see subsubsection 4.5.2.3 above), the small corpus also exhibits a substantially higher number of actual targets. Again, the effect responsible for this may be sought in the frequency cut-off of four, which, however, is already set as low as possible. In case of the large corpus, the situation looks a bit different: as can be seen, not only does the number of candidate types drop with increasing n-gram length but also the proportion of actual terms. In fact, their proportion drops more sharply than can actually be seen from the above data because the various cut-off thresholds have a leveling effect.

Chapter 5

Experimental Results

In this chapter, we will report on the experimental results obtained for both the collocation extraction and the term extraction tasks as it was outlined in their evaluation settings described in subsections 4.5.2 and 4.5.3 above. In particular, we will examine – in section 5.1 for collocation extraction and in section 5.2 for term extraction – whether and to what degree the linguistically motivated statistical association measures LSM and LPM perform better than their standard counterparts. We will illuminate these issues from various aspects, relying on the quantitative and qualitative performance metrics introduced in subsection 4.5.1.4 and 4.5.1.6, respectively. Another aspect we focus on is that we run our performance metrics both on the large and on the small text corpora which we assembled and preprocessed linguistically. While the reason for doing so is to ensure that the observed empirical results and differences are not mainly due to corpus size, this aspect also has some practical relevance. Whereas the availability of large general-language text corpora is typically not a problem for the collocation extraction task, there are certainly subject domains for which the amount of electronically available text resources is not as abundant as for the biomedical domain. Finally, section 5.3 offers a comprehensive overall assessment of our experimental results and summarizes the commonalities and differences between our linguistically motivated association measures, with respect to the collocation and the term extraction tasks as well as with respect to the comparative performance evaluations against the standard statistical and information-theoretic measures.

5.1 Experimental Results for Collocation Extraction

In this section, we will present and examine the evaluation results for our performance experiments conducted for the task of collocation extraction from German newspaper text. For this purpose, we will compare our linguistically motivated statistical association measure LSM to the array of standard statistical and information-theoretic association measures presented in section 3.3, *viz.* t-test, frequency of co-occurrence,[1] log-likelihood and pointwise mutual information (PMI). We also experimented with Daille (1994)'s variants of PMI (see subsection 3.3.4) but did not find any substantial difference and thus excluded them from our discussion – also for the sake of maintaining clarity and not overloading the presentation of our results with non-telling association measure variants.

For both our quantitative and our qualitative results (see subsections 5.1.1 and 5.1.2), we present the performance metric data in form of tables and figures, in order to allow different views and perspectives on the results. As for the qualitative results, in particular with respect to the four qualitative criteria, we refrain from visualizing them for all association measures – due to severe "clarity of presentation" concerns – and instead limit ourselves to some illustrative samples. Finally, subsection 5.1.3 also describes the results for the question to what degree there is a marked difference with respect to the linguistic LSM property, both between collocations and non-collocations and among the different types of collocations.

5.1.1 Quantitative Results

We will carry out quantitative performance evaluation for the PNV triple collocation candidates extracted both from our large 114 million word and our small 10 million word German newspaper language corpus. As laid out in subsubsection 4.5.1.4, this kind of evaluation is performed by incrementally examining increasing portions of the ranked output list returned by each of the five association measures examined. For this purpose, we evaluate their performance in terms of precision, recall, F-score, and ROC in a series of four experiments (subsubsection 5.1.1.1). For evaluating precision,

[1] For the sake of brevity, we will label *frequency of co-occurrence* with *frequency* when discussing our experimental results.

it is possible to determine a lower baseline or bound by determining the proportion of targets in the candidate set. Another much more challenging baseline, which also fits for the other performance metrics (i.e. recall, F-score, and ROC), is the actual easy-to-implement frequency measure. Conversely, the performance of the various association measures is also compared against an optimal measure which gives a sort of upper bound for the collocation extraction task. Finally, employing the McNemar test, we also compare the ranked outputs of various association measures and test whether the differences are statistically significant (subsubsection 5.1.1.2).

5.1.1.1 Results on Performance Metrics

In the first series of quantitative experiments, we incrementally measured the performance of the various association scores in terms of their precision. For our large corpus, the results are visualized in figure 5.1 whereas the corresponding scores are given in the upper part of table 5.1 at incremental intervals of 10 percentage points of the ranked output list. As can be seen from both views on the results, the linguistically motivated LSM association measure holds a constant advantage over the next placed frequency measure, starting from 10 points at 1% of the ranked output list ($\approx$ 86 candidates) and gradually decreasing. Whereas the precision values for frequency and t-test cluster together, log-likelihood is below and PMI actually even underperforms the baseline.

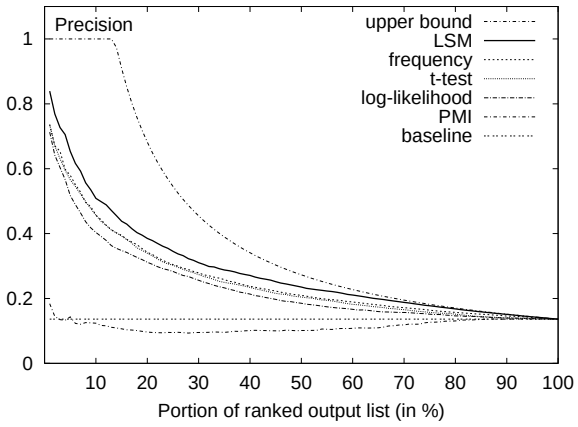

Figure 5.1: Collocation precision on 114 million word corpus

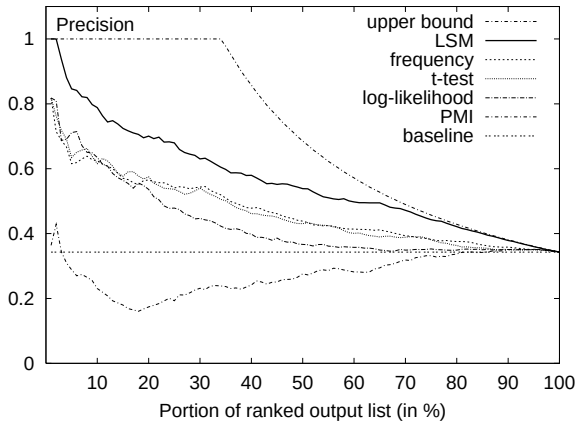

Figure 5.2: Collocation precision on 10 million word corpus

Ranked list portion	Precision scores (8,644 PNV triple candidates; 1,180 targets						
	upper bound	LSM	frequency	t-test	log likelihood	PMI	baseline
1%	1.00	0.84	0.74	0.74	0.71	0.18	0.14
10%	1.00	0.51	0.46	0.46	0.40	0.12	0.14
20%	0.68	0.39	0.34	0.34	0.31	0.10	0.14
30%	0.44	0.30	0.27	0.27	0.25	0.10	0.14
40%	0.34	0.27	0.24	0.23	0.21	0.10	0.14

Ranked list portion	Precision scores (1,035 PNV triple candidates; 355 targets						
	upper bound	LSM	frequency	t-test	log likelihood	PMI	baseline
1%	1.00	1.00	0.82	0.82	0.82	0.36	0.34
10%	1.00	0.79	0.62	0.62	0.63	0.23	0.34
20%	1.00	0.70	0.57	0.57	0.54	0.17	0.34
30%	1.00	0.63	0.55	0.53	0.44	0.23	0.34
40%	0.86	0.58	0.48	0.46	0.40	0.24	0.34

Table 5.1: Precision scores of association measures for collocation extraction on the 114 million word (upper table) and 10 million word (lower table) German newspaper corpus.

The precision score results for our smaller 10 million word corpus are given in the lower part of table 5.1 and the corresponding visualization is shown in figure 5.2. As can be clearly seen, the advantage of LSM with respect to the statistical and information-theoretic association measures is much more pronounced. At 1% of the ranked output list ($\approx$ 11 candidates), LSM runs almost 20 points higher than the next placed measures frequency, t-test and log-likelihood. In fact, its performance is optimal, as can be seen from the comparison with the upper bound limit. Although the advantage becomes smaller with increasing output portions, the lead still runs about 11 points at the 60% portion. At later portions again, all the measures converge toward the lower baseline, which here runs 20 points higher than for the large corpus (0.34 versus 0.14). In general, it is interesting to note that after 20% ($\approx$ 207 candidates) to 30% ($\approx$ 311 candidates) of the output scanned, the precision scores for frequency and t-test may be grouped together followed by log-likelihood, as can be also seen in the

plotted precision graphs. This tendency, although less pronounced, is also observable on the large corpus (see figure 5.1 and the upper table 5.1 above). In a similar vein than for the large corpus, it is again the PMI measure which underperforms the baseline, here even more substantially.

In a second series of quantitative experiments, we incrementally measured the performance of the various association measures in terms of their recall, again both for our large and for our small corpus. Because recall measures the proportion of selected targets at a certain point in the ranked output list, it has a particularly practical relevance because the output lists produced by various association measures are typically post-examined by a human. Hence, the earlier a large proportion of targets is returned, the more efficient an association method may be considered.

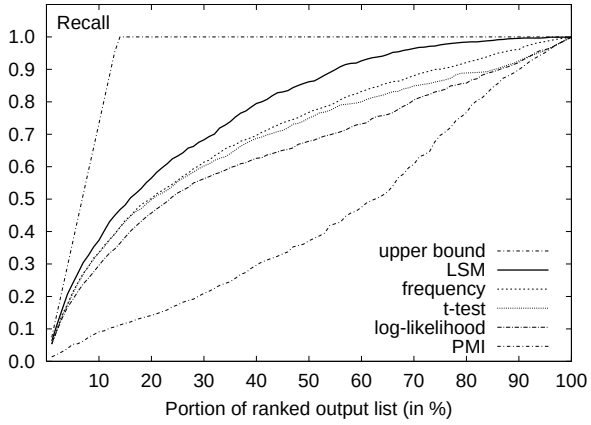

Figure 5.3: Collocation recall on 114 million word corpus

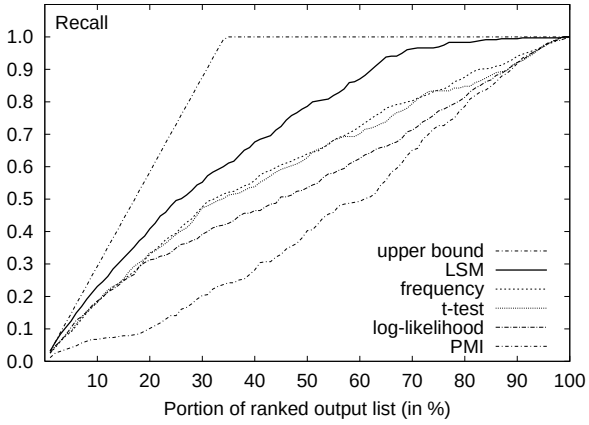

Figure 5.4: Collocation recall on 10 million word corpus

The recall results of the 114 million word corpus are visualized in figure 5.3 and the corresponding scores are given in the upper part of table 5.2, again at incremental intervals of 10 percentage points of the ranked output list. As can be seen from the plot, our linguistically motivated LSM association measure again has a clear advantage compared to the standard statistical and information-theoretic measures. This advantage is most pronounced at the 60% portion of the ranked list where LSM exhibits a 10-point higher recall than the second-placed frequency measure. But, as can be seen both from the plot and recall scores, this clear advantage already prevails at much lower portions of the output. When examining 20% ($\approx$ 1729 candidates), 30% ($\approx$ 2593 candidates) and 40% ($\approx$ 3458 candidates), LSM already identifies almost

60%, 70% and 80% of all targets. In order to return 90% of all targets, LSM only needs to scan 55% of the output (≈ 4754) whereas frequency needs to examine 75% (≈ 6483) and t-test 85% (≈ 7347) to reach such a high level of recall. In addition, a similar pattern like the one observed in the lower part of table 5.1 and figure 5.2 for the precision results on the small corpus appears in that the scores (and thus the curves) for frequency and t-test cluster together, indicating these measures perform similarly. Whereas log-likelihood runs below these two, the information-theoretic PMI again severely underperforms compared to all the other association measures.

Ranked list portion	Recall scores (8,644 PNV triple candidates; 1,180 targets					
	upper bound	LSM	frequency	t-test	log likelihood	PMI
20%	1.00	0.56	0.50	0.50	0.46	0.14
30%	1.00	0.69	0.62	0.61	0.57	0.22
40%	1.00	0.79	0.70	0.69	0.63	0.30
50%	1.00	0.86	0.77	0.75	0.68	0.37
60%	1.00	0.93	0.83	0.80	0.73	0.47
70%	1.00	0.96	0.88	0.85	0.81	0.61
80%	1.00	0.98	0.92	0.89	0.86	0.77
90%	1.00	1.00	0.96	0.93	0.92	0.90

Ranked list portion	Recall scores (1,035 PNV triple candidates; 355 targets					
	upper bound	LSM	frequency	t-test	log likelihood	PMI
20%	0.58	0.41	0.33	0.34	0.31	0.10
30%	0.90	0.57	0.49	0.48	0.40	0.21
40%	1.00	0.68	0.56	0.54	0.46	0.28
50%	1.00	0.79	0.64	0.63	0.54	0.40
60%	1.00	0.87	0.72	0.70	0.63	0.49
70%	1.00	0.96	0.80	0.79	0.71	0.65
80%	1.00	0.98	0.88	0.85	0.81	0.79
90%	1.00	0.99	0.94	0.92	0.92	0.92

Table 5.2: Recall scores of association measures for collocation extraction on the 114 million word (upper table) and 10 million word (lower table) German newspaper corpus.

The recall scores for our smaller 10 million word corpus are given in the lower part of table 5.2 and the corresponding plot visualization is shown in figure 5.4. Similar to the large corpus, frequency and t-test may be grouped together followed by log-likelihood, while PMI again performs substantially worse. Compared to the second best frequency measure, the point advantage of LSM is even much more pronounced for the smaller corpus than for the larger one and reaches its peak with 16 points (0.96 vs. 0.8 recall) at the 70% portion ($\approx$ 725 candidates). Similar to the large corpus, LSM needs to scan a much smaller portion of the ranked output list (62% – $\approx$ 642 candidates) in order to reach 0.9 recall than the next placed frequency and t-test with 83% ($\approx$ 859) and 88% ($\approx$ 911), respectively. As already observed with respect to the precision and recall scores above, it can be also seen here that, with respect to the upper bound optimum, there is still room for substantial improvement.

The third series of quantitative experiments combines both the precision and recall scores into the balanced F-score, the results of which are given in figure 5.5 and the upper table 5.3 for our large corpus. As can be seen, besides PMI, all association measures reach the peak F-score already at an early portion of the ranked output list, i.e. at around 15% ($\approx$ 1383 candidates). Naturally, the order of performance at this portion reflects the singular precision and recall scores, with LSM performing best (0.46 F-score) followed by frequency (0.41) and t-test (0.4) and then log-likelihood (around 0.37).

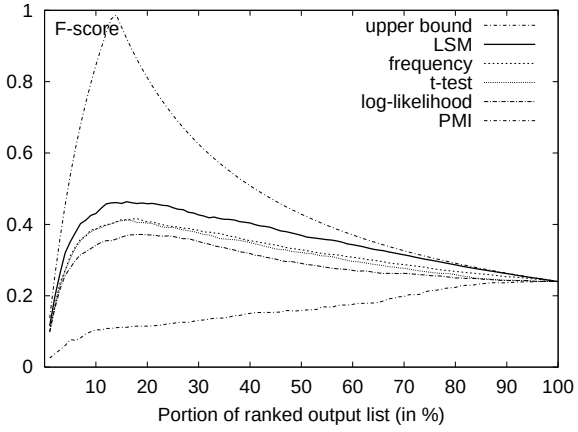

Figure 5.5: Collocation F-score on 114 million word corpus

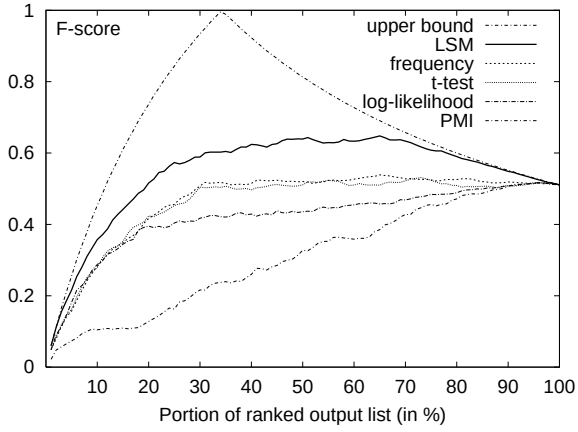

Figure 5.6: Collocation F-score on 10 million word corpus

Ranked list portion	F-scores (8,644 PNV triple candidates; 1,180 targets					
	upper bound	LSM	frequency	t-test	log likelihood	PMI
10%	0.85	0.43	0.39	0.39	0.34	0.10
20%	0.81	0.46	0.41	0.40	0.37	0.11
30%	0.61	0.42	0.38	0.37	0.35	0.13
40%	0.51	0.40	0.35	0.35	0.32	0.15
50%	0.43	0.37	0.33	0.32	0.29	0.16
60%	0.37	0.34	0.31	0.30	0.27	0.18
70%	0.33	0.31	0.29	0.28	0.26	0.20
80%	0.29	0.29	0.27	0.26	0.25	0.22

Ranked list portion	F-scores (1,035 PNV triple candidates; 355 targets					
	upper bound	LSM	frequency	t-test	log likelihood	PMI
10%	0.45	0.36	0.28	0.28	0.29	0.10
20%	0.74	0.52	0.42	0.42	0.40	0.13
30%	0.95	0.60	0.52	0.50	0.42	0.22
40%	0.92	0.62	0.51	0.50	0.43	0.26
50%	0.81	0.64	0.52	0.51	0.44	0.33
60%	0.73	0.63	0.53	0.51	0.45	0.36
70%	0.66	0.63	0.53	0.52	0.47	0.43
80%	0.60	0.59	0.53	0.51	0.49	0.47

Table 5.3: F-scores of association measures for collocation extraction on the 114 million word (upper table) and 10 million word (lower table) German newspaper corpus.

Comparing the F-score results of the large corpus to those of the small corpus (see lower part of table 5.3) shows that, in case of the latter one, all the association measures reach their peak at a much later portion (at around 65% and more [i.e. after $\approx$ 673 candidates] of the output list scanned) and the corresponding F-scores are much higher. Whereas LSM reaches a peak F-score of around 0.65, frequency attains its 0.54 peak at the same portion as LSM but the t-test arrives at its peak of 0.53 F-score only at the 73% portion (after $\approx$ 756 candidates). Whereas these three

measures decrease in F-scores after reaching their peaks, log-likelihood gradually rises almost until the very end of the output list is reached. Again, PMI underperforms all other measures considerably.

There is an obvious reason why the F-score peaks are much higher in the case of the small corpus. Because the proportion of targets (i.e. actual collocations) is much higher for the 10 million word corpus (34.3%) than for the 114 million word corpus (13.7%), the higher ratio of selected non-targets depresses the precision values for the large corpus much more noticeably, which, in turn, has an effect on the balanced F-score. This is very nicely visualized both in plot 5.6 and plot 5.5 in which it can be seen that the optimal upper bound F-score is only reached at the 34% portion for the small corpus whereas this is already the case at 14% portion for the large corpus.[2] In other words, because an optimal upper bound selects all targets to the top of the output list, this idealized status is reached much earlier for the large corpus than for the small one.

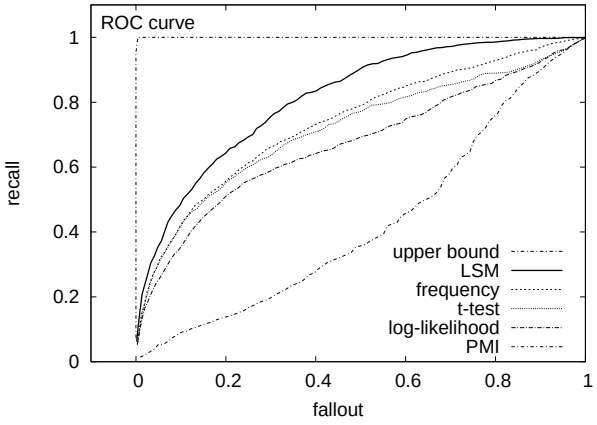

Figure 5.7: Collocation ROC curve on 114 million word corpus

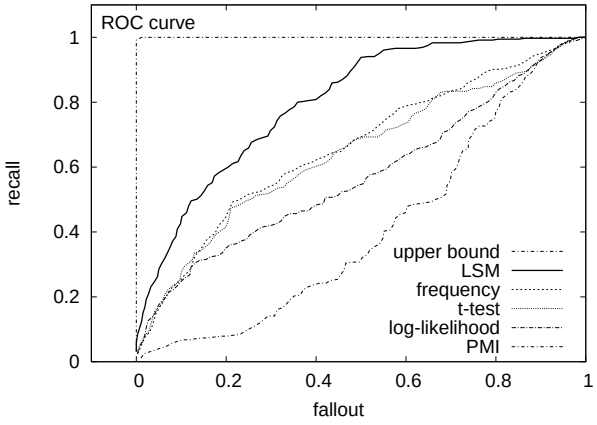

Figure 5.8: Collocation ROC curve on 10 million word corpus

Finally, the fourth series of quantitative experiments computes the fallout scores and plots these against the recall scores into ROC curves. When examining the fallout scores of the various association measures, as given in the upper part of table 5.4, the high proportion of non-targets among the 8,644 PNV triple candidates from the large corpus is reflected.

[2]Contrary to what the plots might suggest, an optimal F-score of 1.0 is actually never reached.

Ranked list portion	Fallout scores (8,644 PNV triple candidates; 1,180 targets					
	upper bound	LSM	fre-quency	t-test	log likelihood	PMI
10%	0.00	0.06	0.06	0.06	0.07	0.10
20%	0.07	0.14	0.15	0.15	0.16	0.21
30%	0.20	0.24	0.26	0.26	0.27	0.32
40%	0.31	0.34	0.35	0.35	0.36	0.42
50%	0.42	0.44	0.46	0.46	0.47	0.52
60%	0.54	0.55	0.56	0.57	0.58	0.62
70%	0.65	0.66	0.67	0.68	0.68	0.71
80%	0.77	0.77	0.78	0.79	0.79	0.81

Ranked list portion	Fallout scores (1,035 PNV triple candidates; 355 targets					
	upper bound	LSM	fre-quency	t-test	log likelihood	PMI
10%	0.00	0.03	0.06	0.06	0.06	0.12
20%	0.00	0.09	0.13	0.13	0.14	0.25
30%	0.00	0.17	0.21	0.22	0.26	0.36
40%	0.09	0.26	0.32	0.33	0.37	0.46
50%	0.24	0.35	0.43	0.43	0.48	0.55
60%	0.39	0.46	0.54	0.55	0.59	0.66
70%	0.54	0.56	0.65	0.65	0.69	0.73
80%	0.70	0.70	0.76	0.78	0.79	0.81

Table 5.4: Fallout scores of association measures for collocation extraction on the 114 million word (upper table) and 10 million word (lower table) German newspaper corpus.

As can be seen, there is no big difference in the fallout scores (i.e. the rate of non-targets selected) among the various association measures examined at increasing portions of the ranked output list.[3] In fact, after a third of the output, there is not even a big difference between the optimal upper limit and the other association measures. Hence, only looking at the mere fallout scores is not very informative and for this reason (see subsection 4.5.1.4), fallout is typically plotted against recall thus yielding the so-called receiver operation characteristic (ROC) curve, which is visualized in figure 5.7 for the large corpus. As can be seen, the effect of the superior recall scores of

[3]Of course, the PMI measure is an exception falling out of the line again and illustrating its poor performance by higher fallout scores.

LSM compared to the other association measures (as previously outlined in figure 5.3 and the upper table 5.2) come to full effect here, as LSM follows the left-hand border and then the top border of the ROC space much more closely than its competitors.

Looking at the ROC curves for the small 10 million word corpus shown in figure 5.8, a similar picture is depicted with LSM again following the left border of the ROC space much more closely and its competitors even falling more behind. If we look at the actual fallout scores from the lower table 5.4, however, it can be seen that the higher proportion of targets has its effects on the various fallout rates in that there is again a noticeable difference between the top-performing LSM measure, on the one hand, and the next placed frequency and t-test and log-likelihood measures, on the other hand. In any case, these results show that for the small corpus, the fallout scores have an effect on the actual ROC curve whereas for the large corpus, the differences among the association measures are almost exclusively driven by the recall scores.

5.1.1.2 Results on Significance Testing

The main reason for applying significance testing to the results of quantitative performance evaluations, such as the ones we have done in the previous subsubsection, is that it may be the case that in comparing association measures (or systems in general), the observed differences are rather small thus raising the question whether they are existent at all or merely due to chance (see subsection 4.5.1.5). Although the results reported above already point towards a clear advantage of the linguistically motivated LSM measure, we will corroborate these findings by applying the McNemar test as a significance test of differences to incremental portion measure points of the ranked output list, both on the large and on the small corpus, as laid out in subsection 4.5.1.5. For this purpose, we selected 100 measure points in the ranked list, one after each increment of one percent, and then applied the two-tailed test for a (very strict) confidence interval of 99%. In particular, we tested the significance of differences between LSM and the two next best performing association measures, frequency and t-test. For the large and the small corpus, the results at 10-point increments are given in table 5.5.

As can be seen from the number of significant differences, the clear advantage that LSM exhibited in comparison to its competitors on all the quantitative performance metrics translates right into the results for the McNemar test, both for the large and

# of measure points considered	Large corpus: # of significant differences comparing LSM with		Small Corpus: # of significant differences comparing LSM with	
	frequency	t-test	frequency	t-test
10	9	9	6	6
20	19	19	16	16
30	29	29	27	26
40	39	39	36	36
50	49	49	46	46
60	59	59	56	56
70	69	69	66	66
80	79	79	76	76
90	89	89	86	86
100	95	97	90	90

Table 5.5: Collocation extraction: significance testing of differences using the two-tailed McNemar test at 99% confidence interval on the large and the small corpus

the small corpus. In both cases, at measure point 10 the small corpus exhibits less points of differences than the big one (4 points less differences versus 1 point less), both comparing LSM to frequency and to t-test. But then, until measure point 90, this ratio is completely kept and it is not until the last measure point 100 that again a comparatively small amount of significant differences is taken away in all cases. This, of course, is not astonishing because as can be seen in the figures from the previous subsection, the various association measure performance curves converge near the end of the ranked output list for virtually all performance metrics.

5.1.2 Qualitative Results

In subsubsection 4.5.1.6, we have formulated four achievement objectives for the qualitative performance evaluation of lexical association measures and took frequency of co-occurrence as as a sort of baseline against which a particular association measure should re-rank (or not) the targets and non-targets of the candidate set. Accordingly, the four objectives may be taken to be two static criteria (subsubsection 5.1.2.1) and two dynamic criteria (subsubsection 5.1.2.2). As explained, we choose the middle

rank as a mark to divide a ranked output list into an upper portion and a lower portion. Then the targets and non-targets assigned to these portions by frequency may be examined and quantified, according to the four criteria, to what degree the other association measures changed these rankings or not. In order to better quantify the degrees of movement, we partitioned both the upper and the lower portions into three further subportions. Of course, frequency is quite a competitive baseline, given its quantitative performance shown in the previous section.

5.1.2.1 Results on the Static Criteria

The first two criteria examine how static an association measure is in that a qualitatively superior association measure should at least keep the status quo with respect to frequency. In this respect, criterion 1 examines whether a measure is able to keep the targets (i.e. the true collocations) in the upper portion. The first part of table 5.6 shows this for the large corpus.

	AM	upper portion (ranks 1 - 4322)			lower portion (ranks 4323 - 8644)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 1	freq	545 (60.2%)	216 (23.9%)	144 (15.9%)	0	0	0
	t-test	540 (59.7%)	198 (21.9%)	115 (12.7%)	9 (1.0%)	12 (1.3%)	12 (1.3%)
905	logL	482 (53.3%)	168 (18.6%)	81 (9.0%)	69 (7.6%)	45 (5.0%)	60 (6.6%)
Ts	PMI	103 (11.4%)	96 (10.6%)	136 (15.0%)	176 (19.4%)	240 (26.5%)	154 (17.0%)
	LSM	606 (67.0%)	237 (26.2%)	35 (3.9%)	10 (1.1%)	12 (1.3%)	5 (0.6%)
Crit. 2	freq	0	0	0	1326 (32.8%)	1357 (33.5%)	1365 (33.7%)
	t-test	0	0	362 (8.9%)	1247 (30.8%)	1323 (32.7%)	1116 (27.6%)
4048	logL	4 (0.1%)	366 (9.0%)	818 (20.2%)	952 (23.5%)	968 (23.9%)	940 (23.2%)
NTs	PMI	820 (20.3%)	725 (17.9%)	687 (16.9%)	581 (14.4%)	589 (14.6%)	646 (16.0%)
	LSM	0	41 (1.0%)	877 (21.7%)	1163 (28.7%)	977 (24.1%)	990 (24.5%)

Table 5.6: Results on the two static criteria for upper-portion targets (Ts) and lower-portion non-targets (NTs) on the large corpus.

Compared to the frequency baseline, none of the association measures is able to keep all their targets in the upper portion; rather, they tend to demote some of them to the lower subportions, in quite different degrees, however. Log-likelihood demotes a higher percentage down to the lower subportions than LSM and t-test. By far the biggest proportion is demoted by PMI which almost demotes two thirds (63%) of the

targets. Hence, the reason for its severe underperformance regarding the quantitative performance metrics in the last subsections already becomes evident at this point. The frequency baseline places 60.2% of the targets into the first upper subportion and 24% into the second one. Only LSM is able to even improve on this by placing 67% of the targets into the first upper subportion and 26.2% into the second one.

The second part of table 5.6 gives the results for criterion 2, i.e. keeping the non-targets in their lower portion place. As can be seen, none of the association measures considered is able to achieve this goal completely whereby some again perform better than others. The best performance is delivered by t-test which only places 8.9% of its lower-portion non-targets into the third upper subportion whereas LSM already moves more non-targets upwards (22.%). Still, log-likelihood performs worse and by far the most non-targets are promoted by PMI which actually places 55% of its lower-portion non-targets into the upper subportions.

	AM	upper portion (ranks 1 - 517)			lower portion (ranks 518 - 1035)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 1	freq	95 (41.9%)	86 (37.9%)	46 (20.3%)	0	0	0
	t-test	101 (44.5%)	76 (33.5%)	34 (15.0%)	19 (8.4%)	3 (1.3%)	3 (1.3%)
227	logL	93 (41.0%)	45 (19.8%)	20 (8.8%)	30 (13.2%)	21 (9.3%)	18 (7.9%)
Ts	PMI	7 (3.1%)	41 (18.1%)	38 (16.7%)	46 (20.3%)	58 (25.6%)	37 (16.3%)
	LSM	122 (53.7%)	88 (38.8%)	11 (4.8%)	0	5 (2.2%)	1 (0.4%)
Crit. 2	freq	0	0	0	138 (35.5%)	132 (33.9%)	119 (30.1%)
	t-test	0	0	39 (10.0%)	110 (28.3%)	132 (33.9%)	108 (27.8%)
389	logL	0	42 (10.8%)	98 (25.2%)	80 (20.6%)	84 (21.6%)	85 (21.9%)
NTs	PMI	102 (26.2%)	71 (18.3%)	62 (15.9%)	51 (13.1%)	51 (13.1%)	52 (13,4%)
	LSM	0	0	80 (20.6%)	102 (26.2%)	94 (24.2%)	113 (29.0%)

Table 5.7: Results on the two static criteria for upper-portion targets (Ts) and lower-portion non-targets (NTs) on the small corpus.

For our small corpus, the result scores given in table 5.7 for the two static criteria show quite similar patterns. Concerning criterion 1, LSM is again best at keeping the targets in the upper portion of the ranked output followed by t-test. As is the case for the large corpus, log-likelihood performs similar and already loses a higher proportion of its targets to the lower subportions. By far the worst performance is again delivered by PMI which demotes 62.2% of the targets to the lower three subportions. Unlike

all other association measures and in line with the large corpus results, LSM is again able to even increase the rate of targets in the third upper subportion by almost 12 points compared to the frequency baseline (from 41.9% to 53.7%).

With respect to criterion 2, in a similar vein to the large corpus, t-test keeps the most non-targets in the lower portion of ranked output (90%) followed by LSM which keeps 79.4%. Next again comes log-likelihood which is able to keep roughly one third of their non-targets in the lower portion (36%). PMI again shows the same poor performance as on the large corpus.

5.1.2.2 Results on the Dynamic Criteria

The third and fourth criteria examine how dynamic an association measure is in that a qualitatively superior association measure should change and improve the rankings with respect to frequency. In this respect, criterion 3 examines whether an association measure is able to demote the non-targets (i.e. the non-collocations) from the upper to the lower portion. Table 5.8 shows the results for this on the large corpus.

	AM	upper portion (ranks 1 - 4322)			lower portion (ranks 4323 - 8644)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 3	freq	896 (26.2%)	1225 (35.9%)	1296 (37.9%)	0	0	0
	t-test	901 (26.4%)	1243 (36.4%)	932 (27.3%)	95 (2.8%)	47 (1.4%)	199 (5.8%)
3417	logL	953 (27.9%)	871 (25.5%)	510 (14.9%)	376 (11.0%)	343 (10.0%)	364 (10.7%)
NTs	PMI	471 (13.8%)	590 (17.3%)	592 (17.3%)	635 (18.6%)	549 (16.1%)	580 (17.0%)
	LSM	835 (24.4%)	1150 (33.7%)	342 (10.0%)	218 (6.4%)	378 (11.1%)	494 (14.5%)

Table 5.8: Results on the dynamic qualitative criteria 3 for upper-portion non-targets (NTs) on the large corpus.

As can be seen, LSM demotes one third (32%) of all its upper-portion non-targets to the lower three subportions, slightly more than log-likelihood with 31.7%. The least degree of re-ranking is performed by t-test which actually keeps 90% of the non-targets in the upper three subportions. PMI actually demotes half of the non-targets into the lower portion but this is of course also what it does on the two static criteria.

A visualized view on the degree of these re-rankings is offered by the scatterplots in figure 5.9 in which the rankings of the upper portion non-targets of frequency are plotted against their ranking in the other association scores. Here it can be seen that,

in terms of the rank subportions considered, the t-test non-targets are concentrated along the same line as frequency non-targets, with only a few being able to break this line and get demoted to a lower subportion. It is interesting to note that LSM demotes the non-targets in quite a structured way, aligning them on distinct lower subportion ranking layers. It also interesting both in the case of LSM of t-test, some upper-portion non-targets are even promoted (instead of demoted) compared to their frequency ranking.

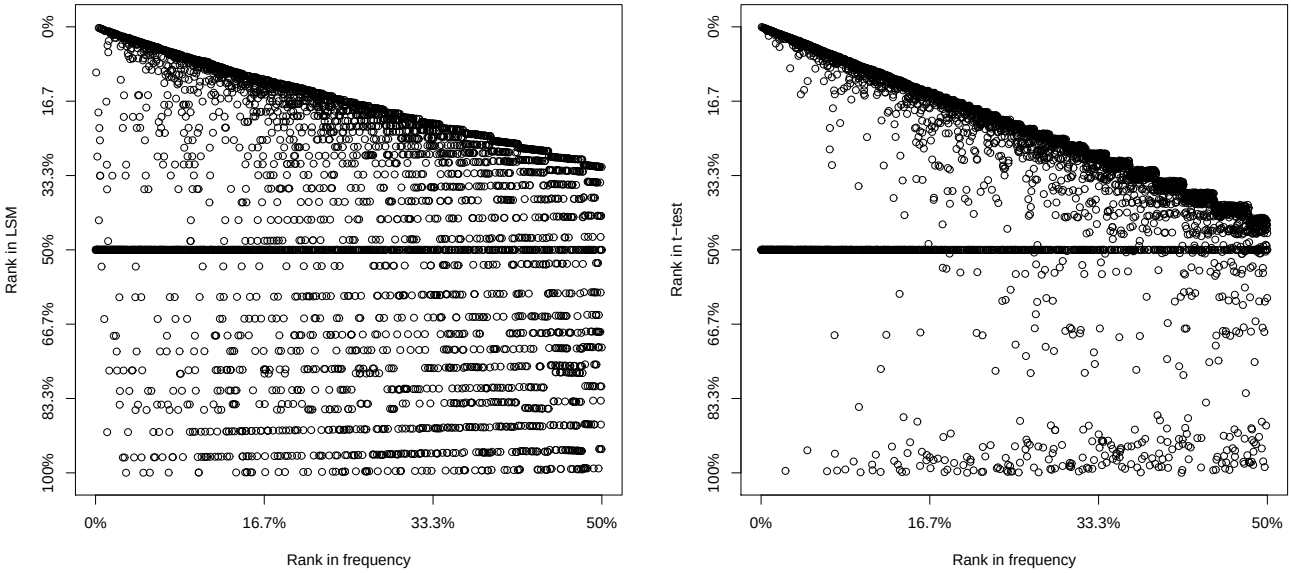

Figure 5.9: Qualitative criterion 3: non-targets moved from upper to lower portion (Left: LSM rank compared to frequency rank. Right: t-test rank compared to frequency rank).

The results for the second dynamic criterion (criterion 4 – promoting the lower portion targets into the upper portion of the ranked output compared to frequency) are given for the large corpus in table 5.9. As can be seen, LSM is able to promote 56.2% of the lower-portion targets into the second and third upper subportion. What's more, also in the first lower subportion a high proportion of targets (30.2%) is concentrated whereas the last two lower subportion only show relatively small amounts of targets. Hence, it is this promotion of lower-portion targets into the upper and middle ranks that seems to trigger the 90% recall that LSM achieves already after scanning only 55% of the ranked output list (see table 5.2 and figure 5.3 in subsubsection 5.1.1.1 above). Again, PMI's behavior for this criterion is as already previously observed for the other criteria.

	AM	upper portion (ranks 1 - 4322)			lower portion (ranks 4323 - 8644)		
Crit. 4	freq	0	0	0	113 (41.2%)	85 (31.0%)	76 (27.7%)
	t-test	0	0	31 (11.3%)	88 (32.1%)	59 (21.5%)	96 (35.0%)
274	logL	2 (0.7%)	36 (13.1%)	31 (11.3%)	42 (15.3%)	85 (31.0%)	78 (28.5%)
Ts	PMI	48 (17.5%)	30 (10.9%)	25 (9.1%)	47 (17.2%)	64 (23.4%)	60 (21.9%)
	LSM	0	10 (3.6%)	144 (52.6%)	84 (30.7%)	27 (9.9%)	9 (3.3%)

Table 5.9: Results on the dynamic qualitative criteria 4 for lower-portion targets (Ts) on the large corpus.

The results for criterion 4 are again also given from a visual perspective in figure 5.10. Regarding the t-test measure, it can be seen in the right scatterplot that is only very modestly successful in meeting this criterion. This, however, appears to be in line with its results on the three other criteria as t-test changes the frequency rankings the least. Also, LSM's promotion of lower-portion targets and their concentration in the upper middle ranks is nicely illustrated in the left plot.

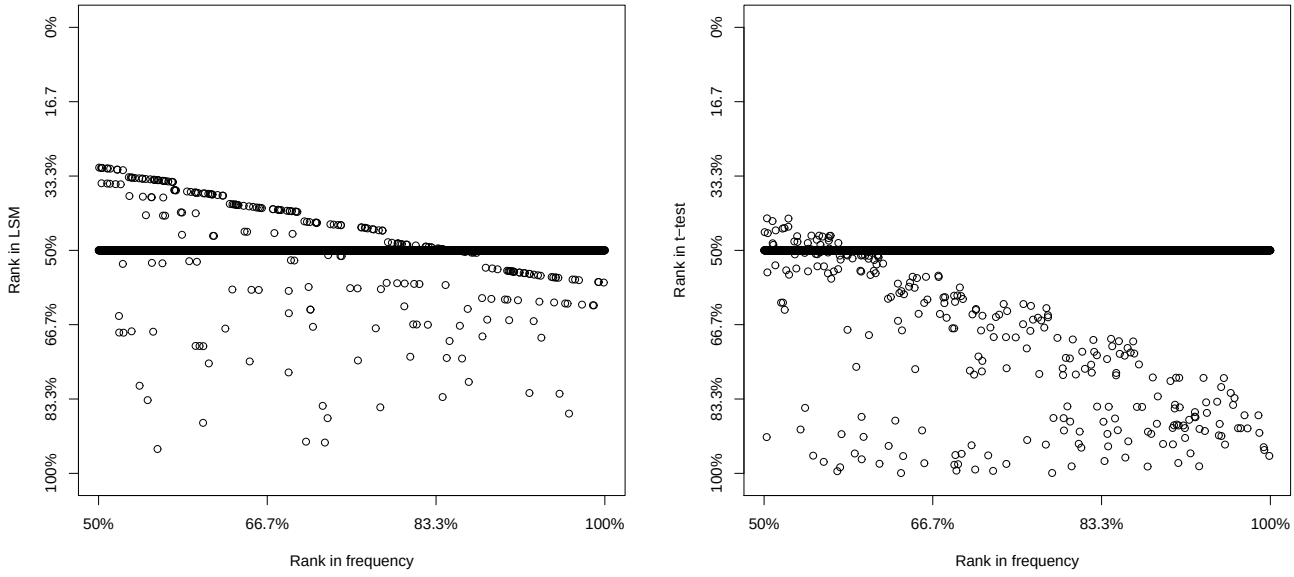

Figure 5.10: Qualitative criterion 4: targets moved from lower to upper portion (Left: LSM rank compared to frequency rank. Right: t-test rank compared to frequency rank).

Last, we examine table 5.10 which shows the results on the two dynamic criteria for the small corpus. For criterion 3 (moving the upper-portion non-targets to the lower portion compared to the frequency baseline), the results exhibit similar patterns

as for the large corpus shown in table 5.8 above. But in particular LSM is even more successful by actually demoting 45.5% of the non-targets to the lower subportions. Log-likelihood manages to demote a small amount more compared to the large corpus and t-test, again, is the association measure which stays the most in line with frequency.

	AM	upper portion (ranks 1 - 517)			lower portion (ranks 518 - 1035)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 3	freq	77 (26.6%)	87 (30.0%)	126 (43.3%)	0	0	0
	t-test	71 (24.5%)	97 (33.4%)	87 (30.0%)	13 (4.5%)	4 (1.4%)	18 (6.2%)
290	logL	79 (27.2%)	76 (26.2%)	32 (11.0%)	42 (14.5%)	24 (8.3%)	37 (12.8%)
NTs	PMI	41 (14.1%)	48 (16.6%)	50 (17.2%)	53 (18.3%)	37 (12.6%)	61 (21.0%)
	LSM	50 (17.2%)	85 (29.3%)	23 (7.9%)	14 (4.8%)	62 (21.4%)	56 (19.3%)
Crit. 4	freq	0	0	0	52 (40.6%)	41 (32.0%)	35 (27.3%)
	t-test	0	0	12 (9.4%)	38 (29.7%)	34 (26.6%)	44 (34.4%)
128	logL	0	10 (7.8%)	22 (17.2%)	19 (14.8%)	44 (34.4%)	33 (25.8%)
Ts	PMI	23 (18.0%)	27 (21.1%)	21 (16.4%)	22 (17.2%)	13 (10.2%)	22 (17.2%)
	LSM	3 (2.3%)	12 (9.4%)	55 (43.0%)	58 (45.3%)	0	0

Table 5.10: Results on the two dynamic qualitative criteria for upper-portion non-targets (NTs) and lower-portion targets (Ts) on the small corpus.

Also for criterion 4 (promoting the lower-portion targets to the upper portion), the results on the small corpus exhibit notable similarities to those on the large corpus from table 5.9 above. Again LSM is able to promote over one half (54.7%) of the targets into the upper three subportions and again the majority is placed in the third upper subportion (43%). At the same time, a big chunk (45.3%) is placed in the first lower subportion and the last two lower subportions actually do not contain any targets any more at all. This again appears to be responsible for LSM obtaining a 0.9 recall after scanning 60% of the ranked output list, as shown in table 5.2 and figure 5.4 in subsection 5.1.1.1 above. Also the statistical (t-test and log-likelihood) and information-theoretic (PMI) association measures show comparable results patterns as on the large corpus.

5.1.3 Limited Syntagmatic Modifiability Revisited

The previous subsections showed that a measure for collocation discovery which takes into account the linguistic property of limited syntagmatic modifiability fares significantly better than linguistically not so founded, purely statistical or information-theoretic measures. Although the LSM property has been stated in linguistic research on collocations (see section 2.1 above), it has not yet been empirically evaluated. Thus, we ran an experiment which took both the PNV triples classified as collocations and the PNV triples classified as non-collocations from our large corpus and counted the numbers of distinct syntagmatic attachments. From this data, we set up a distribution of collocational and non-collocational PNV triples in which the distributional ranking criterion was the number of distinct syntagmatic attachments, which is visualized in figure 5.11.

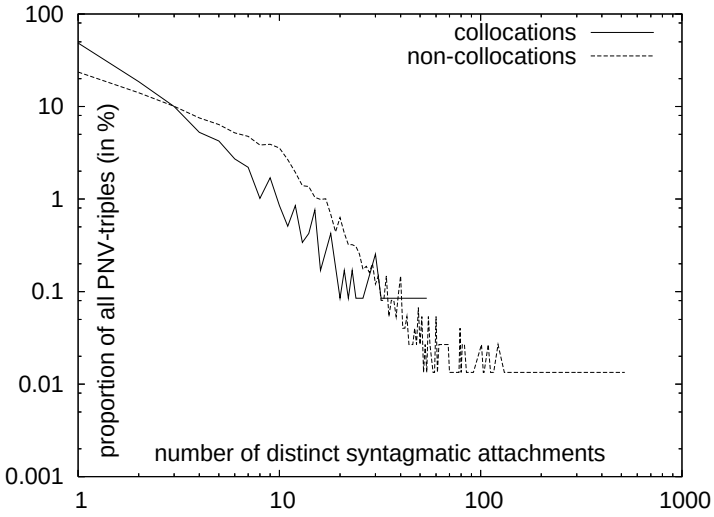

Figure 5.11: Distribution of syntagmatic attachments for collocations and non-collocations. The x- and y-axes are log-scaled to improve visibility.

As figure 5.11 reveals, the proportion of collocational PNV triples with only one distinct syntagmatic attachment[4] almost covers half of all collocational PNV triples (49%). In contrast, the proportion for non-collocational PNV triples with one distinct syntagmatic attachment is only half as high (24%). In addition, with each

[4]In fact, this is actually the zero attachment (see subsection 4.3.1 on the definition of LSM) of the PNV triple, i.e. the one for which no syntagmatic attachment occurs in the first place.

additional syntagmatic attachment, the collocational proportion curve declines more steeply than its non-collocational counterpart. Moreover, the collocational proportion curve already ends with 54 distinct attachments, whereas the non-collocational proportion curve leads up to as much as 520 distinct attachments.

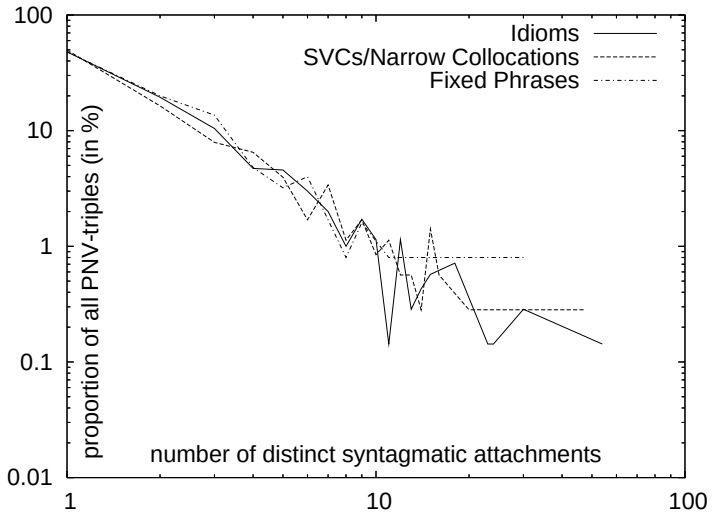

Figure 5.12: Distribution of syntagmatic attachments for the three collocation categories. The x- and y-axes are log-scaled.

It may be also illuminative to examine whether there is any difference in limited syntagmatic modifiability between the three categories of collocations examined in this thesis, i.e. between idioms, support verb constructions/narrow collocations, and fixed phrases. In particular, there appear to be common linguistic perceptions that, as Manning & Schütze (1999, p. 184) claim, limited syntagmatic modifiability (or as they call it: limited or non-modifiability) is especially true for frozen expressions like idioms. In order to investigate whether this is actually the case, we set up a distribution for the three categories of collocational PNV triples in which the distributional ranking criterion was again the number of distinct syntagmatic attachments. The results of this are visualized in figure 5.12.

As can be seen, the notion that idioms exhibit the strongest degree of limited syntagmatic modifiability is not supported by the above analysis. In fact, all three collocational categories have a proportion of almost 50% of their PNV triples which

only possess one distinct syntagmatic attachment.[5] The three curves for the collocational categories then decline in a similar degree although the idioms curve shows more fluctuation and even extends to 54 distinct supplements (which however may well have to do with the fact that idioms form the largest group of all collocations – see subsection 4.5.2.3). In any case, these results corroborate our linguistic assumptions that LSM is an association measure for collocation extraction that best distinguishes collocations from non-collocations but not between different subtypes of collocations.

5.2 Experimental Results for Term Extraction

In this section, we will present and examine the evaluation results for our performance experiments conducted for the task of term extraction from English-(sub)language text from the biomedical subdomain of Hematopoietic Stem Cell Transplantation and Immunology. For this purpose, we compared our linguistically motivated statistical association measure LPM to the standard statistical and information-theoretic association measures presented in section 3.3, *viz.* t-test, frequency, log-likelihood, C-value and PMI. Similar to our experiments for collocation extraction described in the previous section, we excluded Daille (1994)'s PMI variants as their results did not add any further insights. Because we conducted our experiments on bigram, trigram, and quadgram term candidates (see subsubsection 4.5.3.2 above), the log-likelihood association measure could only be utilized on the bigram data as it is not well-defined for larger-sized n-grams (see subsection 3.3.5 above). For both our quantitative and our qualitative results (see subsections 5.2.1 and 5.2.2), we present the performance metric data in form of tables and figures, in order to allow different views and perspectives on the results. Again, concerning the qualitative results, in particular with respect to the four qualitative criteria, due to clarity of presentation concerns, we refrain from visualizing them for all association measures and instead limit ourselves to illustrative samples.

5.2.1 Quantitative Results

The quantitative performance evaluation will be carried out on our bigram, trigram and quadgram term candidate sets, both for our large 100 million word and our

[5]Note that this is again the zero attachment.

small 10 million word biomedical MEDLINE corpus. Analogous to our experiments for collocation extraction (see subsection 5.1.1 above) and as laid out in subsubsection 4.5.1.4, this kind of evaluation is performed by incrementally examining increasing portions of the ranked output list returned by each of the five (for bigrams) / four (for tri- and quadgrams) association measures examined. Again, we evaluate their performance in terms of precision, recall, and ROC in a series of three experiments (subsubsection 5.2.1.1), but we refrain from presenting results for the F-score, as single precision and recall evaluations are equally telling. Finally, employing the McNemar test, we compare the ranked outputs of various association measures and test whether the differences are statistically significant (subsubsection 5.2.1.2).

5.2.1.1 Results on Performance Metrics

Analogous to our experiments for collocation extraction (see subsubsection 5.1.1.1 above), in the first series of quantitative experiments, we incrementally measured the performance of the various association scores in terms of their precision, for all term candidate n-gram sizes considered. For bigrams, the results are visualized in figure 5.13 for the large MEDLINE corpus and in figure 5.14 for the small one. The corresponding scores are given in table 5.11 at incremental intervals of 10 percentage points of the ranked output list. Considering precision scores is only informative at the upper portions[6] of the ranked output list because of their incremental convergence towards the baseline for all association measures considered.

As can be seen from both views on the results, the linguistically motivated LPM association measure holds a constant advantage over the next placed measures which all tend to cluster around the same curve areas and scores. Both for the large corpus and the small one, the difference between LPM and the second-placed t-test starts out with 20 points at one percent of the output list considered (i.e. after $\approx$ 667 and 190 candidates, respectively) and goes down to 3 to 5 points after 30% of the list are considered. It is actually interesting to note that, unlike in the case of collocation extraction, it is the t-test which runs slightly better than the frequency measure. Frequency, log-likelihood and C-value perform almost identically.

Another difference may be noticed with respect to the information-theoretic PMI

[6] In our case up to 30%, i.e. $\approx$ 20,000 candidates on the large corpus and $\approx$ 5,700 candidates on the small one.

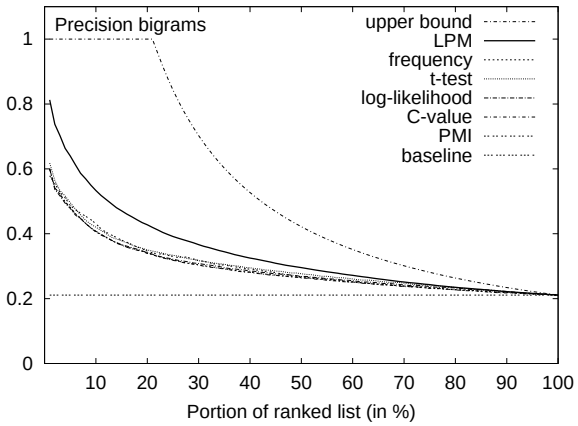

Figure 5.13: Bigram term precision on 100 million word corpus

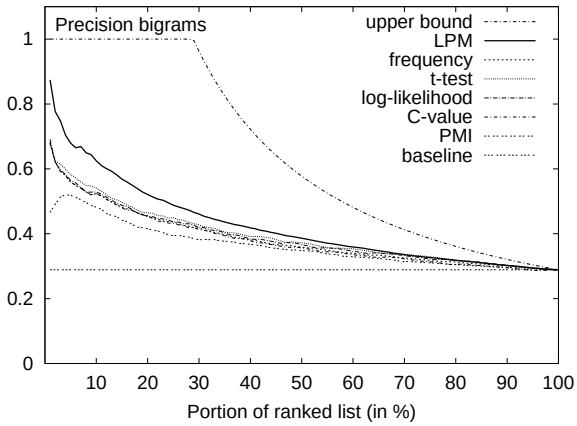

Figure 5.14: Bigram term precision on 10 million word corpus

Ranked list portion	Precision scores (66,669 NP bigram term candidates; 14,054 targets)							
	upper bound	LPM	frequency	t-test	log likelihood	C-value	PMI	baseline
1%	1.00	0.81	0.60	0.62	0.60	0.60	0.58	0.21
10%	1.00	0.53	0.41	0.42	0.41	0.41	0.43	0.21
20%	1.00	0.43	0.34	0.35	0.34	0.34	0.35	0.21
30%	0.68	0.36	0.30	0.31	0.31	0.30	0.32	0.21

Ranked list portion	Precision scores (19,001 NP bigram term candidates; 5,478 targets)							
	upper bound	LPM	frequency	t-test	log likelihood	C-value	PMI	baseline
1%	1.00	0.87	0.68	0.68	0.69	0.68	0.47	0.29
10%	1.00	0.62	0.52	0.54	0.53	0.52	0.48	0.29
20%	1.00	0.52	0.46	0.46	0.45	0.46	0.42	0.29
30%	0.93	0.45	0.41	0.42	0.42	0.41	0.38	0.29

Table 5.11: Bigram precision scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

measure. Whereas PMI exhibited an almost erratic behavior on the collocation extraction task, it runs in line with the other association measures – at least on the large corpus, it exhibits a noticeably weaker performance than its competitors on the small corpus.

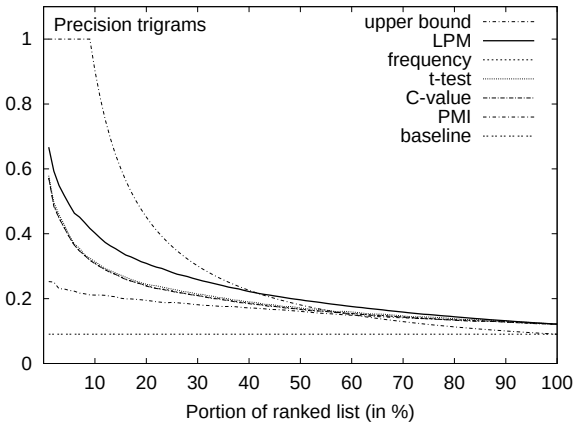

Figure 5.15: Trigram precision on 100 million word corpus

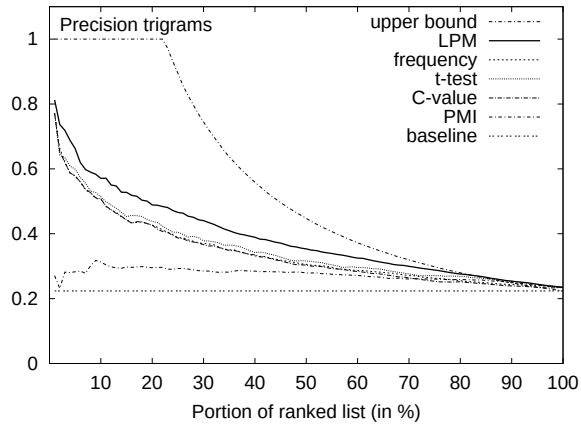

Figure 5.16: Trigram precision 10 million word corpus

The superiority of the linguistically motivated LPM measure can also be witnessed with respect to the performance in precision for trigram term candidates, as shown in figures 5.15 and 5.16 as well as in precision scores in table 5.12.

Ranked list portion	Precision scores (28,499 NP trigram term candidates; 3,459 targets)						
	upper bound	LPM	frequency	t-test	C-value	PMI	baseline
1%	1.00	0.67	0.57	0.58	0.57	0.25	0.12
10%	1.00	0.40	0.31	0.31	0.31	0.21	0.12
20%	0.61	0.31	0.24	0.24	0.24	0.19	0.12
30%	0.39	0.25	0.21	0.21	0.21	0.18	0.12

Ranked list portion	Precision scores (4,721 NP trigram term candidates; 1,108 targets)						
	upper bound	LPM	frequency	t-test	C-value	PMI	baseline
1%	1.00	0.81	0.77	0.77	0.77	0.27	0.23
10%	1.00	0.57	0.51	0.51	0.51	0.31	0.23
20%	1.00	0.49	0.43	0.44	0.43	0.30	0.23
30%	0.76	0.43	0.37	0.38	0.36	0.28	0.23

Table 5.12: Trigram precision scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

LPM's advantage with respect to the second-placed t-test starts out with 9 points on the large corpus but only with 4 points on the small one. Whereas this advantage reduces to 4 points on the large corpus at 30% of the output list considered, it even slightly increases on the small one. Another observation to be made is that PMI looses its ability to keep up with the other measures. Both on the large and the small corpus, it runs substantially lower and closer to the baseline, a pattern which has been similar in the case of our experiments for the collocation extraction task.

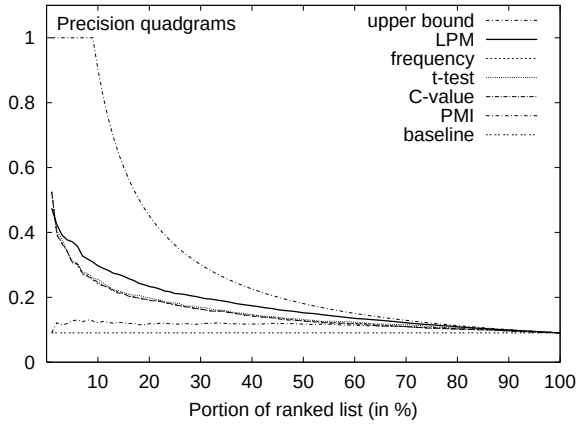

Figure 5.17: Quadgram term precision on 100 million words

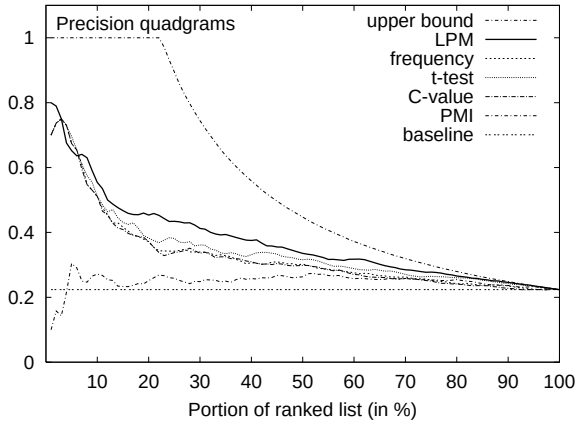

Figure 5.18: Quadgram term precision on 10 million words

The largest n-gram size considered in this study (and relevant for term extraction at all) are quadgrams, whose precision performance is visualized in figures 5.17 and 5.18 along with the corresponding scores up to the 30% portion on ranked output list in table 5.13. Considering 1% of the quadgram list on the large corpus (after ≈ 99 candidates), t-test, C-value and log-likelihood actually lead by 6 points over LPM. At this moment at least, we have no explanation why LPM performs worse than its standard competitors at this portion of the output list. Still, after 10% of the list have been considered, the result pattern has been reversed: it is now LPM which leads by 5 points over t-test.

Concerning the precision performance on the small corpus, it is again LPM which consistently runs above the other standard term extraction measures. The information-theoretic PMI measure, on the other hand, appears to have reached a similarly low performance as for collocation extraction, almost running parallel to the baseline. In between LPM and PMI, t-test, frequency and C-value obtain very

Ranked list portion	Precision scores (9,859 NP quadgram term candidates; 890 targets)						
	upper bound	LPM	fre-quency	t-test	C-value	PMI	base-line
1%	1.00	0.47	0.53	0.53	0.53	0.09	0.09
10%	0.90	0.30	0.24	0.25	0.25	0.12	0.09
20%	0.45	0.23	0.19	0.20	0.19	0.12	0.09
30%	0.29	0.20	0.16	0.17	0.16	0.12	0.09

Ranked list portion	Precision scores (912 NP quadgram term candidates; 204 targets)						
	upper bound	LPM	fre-quency	t-test	C-value	PMI	base-line
1%	1.00	0.80	0.70	0.70	0.70	0.10	0.22
10%	1.00	0.55	0.51	0.51	0.51	0.27	0.22
20%	1.00	0.45	0.37	0.38	0.37	0.25	0.22
30%	0.72	0.41	0.34	0.36	0.34	0.25	0.22

Table 5.13: Quadgram precision scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

similar precision scores. As can be seen from the results for all n-gram sizes and analogous to collocation extraction, compared to the ideal optimum, there is still room for improvement.

In the second series of quantitative experiments, we incrementally measured the performance of the various term extraction measures in terms of their recall, again both for our large and for our small corpus. Because recall measures the proportion of selected targets at a certain point in the ranked output list, it has a particularly practical relevance because the output lists produced by various association measures are typically post-examined by a human. For this reason, it is the in middle portions of the ranked output list where recall is most telling because, first, the earlier a large proportion of targets is returned the more efficient an association method may be considered, and second, toward the end of a ranked output list all association measures naturally approach the recall of 1.0 anyway.

For bigram term extraction, the recall results of the large 100 million and the small 10 million word corpus are visualized in figures 5.19 and 5.20, respectively, while the corresponding scores are given in table 5.14, again at incremental intervals of 10 percentage points of the ranked output list.

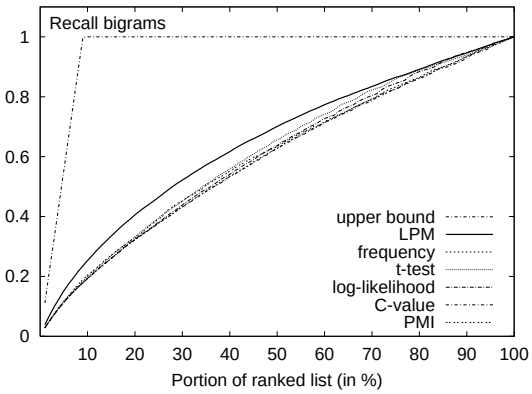

Figure 5.19: Bigram term recall on 100 million words

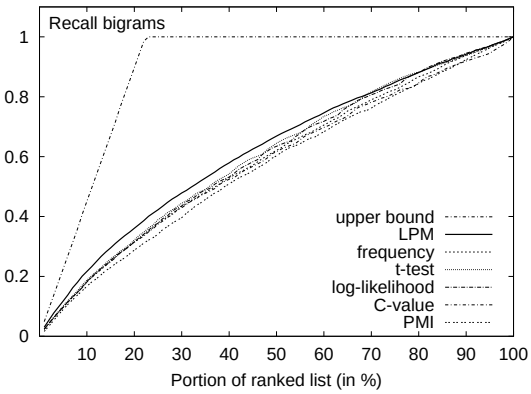

Figure 5.20: Bigram term recall on 10 million words

Ranked list portion	Recall scores (66,669 NP bigram term candidates; 14,054 targets)						
	upper bound	LPM	frequency	t-test	log likelihood	C-value	PMI
30%	1.00	0.53	0.44	0.46	0.45	0.44	0.46
40%	1.00	0.62	0.53	0.56	0.54	0.53	0.55
50%	1.00	0.70	0.63	0.66	0.64	0.63	0.63
60%	1.00	0.77	0.72	0.74	0.73	0.72	0.71
70%	1.00	0.83	0.79	0.82	0.80	0.79	0.79

Ranked list portion	Recall scores (19,001 NP bigram term candidates; 5,478 targets)						
	upper bound	LPM	frequency	t-test	log likelihood	C-value	PMI
30%	1.00	0.49	0.44	0.45	0.45	0.44	0.41
40%	1.00	0.58	0.53	0.54	0.53	0.52	0.51
50%	1.00	0.67	0.62	0.64	0.63	0.62	0.60
60%	1.00	0.75	0.71	0.73	0.72	0.70	0.68
70%	1.00	0.81	0.79	0.82	0.81	0.78	0.76

Table 5.14: Bigram recall scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

As can be seen from the plots and the corresponding scores, LPM maintains a consistent advantage in recall compared to the other measures – both on the large and the small corpus although the lead is bigger on the former one. After 40% of the

output list considered on the large corpus, while LPM already scores a recall of 0.62, the second-placed t-test runs substantially below at 0.56. On both corpora, the recall curves start to converge after 50% of the list have been examined.

As can be seen from the recall scores in table 5.14, t-test noticeably fares better than frequency, a tendency which already has been observed with respect to the precision performance and stays in contrast to results obtained for collocation extraction. Similarly, while PMI breaks even with C-value and log-likelihood on the large corpus, its starts running below these on the small one.

With respect to the recall results for trigrams, as presented in figures 5.21 and 5.22 as well as in table 5.15, it can be seen that LPM even fares better compared to its competitors than was the case for bigrams. On the large corpus, it leads the second-placed t-test by approximately 10 points up to the 70% portion. This has the effect that, in order to obtain a 0.8 recall, LPM only has to consider 49% of the ranked list whereas the second-placed t-test already has to scan up to the 64% portion. In a similar vein, in order to obtain a 0.9 recall, LPM only has to look at 66% of the ranked output list whereas t-test has scan up to the 78% portion and both frequency and C-value even up to the 84% portion.

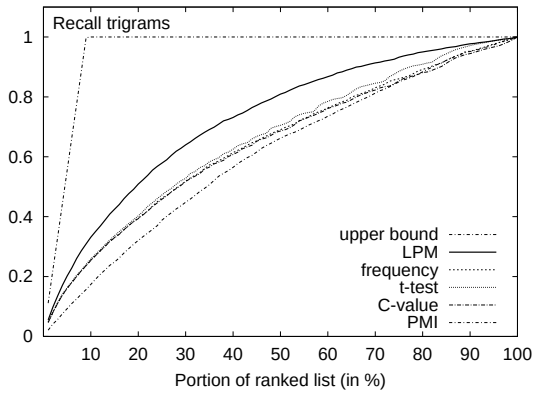

Figure 5.21: Trigram term recall on 100 million words

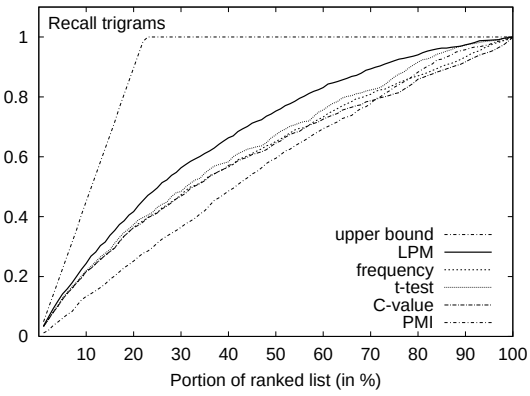

Figure 5.22: Trigram term recall on 10 million words

On the small corpus, as shown in the lower table 5.15, LPM's advantage to t-test in recall scores runs about 7 to 8 points up to the 70% portion of the output list. This still has the effect that, in order to obtain a 0.8 recall, LPM only has to scan 56% of the list while t-test has to consider 65% and both frequency and C-value even

Ranked list portion	Recall scores (28,499 NP trigram term candidates; 3,459 targets)					
	upper bound	LPM	fre-quency	t-test	C-value	PMI
30%	1.00	0.65	0.53	0.54	0.53	0.46
40%	1.00	0.73	0.61	0.63	0.61	0.56
50%	1.00	0.81	0.69	0.71	0.69	0.66
60%	1.00	0.87	0.76	0.78	0.76	0.74
70%	1.00	0.91	0.83	0.85	0.82	0.81

Ranked list portion	Recall scores (4,721 NP trigram term candidates; 1,108 targets)					
	upper bound	LPM	fre-quency	t-test	C-value	PMI
30%	1.00	0.57	0.49	0.50	0.48	0.38
40%	1.00	0.66	0.57	0.58	0.57	0.49
50%	1.00	0.75	0.65	0.67	0.65	0.59
60%	1.00	0.83	0.73	0.76	0.73	0.69
70%	1.00	0.89	0.81	0.82	0.79	0.78

Table 5.15: Trigram recall scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

73%. Similarly, LPM obtains the 0.9 recall threshold already at the 71% portion while t-test needs to go up to the 79% portion; C-value and frequency even need to crawl up to the 87% portion. In addition, on the large corpus and, in particular, on the small corpus, it can be seen that the information-theoretic PMI measure substantially underperforms the other measures in terms of its recall capacity.

Concerning the recall performance on quadgrams, as given in figures 5.23 and 5.24 as well as in table 5.16, it may be seen that LPM also substantially outperforms the other association measures in the critical middle portions of the ranked output list. From the 30% to the 60% portion on the large corpus, LPM leads t-test in the range of 8 to 12 points. This means that in order to obtain a 0.7 recall, LPM only needs to scan 33% of the ranked list whereas t-test needs to go as far as 48%. Likewise, in order to reach 0.8 recall, LPM needs to scan 14 percentage points less of ranked output compared to t-test, *viz.* 44% compared to 58%.

Taking the same portion range (30% to 60%) on the small corpus, LPM's lead compared to t-test runs between 8 points (at the 40% and at the 60% portion) and 4

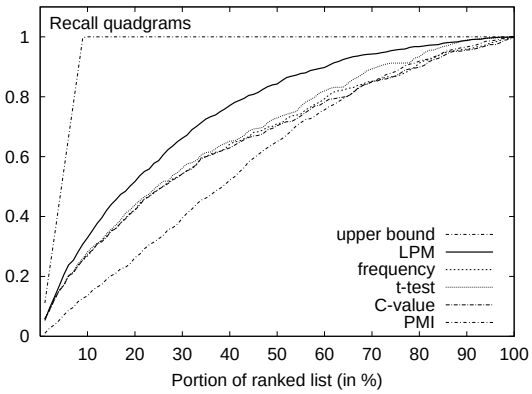

Figure 5.23: Quadgram term recall on 100 million words

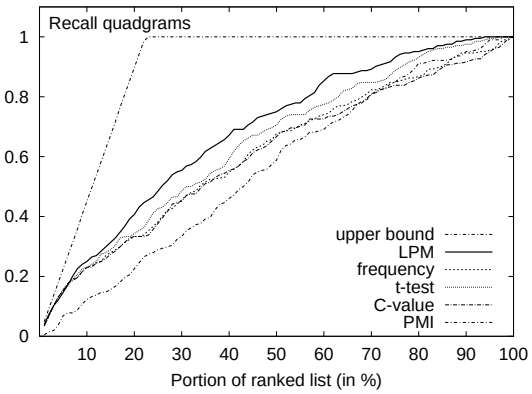

Figure 5.24: Quadgram term recall on 10 million words

Ranked list portion	Recall scores (9,859 NP quadgram term candidates; 890 targets)					
	upper bound	LPM	frequency	t-test	C-value	PMI
30%	1.00	0.67	0.55	0.58	0.55	0.41
40%	1.00	0.77	0.64	0.65	0.63	0.52
50%	1.00	0.84	0.71	0.73	0.70	0.65
60%	1.00	0.90	0.79	0.82	0.78	0.76
70%	1.00	0.94	0.85	0.89	0.85	0.85

Ranked list portion	Recall scores (912 NP quadgram term candidates; 204 targets)					
	upper bound	LPM	frequency	t-test	C-value	PMI
30%	1.00	0.57	0.47	0.50	0.47	0.35
40%	1.00	0.67	0.55	0.59	0.55	0.46
50%	1.00	0.75	0.68	0.71	0.67	0.59
60%	1.00	0.85	0.74	0.77	0.73	0.69
70%	1.00	0.89	0.82	0.85	0.81	0.81

Table 5.16: Quadgram recall scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

points (at the 50% portion). For obtaining a 0.7 recall, LPM needs to consider 44% of the output while t-test needs to go further up to 50%. Likewise LPM reaches the 0.8 recall threshold after 56% of the output and t-test does so after 63%. Still, as it was

the case for precision performance of quadgrams on the small corpus described above, the curves indicate some fluctuations which appear to due to the small candidate set. Again, both on the large and the small corpus, PMI substantially runs below the other measures.

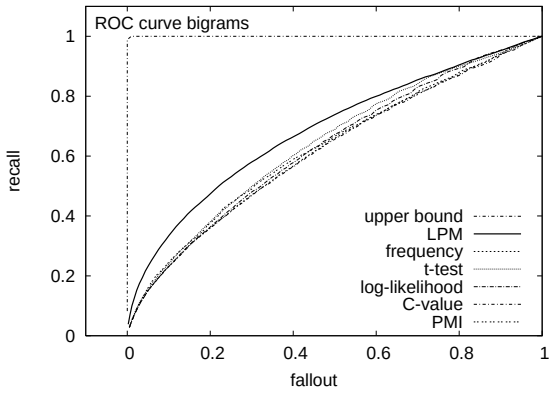

Figure 5.25: Bigram term ROC on 100 million words

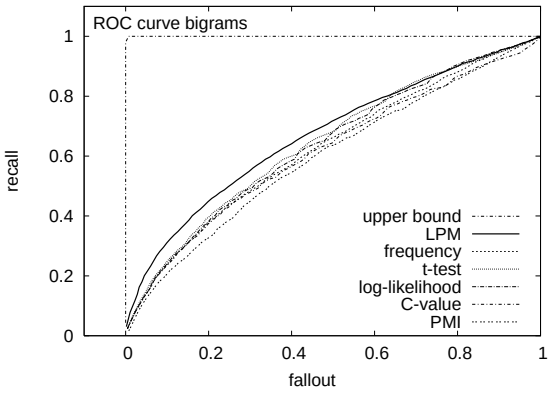

Figure 5.26: Bigram term ROC on 10 million words

When examining the fallout scores of the various association measures on bigrams, as given in table 5.17, it can be seen that there is no big difference in the fallout scores (i.e. the rate of non-targets selected) among the various association measures examined at increasing portions of the ranked output list. Hence, only looking at the mere fallout scores is not very informative and for this reason (see subsubsection 4.5.1.4 and also the results on collocation extraction in subsubsection 5.1.1.1), fallout is typically plotted against recall thus yielding the so-called receiver operation characteristic (ROC) curve, which is visualized in figures 5.25 and 5.26 for the large and the small corpus, respectively. As can be seen, the effect of the superior recall scores of LPM compared to the other association (as previously outlined in figures 5.19 and 5.20 as well as table 5.14) come to effect here, as LPM follows the left-hand border and then the top border of the ROC space more closely than its competitors.

This effect is still much more pronounced when looking at the the ROC curves for the trigram extraction task. As can be seen in figures 5.27 and 5.28, the LPM ROC curve follows the left-hand border and then the top border of the ROC space much more closely than its competitors – exhibiting a substantial area difference to the second-placed t-test. In a similar vein, the ROC curves for the quadgram extraction task (see figures 5.29 and 5.28) show a comparable pattern, although the fluctuations

Ranked list portion	Fallout scores (66,669 NP bigram term candidates; 14,054 targets)						
	upper bound	LPM	fre-quency	t-test	log likelihood	C-value	PMI
30%	0.13	0.25	0.27	0.27	0.27	0.27	0.27
40%	0.24	0.34	0.36	0.36	0.36	0.36	0.36
50%	0.37	0.45	0.47	0.46	0.46	0.47	0.46
60%	0.49	0.55	0.57	0.56	0.57	0.57	0.57
70%	0.62	0.66	0.68	0.67	0.67	0.68	0.68

Ranked list portion	Fallout scores (19,001 NP bigram term candidates; 5,478 targets)						
	upper bound	LPM	fre-quency	t-test	log likelihood	C-value	PMI
30%	0.03	0.24	0.26	0.25	0.25	0.26	0.27
40%	0.16	0.33	0.35	0.34	0.35	0.35	0.36
50%	0.30	0.43	0.45	0.44	0.45	0.45	0.46
60%	0.44	0.54	0.56	0.55	0.55	0.56	0.57
70%	0.58	0.65	0.66	0.65	0.66	0.67	0.67

Table 5.17: Bigram fallout scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

on the small corpus show the same behavior as in the other quantitative performance results above.

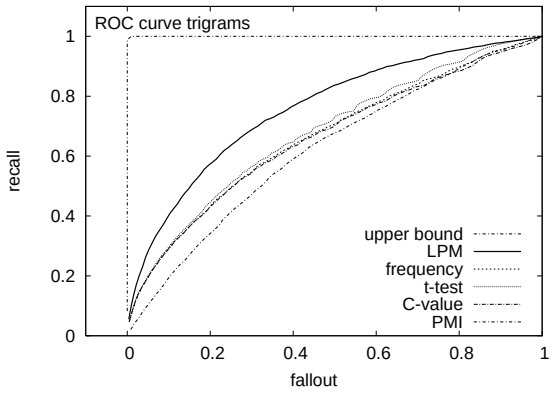

Figure 5.27: Trigram term ROC on 100 million words

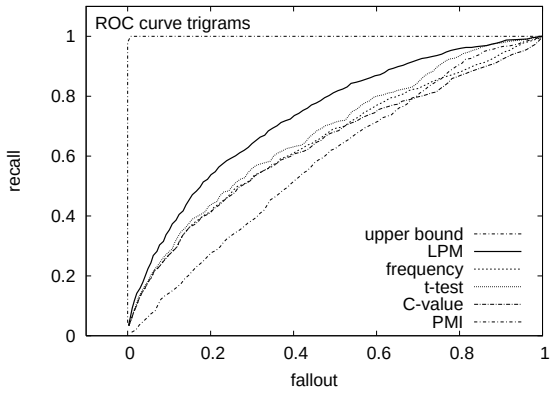

Figure 5.28: Trigram term ROC on 10 million words

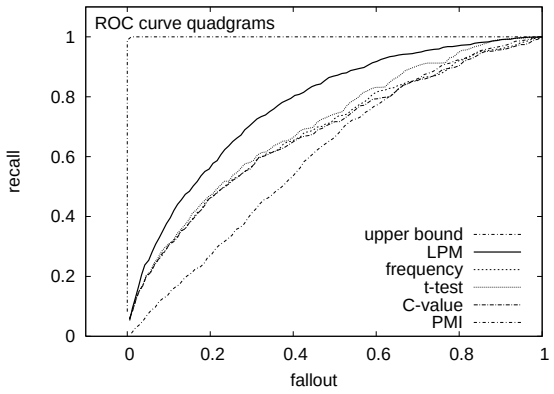

Figure 5.29: Quadgram term ROC on 100 million words

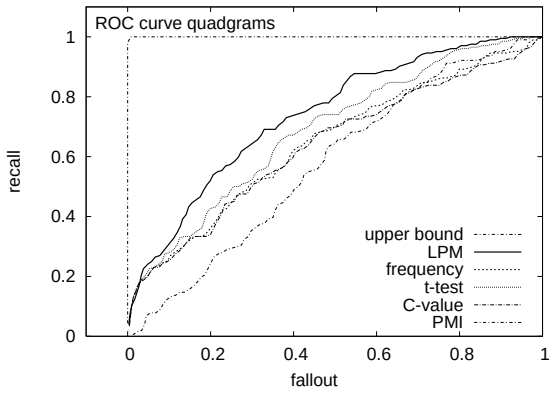

Figure 5.30: Quadgram term ROC on 10 million words

Ranked list portion	Fallout scores (28,499 NP trigram term candidates; 3,459 targets)					
	upper bound	LPM	frequency	t-test	C-value	PMI
30%	0.21	0.26	0.28	0.28	0.28	0.29
40%	0.32	0.35	0.37	0.37	0.37	0.38
50%	0.43	0.46	0.47	0.47	0.47	0.48
60%	0.54	0.56	0.58	0.57	0.58	0.58
70%	0.66	0.67	0.68	0.68	0.68	0.68

Ranked list portion	Fallout scores (4,721 NP trigram term candidates; 1,108 targets)					
	upper bound	LPM	frequency	t-test	C-value	PMI
30%	0.10	0.23	0.26	0.25	0.26	0.29
40%	0.22	0.32	0.35	0.34	0.35	0.37
50%	0.35	0.42	0.45	0.45	0.46	0.47
60%	0.48	0.53	0.56	0.55	0.56	0.57
70%	0.61	0.64	0.67	0.66	0.67	0.68

Table 5.18: Trigram fallout scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

Inspecting the corresponding fallout scores for trigrams and quadgrams in tables 5.18 and 5.19 again shows that their impact on the ROC curves is less than that of the recall scores given above. On the large corpus, although LPM reveals a consistently

Ranked list portion	Fallout scores (9,859 NP quadgram term candidates; 890 targets)					
	upper bound	LPM	fre-quency	t-test	C-value	PMI
30%	0.24	0.27	0.29	0.28	0.29	0.30
40%	0.34	0.36	0.38	0.38	0.38	0.39
50%	0.45	0.47	0.48	0.48	0.48	0.49
60%	0.56	0.57	0.58	0.58	0.58	0.58
70%	0.67	0.68	0.69	0.68	0.69	0.69

Ranked list portion	Fallout scores (912 NP quadgram term candidates; 204 targets)					
	upper bound	LPM	fre-quency	t-test	C-value	PMI
30%	0.11	0.24	0.26	0.26	0.26	0.30
40%	0.23	0.32	0.36	0.34	0.36	0.38
50%	0.36	0.43	0.45	0.44	0.45	0.47
60%	0.49	0.53	0.56	0.55	0.56	0.57
70%	0.61	0.65	0.67	0.66	0.67	0.67

Table 5.19: Quadgram fallout scores of association measures for term extraction on the 100 million word (upper table) and 10 million word (lower table) MEDLINE corpus.

lower fallout rate than its competitors, the difference to the second-placed t-test runs around 2 points. As was shown for the recall scores in tables 5.15 and 5.16 above, LPM's lead amounts to up to 10 points for the corresponding portions of the ranked output list. On the small corpus, both the trigram and quadgram fallout scores show a slightly bigger difference of 3 points, but still, it is the corresponding recall scores which cause the LPM ROC curves to cover more of the available ROC space.

5.2.1.2 Results on Significance Testing

Although the observed differences for the results on the quantitative performance evaluation (see the last subsubsection 5.2.1.1) are substantial with respect to LPM's superiority, and thus there is reason to believe that they are not merely due to chance (see subsection 4.5.1.5), we will corroborate these findings by applying McNemar as a significance test of differences to incremental portion measure points of the ranked output list, both on the large and on the small corpus for all n-gram sizes. Analogously to our experiments for collocation extraction (see subsubsection 5.1.1.2), we selected

100 measure points in the ranked list, one after each increment of one percent, and then used the two-tailed test for a (very strict) confidence interval of 99%. In particular, we tested the significance of differences between LPM and the two next best performing association measures, t-test and frequency.

For bigrams, the results at 10-point increments are given in table 5.20 both for the large and the small corpus.

# of measure points considered	Large corpus: # of significant differences comparing LPM with		Small Corpus: # of significant differences comparing LPM with	
	frequency	t-test	frequency	t-test
10	10	10	10	10
20	20	20	20	20
30	30	30	30	30
40	40	40	40	40
50	50	50	50	50
60	60	60	60	60
70	70	70	70	67
80	80	76	80	70
90	90	85	85	75
100	92	94	85	80

Table 5.20: Bigram term extraction: significance testing of differences using the two-tailed McNemar test at 99% confidence interval on the large and the small MEDLINE corpus

As can be seen from the number of significant differences, the clear advantage that LPM exhibited in comparison to its competitors on all the quantitative performance metrics translates right into the results for the McNemar test. Up to measure point 70 (on the large corpus) and measure point 60 (on the small corpus), all differences are significant, and only then some differences to the second-placed t-test turn out to be not significant any more. As can be seen from the higher number of significant differences to LPM at later measure points, the weaker performance of frequency compared to t-test on the performance evaluations is also corroborated here.

As for the McNemar results on the trigram extraction task (see table 5.21), up to measure point 90 all the differences but one up are significant for LPM on the large corpus, both with respect to t-test and frequency. Curiously, on the small corpus the first 10 measure points show (almost) no significant differences, which

# of measure points considered	Large corpus: # of significant differences comparing LPM with		Small Corpus: # of significant differences comparing LPM with	
	frequency	t-test	frequency	t-test
10	9	9	1	0
20	19	19	11	10
30	29	29	21	20
40	39	39	31	30
50	49	49	41	40
60	59	59	51	50
70	69	69	61	60
80	79	79	71	70
90	89	89	81	75
100	97	89	87	75

Table 5.21: Trigram term extraction: significance testing of differences using the two-tailed McNemar test at 99% confidence interval on the large and the small MEDLINE corpus

seems to indicate that during the first 10% of the ranked output list the three measures considered do not rank their term candidates in a significantly different way – a fact that at least partially also seems to be reflected in the trigram performance results given in figure 5.16 and the lower part of table 5.12. Then, however, all the consecutive measures points up to 80 are significantly different, thus reflecting LPM's superiority in performance at the crucial portions of the ranked output list.

As for the McNemar results on the quadgram extraction task (see table 5.21), the first 10 measures points show 6 significant differences for LPM on the large corpus. This corresponds most visibly to the precision performance results shown in figure 5.17 and the upper table 5.13 in which it may be seen that LPM starts out weakly in comparison to t-test and frequency. Then, however, all measure points up to 80 show significant differences and thus corroborate the performance advantage of LPM on the decisive middle portions of the ranked output list.

The picture looks rather different on the small corpus. Comparing LPM with t-test, all together only 13 of the 100 measure points are significantly different. This means that the visible performance advantages of LPM on quadgrams shown in the previous subsubsection are not borne out by statistical significance testing. This is in line with the already made observations that due to the substantially smaller size of

# of measure points considered	Large corpus: # of significant differences comparing LPM with		Small Corpus: # of significant differences comparing LPM with	
	frequency	t-test	frequency	t-test
10	6	6	0	0
20	16	16	0	0
30	26	26	10	2
40	36	36	20	9
50	46	46	29	9
60	56	56	37	10
70	66	66	47	13
80	76	76	57	13
90	86	78	67	13
100	93	78	73	13

Table 5.22: Quadgram term extraction: significance testing of differences using the two-tailed McNemar test at 99% confidence interval on the large and the small MEDLINE corpus

the quadgram candidate set for the small corpus (i.e. only 912 term candidates) the results are rather brittle and difficult to interpret.

5.2.2 Qualitative Results

In section 4.5.1.6, we have formulated four achievement objectives (i.e. two static ones and two dynamic ones) for qualitative performance evaluation of lexical association measures and took frequency of co-occurrence as as a sort of baseline against which a particular association measure should re-rank (or not) the targets and non-targets of the candidate set. These four objectives are divided into two static criteria (subsubsection 5.2.2.1) and two dynamic criteria (subsubsection 5.2.2.2). Similar to the results reported for collocation extraction in subsection 5.1.2, we choose the middle rank as a mark to divide a ranked output list into an upper portion and a lower portion and partitioned these into three further subportions each. Then, again, the targets and non-targets assigned to these portions by frequency will be examined and quantified, according to the four criteria, to what degree the other association measures changed these rankings or not.

5.2.2.1 Results on the static criteria

The first two criteria examine how static an association measure is in that a qualitatively superior measure should at least keep the status quo with respect to frequency. In this respect, criterion 1 examines whether an association measure is able to keep the targets (i.e. the true terms) in the upper portion, whereas criterion 2 checks to which degree a measure is able to keep the non-targets in their lower portion place

	AM	upper portion (ranks 1 - 33,334)			lower portion (ranks 33,335 - 66,669)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 1	freq	3951 (44.8%)	2619 (29.7%)	2254 (25.5%)	0	0	0
	t-test	4079 (46.2%)	2768 (31.4%)	1272 (14.4%)	145 (1.6%)	86 (1.0%)	474 (5.4%)
8824	logL	3950 (44.8%)	2674 (30.3%)	1868 (21.2%)	158 (1.8%)	47 (0.5%)	127 (1.4%)
Ts	C-value	3940 (44.7%)	2611 (29.6%)	2177 (24.7%)	96 (1.1%)	0	0
	PMI	2346 (26.6%)	1755 (19.9%)	1407 (15.9%)	1257 (14.2%)	1155 (13.1%)	904 (10.2%)
	LPM	3967 (45.0%)	1981 (22.5%)	1188 (13.5%)	825 (9.3%)	548 (6.2%)	315 (3.6%)
Crit. 2	freq	0	0	0	9132 (32.5%)	9390 (33.5%)	9583 (34.0%)
	t-test	0	0	3537 (12.6%)	8254 (29.4%)	8970 (31.9%)	7344 (26.1%)
28105	logL	205 (0.7%)	201 (0.7%)	1367 (4.9%)	8182 (29.1%)	9093 (32.4%)	9057 (32.2%)
NTs	C-value	0	0	340 (1.2%)	8834 (31.4%)	9389 (33.4%)	9542 (34.0%)
	PMI	4365 (15.5%)	4605 (16.4%)	4686 (16.4%)	4372 (15.6%)	4637 (16.5%)	5440 (19.4%)
	LPM	1499 (5.3%)	2906 (10.3%)	3862 (13.7%)	5201 (18.5%)	6501 (23.1%)	8136 (29.0%)

Table 5.23: Results on the two static qualitative criteria for bigram term extraction on the large MEDLINE corpus.

Table 5.23 shows the results for the two criteria for bigram extraction on the large MEDLINE corpus. As can be seen for criteria 1, t-test, log-likelihood, and, in particular, C-value only demote few of their targets to the lower portion, although t-test places 5.4% of its targets to the lowest subportion. LPM, on the other hand, places a bigger chunk of its targets (9.3%, 6.2% and 3.5%) to the lower three subportions. In a pattern reminiscent for the collocation extraction task, PMI demotes the highest number of targets (37.5%) to the lower three subportions, compared to frequency.

Concerning the other static criterion 2, it is again the case that none of the association measures is able to achieve this goal completely whereby some measures fare better than others. Again, t-test, log-likelihood and C-value are best able to fulfill this criterion although t-test still places 12.6% of its non-targets into the third

upper subportion. Conversely to the first criterion above, PMI is least able to meet this criterion and puts the biggest proportion of its lower-portion non-targets out of place, with almost half (48.3%) promoted to the upper three subportions. Although LPM, by only keeping 70% of its non-targets in the lower portion, is less able to fulfill this criterion than the frequency-like behaving measures t-test, log-likelihood, and C-value, it still does so considerably better than PMI.

	AM	upper portion (ranks 1 - 9,500)			lower portion (ranks 9,501 - 19,001)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 1	freq	1500 (44.0%)	1050 (30.8%)	858 (25.2%)	0	0	0
	t-test	1534 (45.0%)	1096 (32.2%)	508 (14.9%)	89 (2.6%)	24 (0.7%)	157 (4.6%)
3408	logL	1509 (44.3%)	1077 (31.6%)	705 (20.7%)	69 (2.0%)	11 (0.3%)	37 (1.1%)
Ts	C-value	1523 (44.7%)	1038 (30.5%)	812 (23.8%)	35 (1.0%)	0	0
	PMI	723 (21.2%)	633 (18.6%)	592 (17.4%)	516 (15.1%)	486 (14.3%)	458 (13.4%)
	LPM	1344 (39.4%)	733 (21.5%)	530 (15.6%)	380 (11.2%)	281 (8.2%)	140 (4.1%)
Crit. 2	freq	0	0	0	2380 (32.1%)	2492 (33.6%)	2550 (34.0%)
	t-test	0	0	796 (10.7%)	2115 (29.3%)	2423 (32.6%)	2088 (28.1%)
7422	logL	12 (0.2%)	21 (0.3%)	368 (5.0%)	2048 (27.6%)	2426 (32.7%)	2547 (34.3%)
NTs	C-value	0	0	88 (1.2%)	1905 (25.7%)	1999 (26.9%)	3430 (42.2%)
	PMI	980 (13.2%)	1002 (13.5%)	969 (13.1%)	1242 (16.7%)	1236 (16.7%)	1993 (26.9%)
	LPM	492 (6.6%)	868 (11.7%)	1011 (13.6%)	1360 (18.3%)	1594 (21.5%)	2097 (28.3%)

Table 5.24: Results on the two static qualitative criteria for bigram term extraction on the small MEDLINE corpus.

Looking at the two static criteria for bigrams on the small MEDLINE corpus, table 5.24 shows a quite similar picture. Of the three frequency-like behaving measures C-value, log-likelihood and C-value, it is t-test which still demotes most of its targets to the lower portion whereas log-likelihood and, even more, C-value keep them in their respective upper portions. PMI again falls out of line by demoting 43% of its upper-portion targets to the lower three subportions while LPM is able to fulfill criterion 1 to a much higher degree. Conversely for criterion 2, although LPM is not able to keep as many of its non-targets in their lower-portion place as t-test, log-likelihood and C-value, this deplacement is not substantial compared to the one PMI causes to its lower-portion non-targets.

For the trigram term extraction task on the large corpus, table 5.25 shows that for criterion 1 t-test and C-value are hardly distinguishable from frequency. Also LPM

demotes fewer targets to the lower subportions than in the bigram case described above. Only PMI, demoting 36% of its targets to one of the lower three subportions, underperforms in a similar pattern already observed before. One notable result about LPM is that it is even to promote more targets into the first upper portion compared to frequency (55% vs. 51.2%, respectively).

	AM	upper portion (ranks 1 - 14,249)			lower portion (ranks 14,250 - 28,499)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 1	freq	1223 (51.2%)	1025 (42.9%)	140 (5.9%)	0	0	0
	t-test	1244 (52.1%)	1049 (43.9%)	74 (3.1%)	10 (0.4%)	2 (0.1%)	9 (0.4%)
2388	C-value	1214 (50.8%)	1013 (42.4%)	144 (6.0%)	17 (0.7%)	0	0
Ts	PMI	579 (24.2%)	838 (35.1%)	123 (5.2%)	315 (13.2%)	304 (12.7%)	229 (9.6%)
	LPM	1314 (55.0%)	702 (29.4%)	79 (3.3%)	166 (6.9%)	92 (3.8%)	35 (1.5%)
Crit. 2	freq	0	0	0	4340 (32.9%)	4401 (33.4%)	4438 (33.7%)
	t-test	0	0	581 (4.4%)	4000 (30.4%)	4329 (32.8%)	4269 (32.4%)
13179	C-value	0	0	115 (0.9%)	4235 (32.1%)	4435 (33.7%)	4394 (33.3%)
NTs	PMI	2349 (17.8%)	3622 (27.5%)	701 (5.3%)	2216 (16.8%)	2171 (16.5%)	2120 (16.1%)
	LPM	897 (6.8%)	2863 (21.7%)	679 (5.1%)	2348 (17.8%)	2808 (21.3%)	3584 (27.2%)

Table 5.25: Results on the two static qualitative criteria for trigram term extraction on the large MEDLINE corpus.

With respect to criterion 2, the lower part of table 5.25 exhibits that LPM places a high proportion of non-targets to the second upper portion (21.7%), compared to the same criterion for the bigram extraction task. That t-test and especially C-value exhibit similar characteristics to frequency is corroborated by the observation that no non-targets get promoted to the first two upper subportions.

On the small corpus, the results for the two static criteria fall into the same patterns as on the large corpus. For the first criterion given in the upper part of table 5.26, LPM again demotes less trigram targets to the lower three subportions that it was the case for the bigram task and again, the linguistically motivated association is even able to promote a higher proportion of targets into the top upper subportion. T-test and C-value hardly change the target rankings of frequency while PMI demotes 41% of its upper-portion targets to the lower portion.

Like it was the case for the trigram term extraction on the large corpus, the result patterns for the second static criterion on the small corpus (in the lower part of table

	AM	upper portion (ranks ranks 1 - 2,360)			lower portion (ranks 2,361 - 4,721)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 1	freq	343 (47.5%)	219 (30.3%)	160 (22.2%)	0	0	0
	t-test	361 (50.0%)	221 (30.6%)	124 (17.2%)	10 (1.4%)	1 (0.1%)	5 (0.7%)
722	C-value	340 (47.1%)	219 (30.3%)	152 (21.1%)	9 (1.2%)	0	2 (0.2%)
Ts	PMI	125 (17.3%)	123 (17.0%)	140 (19.4%)	122 (18.9%)	142 (19.7%)	70 (9.7%)
	LPM	342 (47.4%)	159 (22.0%)	92 (12.7%)	68 (9.4%)	51 (7.1%)	10 (1.4%)
Crit. 2	freq	0	0	0	633 (32.1%)	676 (34.2%)	668 (33.8%)
	t-test	0	0	93 (4.7%)	580 (29.4%)	637 (32.3%)	665 (33.6%)
1975	C-value	0	0	36 (1.8%)	621 (31.4%)	658 (33.3%)	660 (33.4%)
NTs	PMI	382 (19.3%)	339 (17.2%)	332 (16.8%)	318 (16.1%)	268 (13.6%)	336 (17.0%)
	LPM	143 (7.2%)	253 (12.8%)	300 (15.2%)	361 (18.3%)	388 (19.6%)	530 (26.8%)

Table 5.26: Results on the two static qualitative criteria for trigram term extraction on the small MEDLINE corpus.

5.26) show that t-test and C-value fulfill this criterion well in that they do not promote any non-targets into the upper two subportions and only very few into the third upper subportion. LPM, on the other hand, places a substantial number of non-targets into the upper three subportions (39.2% all together). This, however, is again topped by PMI which even promotes 53.3% of its lower-portion non-targets into the upper portions.

	AM	upper portion (ranks 1 - 4,929)			lower portion (ranks 4,930 - 9,859)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 1	freq	331 (52.4%)	186 (29.4%)	115 (18.2%)	0	0	0
	t-test	341 (53.9%)	188 (29.7%)	100 (15.8%)	3 (0.5%)	0	0
632	C-value	331 (52.4%)	186 (29.4%)	109 (17.2%)	5 (0.8%)	1 (0.2%)	0
Ts	PMI	115 (18.2%)	129 (20.4%)	127 (20.1%)	127 (20.1%)	85 (13.4%)	49 (7.8%)
	LPM	343 (54.3%)	146 (23.1%)	72 (11.4%)	51 (8.1%)	13 (2.0%)	7 (1.1%)
Crit. 2	freq	0	0	0	1583 (33.9%)	1558 (33.3%)	1531 (32.8%)
	t-test	0	0	183 (3.9%)	1401 (30.0%)	1539 (32.9%)	1549 (33.2%)
4672	C-value	0	0	58 (1.2%)	1487 (31.8%)	1551 (33.2%)	1576 (33.7%)
NTs	PMI	931 (19.9%)	817 (17.5%)	752 (16.1%)	737 (15.8%)	717 (15.3%)	718 (15.4%)
	LPM	449 (9.6%)	632 (13.5%)	747 (16.0%)	776 (16.6%)	916 (19.6%)	1153 (24.7%)

Table 5.27: Results on the two static qualitative criteria for quadgram term extraction on the large MEDLINE corpus.

Finally, the results for the two static criteria with respect to quadgram term extraction on the large corpus are given in table 5.27. Concerning criterion 1, both t-test and C-value demote very few of their targets to the lower portion (0.5% and 1%, respectively) and also LPM places a smaller proportion of its upper-portion targets into the lower three subportion (11.1%) than it was the case for the bi- and trigram extraction task. PMI, on the hand, remains faithful to its pattern of demoting a large chunk (41.3%) to the lower portions. LPM and, to a lesser extent, t-test also augment the proportion of targets in their top upper subportion by 1.9% and 1.5%, respectively, compared to the frequency proportion of 52.4%.

The results for the second criterion are given in the lower part of table 5.27. Also here, similar qualitative result patterns surface compared to the other two n-gram term extraction tables: Whereas t-test and C-value meet the criterion to a very large extent by only promoting a very small margin of its lower-portion non-targets to the third upper subportion, PMI violates the criterion in an enormous way since it moves 53.6% of them upwards. LPM again lies in between these two extremes.

	AM	upper portion (ranks 1 - 456)			lower portion (ranks 457 - 912)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 1	freq	61 (44.2%)	41 (29.7%)	36 (26.1%)	0	0	0
	t-test	65 (47.1%)	39 (28.3%)	34 (24.6%)	0	0	0
138	C-value	60 (43.5%)	39 (28.3%)	37 (26.8%)	2 (1.4%)	0	0
Ts	PMI	19 (13.8%)	26 (18.8%)	31 (22.5%)	29 (21.0%)	22 (15.9%)	11 (8.0%)
	LPM	58 (42.0%)	40 (29.0%)	17 (12.3%)	13 (9.4%)	10 (7.2%)	0
Crit. 2	freq	0	0	0	129 (33.0%)	131 (33.7%)	130 (33.3%)
	t-test	0	0	23 (5.9%)	117 (30.0%)	122 (31.3%)	128 (32.8%)
390	C-value	0	0	6 (1.5%)	127 (32.6%)	128 (32.8%)	129 (33.1%)
NTs	PMI	87 (22.3%)	69 (17.7%)	57 (14.6%)	55 (14.1%)	59 (15.1%)	63 (16.2%)
	LPM	37 (9.5%)	50 (12.8%)	66 (16.9%)	64 (16.4%)	77 (19.7%)	96 (24.6%)

Table 5.28: Results on the two static qualitative criteria for quadgram term extraction on the small MEDLINE corpus

Finally, the results for the two static criteria on the small corpus, given in table 5.28, confirm all the already observed result patterns. Despite this fact, these particular results have to interpreted with care. Because the quantitative results described in previous two subsections had shown that, due to the relatively small size of the

candidate set,[7] these results are more brittle and thus less reliable than those for the n-gram term extraction cases for which a larger amount of candidate set data is available.

5.2.2.2 Results on the dynamic criteria

The third and fourth criteria examine how dynamic an association measure is in that a qualitatively superior measure should change and improve the rankings with respect to frequency. In this respect, criterion 3 examines whether an association measure is able to demote the non-targets (i.e. the non-terms) from the upper to the lower three subportions while criterion 2 determines to which degree a measure is able to promote targets (i.e. actual terms) from the lower to the upper subportions.

Table 5.29 gives the results for the bigram term extraction on the large MEDLINE corpus. As can be seen with respect to criterion 3, the measure which least changes the frequency ranking is C-value, which only demotes 1.3% of the upper-portion non-targets into the first lower subportion. Log-likelihood demotes a slightly higher proportion of non-targets (7.8%); t-test doubles this proportion by already demoting 16% of the non-targets, in particular it demotes 10% of them to the third lower subportion. LPM, however, by far performs much better in meeting this criterion than t-test as it demotes 38% of the upper-portion non-targets to the lower three subportions. It is PMI again which, like in the case of the two static criteria described above, appears to shuffle things around and places 56% of the non-targets into the lower three subportions.

The results for the other dynamic criterion 4, the ability to promote lower-portion targets to the upper subportions, are given in the lower part of table 5.29. As can be seen, also here it is the C-value measure which is the most conservative and retains the frequency ranking to an enormous extent as it is only able to place 1.2% of the targets to the third upper subportion. The t-test measure is at least able to promote 21.1% of the targets to the third upper subportion (but not any further upwards) whereas log-likelihood only manages to do so with 7.6%. LPM, on the other hand, is able to substantially meet this criterion by placing over half of the targets (51.4%) into the upper three subportions, with the largest chunk (20.8%) landing in the top

[7] It should be recalled that the small size of the candidates set is due to fixed frequency cut-off threshold $c \geq 5$ – (see subsection 4.5.3.2).

	AM	upper portion (ranks 1 - 33,334)			lower portion (ranks 33,335 - 66,669)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 3	freq	7160 (29.2%)	8492 (34.6%)	8858 (36.1%)	0	0	0
	t-test	7031 (28.7%)	8343 (34.0%)	5206 (21.2%)	949 (3.9%)	525 (2.1%)	2456 (10.0%)
24510	logL	6923 (28.2%)	8208 (33.5%)	7477 (30.5%)	923 (3.8%)	309 (1.3%)	670 (2.7%)
NTs	C-value	7164 (29.2%)	8496 (34.7%)	8533 (34.8%)	317 (1.3%)	0	0
	PMI	2635 (10.8%)	3808 (15.5%)	4304 (17.6%)	4866 (19.9%)	4783 (19.5%)	4114 (16.8%)
	LPM	4557 (18.6%)	5451 (22.2%)	5211 (21.3%)	4316 (17.6%)	3249 (13.3%)	1726 (7.0%)
Crit. 4	freq	0	0	0	1979 (37.8%)	1721 (32.9%)	1530 (29.3%)
	t-test	0	0	1098 (21.1%)	1763 (33.7%)	1530 (29.2%)	839 (16.0%)
5230	logL	26 (0.5%)	24 (0.4%)	399 (7.6%)	1846 (35.3%)	1658 (31.7%)	1277 (24.4%)
Ts	C-value	0	0	62 (1.2%)	1862 (35.6%)	1719 (32.9%)	1587 (30.3%)
	PMI	1755 (33.6%)	937 (17.9%)	713 (13.6%)	615 (11.8%)	533 (10.2%)	677 (12.9%)
	LPM	1088 (20.8%)	774 (14.8%)	850 (16.3%)	770 (14.7%)	813 (15.5%)	935 (17.9%)

Table 5.29: Results on the two dynamic qualitative criteria for bigram term extraction on the large MEDLINE corpus.

subportion. Only PMI places an even larger proportion of targets into the upper subportions (65.1%).

The analogous results for the two dynamic criteria on the small corpus are given table in 5.30. Similar to the results on the large corpus, t-test manages to place approximately twice as many upper-portion non-targets than log-likelihood to the lower subportions, i.e. 15.2% vs. 7.8%, respectively. LPM again fulfills the criterion to a considerable degree as it is able to demote as much as 43.2% of the non-targets. As almost expected, this is still topped by PMI with 59%, although it appears again that this is merely characteristic of PMI's tendency towards indiscriminate shuffling, as already previously observed.

With regard to criterion 4, C-value even remains more conservative in that only 0.3% of the lower-portion bigram targets are promoted to the third upper subportion. In fact, in the lower subportions C-value even demotes the targets, compared to frequency, in such a way that its lowest subportion exhibits a higher proportion of them than that of frequency (40.7% for C-value vs. 37.8% for frequency). T-test again manages to place one fifth (19.2%) of its lower-portion targets to the into the third upper subportion, more than twice as much than log-likelihood with 8.8% (and 9.1% altogether). LPM is again able to promote more than half of lower-portion targets to

	AM	upper portion (ranks 1 - 9,500)			lower portion (ranks 9,501 - 19,001)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 3	freq	1665 (27.3%)	2117 (34.8%)	2310 (37.9%)	0	0	0
	t-test	1631 (26.8%)	2071 (34.0%)	1464 (24.0%)	307 (5.0%)	79 (1.3%)	540 (8.9%)
6092	logL	1639 (26.9%)	2063 (33.9%)	1911 (31.4%)	298 (4.9%)	63 (1.0%)	118 (1.9%)
NTs	C-value	1624 (26.7%)	2209 (36.3%)	2156 (35.4%)	103 (1.7%)	0	0
	PMI	595 (9.8%)	859 (14.1%)	1037 (17.0%)	1210 (19.7%)	1248 (20.5%)	1143 (18.8%)
	LPM	925 (15.2%)	1239 (20.3%)	1297 (21.3%)	1120 (18.4%)	967 (15.9%)	544 (8.9%)
Crit. 4	freq	0	0	0	617 (29.7%)	675 (32.5%)	787 (37.8%)
	t-test	0	0	400 (19.2%)	655 (31.5%)	641 (30.8%)	383 (18.4%)
2079	logL	3 (0.1%)	4 (0.2%)	183 (8.8%)	750 (36.0%)	666 (32.0%)	473 (22.8%)
Ts	C-value	0	0	7 (0.3%)	589 (28.3%)	637 (30.6%)	846 (40.7%)
	PMI	378 (18.2%)	488 (23.5%)	273 (13.1%)	286 (13.8%)	326 (15.7%)	328 (15.8%)
	LPM	405 (19.5%)	327 (15.7%)	329 (15.8%)	307 (14.8%)	325 (15.6%)	386 (18.6%)

Table 5.30: Results on the two dynamic qualitative criteria for bigram term extraction on the small MEDLINE corpus.

the upper three subportions (51%) and PMI's tendency to reshuffling is exhibited by doing so with even 54.8%.

For trigram term extraction on the large corpus, the qualitative results with respect to criterion 3 are shown in the upper part of table 5.31. It is again C-value whose distribution patterns of non-targets are most identical to those of frequency in that it is only able to promote 0.8% of them to the third upper subportion. Although t-test is a bit more successful in doing so with 5.4%, it underperforms with respect to this criterion compared to the bigram case in which it was able to promote up to 16% of the lower-portion targets. LPM is again very successful in promoting these targets, *viz.* 40.9% altogether, and so is PMI with 54.9%, at least on the surface.

For t-test and LPM, a view from another angle on this criterion is offered by the scatterplots in figures 5.31 and 5.32, in which the rankings of the upper-portion non-targets of frequency are plotted against their ranking in t-test and LPM, respectively. Here it can be seen that, in terms of the rank subportions considered, the t-test non-targets are concentrated along the same line as the frequency non-targets, with only a few being able to break this line and get demoted to a lower subportion. On the other hand, LPM completely breaks the original frequency ranking pattern and scatters the upper portion non-targets in the two possible directions, but with the vast majority

	AM	upper portion (ranks 1 - 14,249)			lower portion (ranks 14,250 - 28,499)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 3	freq	3525 (29.7%)	4064 (34.3%)	4272 (36.0%)	0	0	0
	t-test	3504 (29.5%)	4040 (34.1%)	3685 (31.1%)	307 (2.6%)	78 (0.7%)	247 (2.1%)
11861	C-value	3534 (29.8%)	4068 (34.3%)	4159 (35.1%)	99 (0.8%)	1 (<0.1%)	0
NTs	PMI	1446 (12.2%)	1806 (15.2%)	2027 (17.1%)	2101 (17.1%)	2160 (18.2%)	2321 (19.6%)
	LPM	2283 (19.2%)	2369 (19.9%)	2358 (19.8%)	2081 (17.6%)	1748 (14.7%)	1022 (8.6%)
Crit. 4	freq	0	0	0	409 (38.2%)	349 (32.6%)	313 (29.2%)
	t-test	0	0	72 (6.7%)	432 (40.3%)	342 (31.9%)	225 (21.0%)
1071	C-value	0	0	2 (0.2%)	398 (37.2%)	314 (29.3%)	357 (33.3%)
Ts	PMI	370 (34.5%)	219 (20.4%)	161 (15.0%)	118 (11.0%)	114 (10.6%)	89 (8.3%)
	LPM	255 (23.8%)	233 (21.8%)	216 (20.2%)	153 (14.3%)	104 (9.7%)	110 (10.3%)

Table 5.31: Results on the two dynamic qualitative criteria for trigram term extraction on the large MEDLINE corpus.

of them getting demoted to a lower rank than in frequency.[8]

The results shown in the lower part of table 5.31 again reveal that t-test performs worse with respect to criterion 4 for trigram extraction than for bigram extraction. Whereas it is able to place 21.1% of the bigram lower-portion targets into the third upper subportion (see table 5.29), it only manages to do so with 6.7% in the trigram case. The staggered grouping of lower-portion t-test targets (visualized in the right-hand scatterplot in figure 5.33) actually indicates that there are certain plateaus beyond which the targets cannot get promoted. C-value basically retains the frequency rankings as it only places 2 lower-portion targets into the third upper subportion (0.2%). LPM meets this criterion to an even more substantial degree than it has done for bigrams as it places 65.8% of its lower-portion trigram targets into the three top portions. The respective scatterplot in figure 5.34 additionally shows that this upward movement of targets, like the downward movement of targets in figure 5.32, is bidirectional – but with the vast majority of them getting promoted to a higher rank

[8]Because these scatterplots serve as an illustrative purpose, due to space limitations we restrict ourselves to showing them for the trigrams and for the best-performing association measure, LPM, as well for as the second best-performing one, t-test. Furthermore, the result patterns for both bigrams and quadgrams are very similar and thus additional scatterplots may just clutter the page without purpose. In fact, apart from the dot density (due to more or fewer terms/non-terms) the respective bigram and quadgram scatterplots are almost indistinguishable from the trigram plots.

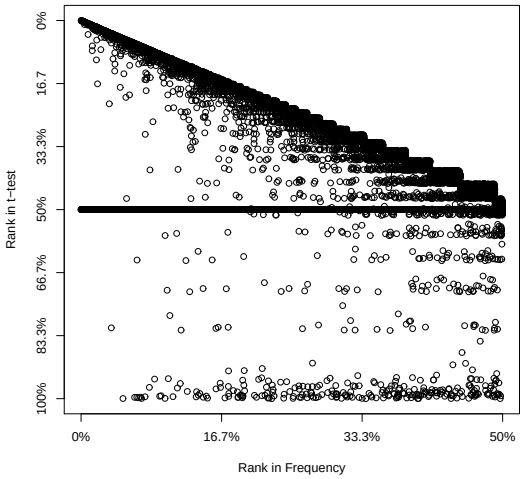

Figure 5.31: Criterion 3 for t-test trigrams on large corpus.

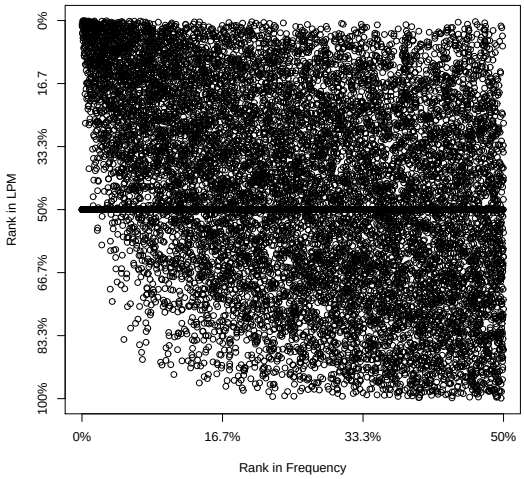

Figure 5.32: Criterion 3 for LPM trigrams on large corpus.

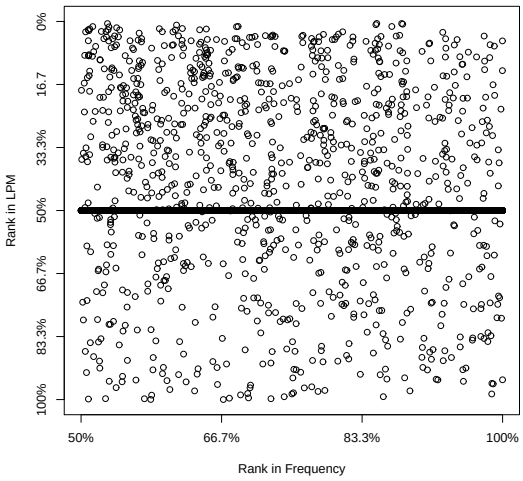

Figure 5.33: Criterion 4 for t-test trigrams on large corpus.

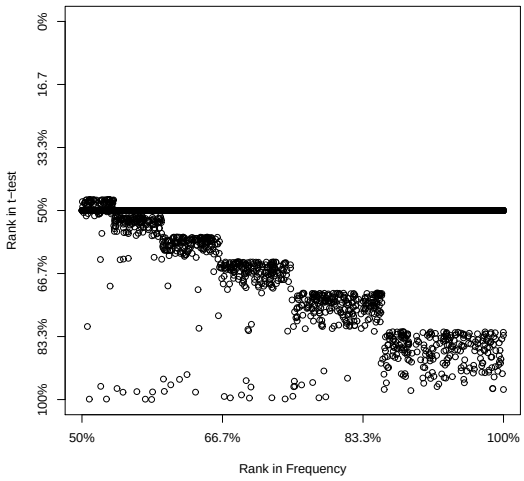

Figure 5.34: Criterion 4 for LPM trigrams on large corpus.

than in frequency.

For trigram term extraction on the small corpus, the qualitative results with respect to criterion 3 are shown in the upper part of table 5.32. T-test is only able to demote 7.1% of its upper-portion non-targets, which is less than half of the proportion it achieves for the bigram term extraction case. On the other hand, LPM even performs better by demoting about 50% of the upper-portion non-targets to the lower three subportions – an increase of over 10 points compared to bigram term extraction.

As already was previously the case, C-value and PMI lie at the two extremes, with C-value basically maintaining the frequency status-quo and PMI shuffling 61.4% of the non-targets downward.

	AM	upper portion (ranks 1 - 2,360)			lower portion (ranks 2,361 - 4,721)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 3	freq	443 (27.0%)	568 (34.7%)	627 (38.3%)	0	0	0
	t-test	424 (25.9%)	566 (34.6%)	531 (32.4%)	55 (3.4%)	9 (0.5%)	53 (3.2%)
1638	C-value	445 (27.2%)	568 (34.7%)	594 (36.3%)	29 (1.8%)	1 (<0.1%)	1 (<0.1%)
NTs	PMI	159 (9.7%)	236 (14.4%)	236 (14.4%)	301 (18.4%)	332 (20.2%)	374 (22.8%)
	LPM	235 (14.3%)	282 (17.2%)	313 (19.1%)	296 (18.1%)	300 (18.3%)	212 (12.9%)
Crit. 4	freq	0	0	0	153 (39.6%)	111 (28.8%)	122 (31.6%)
	t-test	0	0	40 (10.4%)	141 (36.5%)	140 (36.3%)	65 (16.8%)
386	C-value	0	0	4 (1.0%)	150 (38.9%)	131 (33.9%)	101 (26.2%)
Ts	PMI	79 (20.5%)	72 (18.7%)	66 (17.1%)	58 (15.0%)	49 (12.7%)	62 (16.1%)
	LPM	65 (16.8%)	93 (24.1%)	83 (21.5%)	61 (15.8%)	48 (12.4%)	36 (9.3%)

Table 5.32: Results on the two dynamic qualitative criteria for trigram term extraction on the small MEDLINE corpus.

The results shown in the lower part of table 5.32 again reveal that C-value lies at the lower end with respect to criterion 4 for trigram extraction on the small corpus, with only being able to promote a tiny 1% of its lower-portion targets to the third upper subportion. Although t-test is able to promote 10.4% of its lower-portion targets to the third upper subportion, this constitutes only half of the amount achieved in the bigram extraction case. Surprisingly, LPM here even performs best in fulfilling this criterion by promoting 62.4% of its targets. PMI, on the other hand, is only able to do so with 56.3%.

Table 5.33 gives the results on the dynamic criteria for the quadgram term extraction on the large MEDLINE corpus. As can be seen with respect to criterion 3, the measure which again least changes the frequency ranking is C-value, which only demotes 1.1% of the upper-portion non-targets into the first lower subportion and 0.1% into the middle lower subportion. Also here, t-test is less able to demote non-targets to the lower subportions (4.6%) than it is for bigram extraction (16% – see table 5.29). LPM, on the other hand, manages to demote even more here (45.5%) than for bigram extraction (38%) whereas PMI's demotion rate basically remains unchanged with 56.6%.

	AM	upper portion (ranks 1 - 4,929)			lower portion (ranks 4,930 - 9,859)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 3	freq	1312 (30.5%)	1457 (33.9%)	1528 (35.6%)	0	0	0
	t-test	1300 (30.2%)	1456 (33.9%)	1341 (31.2%)	121 (2.8%)	21 (0.5%)	58 (1.3%)
4297	C-value	1310 (30.5%)	1458 (33.9%)	1477 (34.4%)	49 (1.1%)	3 (0.1%)	0
NTs	PMI	550 (12.8%)	614 (14.3%)	702 (16.3%)	753 (17.5%)	833 (19.4%)	845 (19.7%)
	LPM	778 (18.1%)	797 (18.5%)	776 (18.1%)	785 (18.3%)	694 (16.2%)	467 (10.9%)
Crit. 4	freq	0	0	0	112 (43.4%)	86 (33.3%)	60 (23.3%)
	t-test	0	0	20 (7.8%)	117 (45.3%)	84 (32.6%)	37 (14.3%)
258	C-value	0	0	0	101 (39.1%)	89 (34.5%)	68 (26.4%)
Ts	PMI	85 (32.9%)	61 (23.6%)	44 (17.1%)	28 (10.9%)	25 (9.7%)	15 (5.8%)
	LPM	72 (27.9%)	69 (26.7%)	48 (18.6%)	31 (12.0%)	21 (8.1%)	17 (6.6%)

Table 5.33: Results on the two dynamic qualitative criteria for quadgram term extraction on the large MEDLINE corpus.

As regards criterion 4 for quadgram extraction, the results shown in the lower part of table 5.33 even indicate that C-value is not able at all to promote any of the lower-portion targets into the upper portion. Like for the trigram case, t-test is less able to meet this requirement than for bigrams and only promotes 7.8% of the targets. LPM, on the other hand again, manages to promote a record 73.2% of the lower-portion targets into the upper portions – the relative majority of these (27.9%) into the top upper subportion – and thus is on par with the promotion rate of PMI (73.6%).

Finally, for the quadgram extraction task on the small corpus, whose observed brittleness is due the small size of the candidate set, the quantitative results for the two dynamic criteria are given in table 5.34. As concerns criterion 3, t-test manages to demote 9.1% of the upper-portion non-targets to the lower subportions whereas C-value's ranking again remains almost unchanged with only 1.6% of non-targets getting demoted. Both LPM's and PMI's demotion rate (52.8% and 66.6%) for non-targets is higher for this case than for lower-size n-grams.

For criterion 4 presented in the lower portion of table 5.34, the results for C-value and t-test pretty much fall in line with previous ones. Like for the quadgram extraction on the large corpus, C-value is not able to promote any target to the upper portion while t-test does so with 9.1%. Whereas LPM, as expected, promotes 57.6% of its lower-portion-targets to the upper subportions, the results for PMI fall somewhat out

	AM	upper portion (ranks 1 - 456)			lower portion (ranks 457 - 912)		
		0-16.7%	16.7-33.3%	33.3-50%	50-66.7%	66.7-83.3%	83.3-100%
Crit. 3	freq	91 (28.6%)	111 (34.9%)	116 (30.7%)	0	0	0
	t-test	86 (27.0%)	113 (36.5%)	90 (28.3%)	8 (2.5%)	6 (1.9%)	15 (4.7%)
318	C-value	92 (28.9%)	112 (35.2%)	109 (34.3%)	5 (1.6%)	0	0
NTs	PMI	21 (6.6%)	39 (12.3%)	46 (14.5%)	66 (20.7%)	61 (19.2%)	85 (26.7%)
	LPM	47 (14.8%)	50 (15.7%)	53 (16.7%)	62 (19.5%)	57 (17.9%)	49 (15.4%)
Crit. 4	freq	0	0	0	23 (34.8%)	21 (31.8%)	22 (33.3%)
	t-test	0	0	6 (9.1%)	27 (40.9%)	24 (36.4%)	9 (13.6%)
66	C-value	0	0	0	17 (25.8%)	19 (28.8%)	30 (45.5%)
Ts	PMI	8 (12.1%)	9 (13.6%)	8 (12.1%)	15 (22.7%)	12 (18.2%)	14 (21.2%)
	LPM	10 (15.2%)	12 (18.2%)	16 (24.2%)	13 (19.7%)	8 (12.1%)	7 (10.6%)

Table 5.34: Results on the two dynamic qualitative criteria for quadgram term extraction on the small MEDLINE corpus

of line as it promotes the smallest proportion of targets (37.8%), compared to its other results for this criterion. But as already mentioned before, the results reported on this candidate set have to be interpreted with care as they may be rather inconclusive.

5.3 Assessment of Experimental Results

Both the quantitative and the qualitative performance evaluations for PNV triple collocation extraction on German general-language newspaper text data have shown that the linguistically motivated association measure LSM outperforms the standard frequency, statistical and information-theoretic association measures (frequency, t-test, log-likelihood, and PMI) by large margins in every respect. With regard to the quantitative performance evaluation, LSM outperforms its competitors on a variety of established performance measures, such as the precision, recall, and receiver operating curve (ROC) metrics. In addition, this advantage is not due to chance, as it has been shown by applying McNemar as a significance test of differences to the ranked output lists. We have also examined performance from a qualitative perspective. For this purpose, we examined the rankings output by a certain association measure against those given by frequency of co-occurrence. The degree to which an association measure is able to retain or change the rankings of its targets and non-targets is taken to be

a sort of challenging baseline against which the superiority of an association measure may be determined from a qualitative perspective. The results obtained here also clearly indicate that LSM is the only association measure which exhibits superior qualitative characteristics.

The general observations of LSM's quantitative and qualitative superiority hold both when tested on a large corpus sample of 114 million words and on a smaller one of 10 million words and thus are independent of corpus size. Interestingly, the advantage of LSM is even more marked when applied on the small corpus. This may have to do with the fact that on the large corpus a higher candidate frequency plays a bigger role as an ingredient to the other association measures which all have some sort of frequency factor (i.e. O_{11} – see subsection 3.3.1) in their computation. This is most evidenced for the almost erratic behavior of PMI which severly underperforms all other measures and even runs below the precision baseline.

Our performance evaluations also have clearly shown that frequency of co-occurrence is quite a competitive association measure for collocation extraction – a finding which is in line with previous studies (see sections 3.1 and 3.2). Indeed, frequency of co-occurrence performs as the second best association measure for collocation extraction but however turns out to only have a slight advantage compared to t-test. That t-test performs so similar to frequency may be neatly explained by the fact that it actually *behaves* most similar to frequency: as shown for the qualitative evaluation it actually changes the frequency rankings the least compared to all other association measures considered.

Another reason why LSM performs so much better may be given by the qualitative evaluation results which shows that LSM, by promoting its lower-portion targets to the upper portion (in particular, the third upper subportion), yields a big recall boost compared to all other association measures. This is of course also of substantial practical relevance: a measure which returns most actual collocations earlier than others is advantageous as humans (e.g. lexicographers) may want to post-process such a ranked output list. In addition, what may also add to this superior recall is the fact LSM is best at keeping its targets in the upper portion (compared to frequency). Indeed, it is the only measure that is actually able to outperform frequency at all.

Whereas we have shown in section 4.3 why LSM may be regarded as a conceptually sound linguistic property of collocations, subsection 5.1.3 has empirically demonstrated that collocations actually do posses limited syntagmatic modifiability com-

pared to non-collocations. From a linguistic point of view, it is interesting to note that, contrary to linguistic perceptions, this property equally holds for all three subcategories of collocations alike, i.e. for idioms, for support verb constructions/narrow collocations and for fixed phrases.

In a similar vein to our experimental results for collocation extraction, the quantitative and qualitative performance evaluations conducted for term extraction on the English-(sub)language biomedical subdomain of Hematopoietic Stem Cell Transplantation and Immunology have shown that a linguistically motivated term extraction measure, LPM, is also able to clearly outperform the same standard statistical and information-theoretic measures (t-test, log-likelihood, PMI) as well as frequency of co-occurrence and the frequency-based C-value measure by substantial margins in almost every respect. Again, we conducted the quantitative performance evaluation on the precision, recall, and ROC performance metrics and tested for the significance of differences using McNemar. Although the quantitative result effects are superficially very similar for the experiments on collocation and term extraction in that in both cases the linguistically motivated measures evidently outrank their competitors, certain differences in the term extraction case may nevertheless be observed. For once, it is noticeable that for term extraction it is t-test, and not frequency of co-occurrence like for collocation extraction, which runs as the second-best performing association measure. A reason for this may be sought in the qualitative evaluation in which for the dynamic criteria t-test manages to do more favorable re-rankings while at the same time being able to keep its items in place for the static criteria. In fact, also log-likelihood seems to be running partly better than frequency of co-occurrence even though the advantage is admittedly rather small. Moreover, as we have already pointed out on several occasions, because log-likelihood is inherently only applicable on bigram data, it is not a serious candidate measure for practical collocation and term extraction tasks. As for the only association measure actually devised for term extraction, C-value, the qualitative evaluation results clearly show that it is the most conservative of all measures in that all its rankings basically fall in line with those of frequency of co-occurrence.

As far as the the information-theoretic PMI measure is concerned, it is actually interesting to note that for bigram term extraction on the large corpus it performs on par with the other standard association measures. It is only on the small corpora and with increasing n-gram length that its performance scores deteriorate. A look

at the qualitative performance results actually reveals that, in comparison to the collocation extraction task, the ranking patterns for targets and non-targets are not really different as PMI exhibits a strong tendency to reshuffling in both cases. Rather, what seems to be the determining factor here is that PMI does not appear to be able to handle smaller-sized candidate sets well in that the negative effects of massively reshuffling items appear to become stronger. Whereas PMI yields its best results on bigram term extraction, it is interesting to note that this is also by far the biggest candidate set with 66,669 term candidate types. As the sizes of both the candidate sets for the small corpora and for the larger-sized n-grams shrink substantially, so does PMI's ability to keep up with the other association measures.

On the quantitative side of the evaluation, there are virtually no differences between the linguistically motivated association measures for collocation extraction (LSM) and for term extraction (LPM) as both clearly outperform their standard competitors. Still, a look at the results from the qualitative perspective reveals some differences. Although the results for the static criteria show that although LSM performs a modest degree of re-rankings on upper-portion targets and on lower-portion non-targets, this effect is much more pronounced for LPM for criterion 2 (promoting lower-portion non-targets) while it visible but less pronounced for criterion 1 (demoting upper-portion targets).[9] However, with regard to the results on the dynamic criteria, even though both measures manage to demote their upper-portion non-targets and promote their lower-portion targets to a considerable extent, LPM does so to a much more substantial degree and thus remedies any possible negative effects from the static criteria.

Finally, a note is in order concerning the somewhat distinct nature of the results for the extraction of quadgrams on the small corpus. Whereas in this case the qualitative results essentially fall in line with those for the other settings considered, it is the quantitative results which, although showing a distinct advantage for LPM on the standard performance metrics, do not turn out to be statistically significant in terms

[9] As was already pointed out in subsection 4.4.2, LPM will demote true terms in their ranking, if their paradigmatic modifiability is less limited, which seems to the case if one or more of the word tokens of a particular term often occur in the same k-slot of other equal-length n-grams, as it it the case with the word token "*cell*" in the third slot of the trigram term "*bone marrow cell*". Likewise, of course, it will promote items classified as non-terms if their paradigmatic modifiability is limited. But this behavior is greatly outweighed by appropriate re-rankings done on lower-portion targets and upper-portion non-targets.

of the McNemar test.[10] A look at the size of the candidate set reveals that, given a frequency cut-off threshold of five, it only amounts to 912 quadgram term candidates. Given the comparatively long n-gram size and given the smaller size of the corpus, however, this small amount of candidate terms is not unusual but rather presents a well-known phenomenon from other NLP domains such as language modeling (Jurafsky & Martin, 2000). For LSM, on the contrary, small sized candidate sets still lead to statistically significant performance differences (see subsubsection 5.1.1.2 above), although the candidate set here only amounts to 1035 collocation candidates on the small 10 million word corpus. What these differences between the two linguistically enhanced assocation measures seem to indicate, then, is that LSM for collocation extraction is more resistant to small candidate sets than LPM for term extraction.

For term extraction, one final note is in order about n-gram term candidates of size $n > 4$. As has already been described in subsection 2.2.7, Justeson & Katz (1995) found in their study that only 6% of all terms were quadgrams in the first place. Hence, employing more or less sophisticated association measures to the extraction of pentagrams (or even larger-sized n-grams) may not be worth while and the use of frequency of co-occurrence instead may just be a cheaper and equally viable solution. For all smaller-sized n-grams, however, the use of the linguistically enhanced statistical method LPM is certainly advantageous and preferable, given a reasonably sized candidate set.

[10] Although it is certainly desirable in terms of scientific rigor, it should also be noted that we applied the McNemar test for a very strict confidence interval of 99%, whereas many other studies make do with a 95% confidence interval.

Chapter 6

Conclusions and Outlook

The research presented in this thesis has shed new light on how computational approaches should address the extraction of collocations and terms from natural language text corpora. We started out our enterprise with one of the most famous slogans of 20th century linguistics – Firth (1957)'s "*You shall know a word by the company it keeps!*". With the insights gained from this work, we might want to add the phrase "*... in its syntagmatic and paradigmatic context*". And in fact, we have seen that the linguistic research literature on collocations and terms has provided us with some crucial insights about the characteristic properties of collocations and terms.[1] Firstly, they have enabled us to isolate one particular linguistic property shared by both kinds of linguistic expressions, *viz.* limited modifiability. Secondly, they have provided us with an appropriate linguistic frame to structure this property, *viz.* the lexical-collocational layer of Firth's (1957) model of language description, in particular its syntagmatic and paradigmatic contexts. That this embodiment is necessary has been extensively shown because, after all, collocations and terms are different linguistic entities that surface in different linguistic contexts, both syntactically and pragmatically. From the syntactic perspective, it has become clear that while collocations may manifest themselves in a variety of syntactic constructions, the surface manifestation of terms is basically limited to noun phrases. In addition, the fact that collocation and term candidates may be filtered out from pre-definable linguistic structures has also clearly underscored the necessity for linguistic prepro-

[1] This also includes, to a certain extent at least, the non-linguistic research literature on terms and terminology – cf. section 2.2.

cessing of text corpora. From a pragmatic perspective, we have demonstrated that collocations may be best conceived of as general-language constructions whereas terms are relevant in subject-specific sublanguage domains.

From these linguistic findings, we were then able to structure the property of limited modifiability in such a way that we could forge it in to observable, formalizable and quantifiable terms in order to serve as linguistic parameters for two distinct statistical computations measuring lexical association in a language- and domain-independent manner. In order to compute the degree of collocaticity of a collocation candidate, limited syntagmatic modifiability (LSM) incorporates the tendency of collocations to limit the number of potential syntagmatic attachments, while, in order to determine the degree of termhood on a term candidate, limited paradigmatic modifiability (LPM) incorporates the tendency of terms to limit the number of potential paradigmatic substitutions. One notable feature of both lexical association measures is that, besides utilizing limited modifiability, they also exploit frequency of co-occurrence as another linguistically prominent property of both collocations and terms. Frequency of co-occurrence is in fact the one linguistic attribute from the British contextualist linguistic tradition that has greatly influenced the use of various standard statistical and information-theoretic association measures (or test statistics) for collocation and term extraction approaches in the first place, as they basically exploit this information in different forms in their statistical computations. In fact, our review of the research literature on computational approaches to collocation and term extraction has actually shown that frequency of co-occurrence fares competitively well against the more complex statistical association measures.

While we were able to put our newly devised lexical association measures on a theoretically and definitionally sound base, we also had to validate that our whole endeavor would fulfill our assumptions, i.e. that linguistically more informed lexical association measures would outperform their standard competitors to a substantial degree – thus making the whole enterprise worthwhile in the first place. For this purpose, we established a comprehensive comparative performance evaluation setting, which not only ran a wide array of standard quantitative performance metrics, but also applied a new qualitative performance evaluation metric that compared the output rankings of an association measure to frequency of co-occurrence as a challenging baseline. For collocation extraction, the evaluation setting was on German-language preposition-noun-verb collocation candidates and for term extraction it was

on English-language noun phrase term candidates from the biomedical subdomain of immunology. Both the quantitative and the qualitative performance evaluations showed that our assumptions were correct. For the tasks of collocation and term extraction from text corpora, the linguistically motivated association measures LSM and LPM outperformed their standard frequency-based, statistical and information-theoretic lexical association measures by large margins in every respect.

The research presented in this thesis may, of course, be built upon in various further research directions. Although this work corroborates the essential assumption held by many other researchers (e.g. Evert (2005), Jacquemin (2001), and others), *viz.* that the crucial backbone of any approach to collocation and term extraction is a high-performance lexical association measure, it is expansible in many different ways which have already been hinted at in this work. For the task of collocation extraction, a logical next step would be to transfer the LSM approach to other types of syntactic constructions which may harbor potential collocation candidates, such as noun-verb or noun-noun-verb constructions. Whether or not an equally comprehensive (and labor-intensive) evaluation as presented here is necessary remains to be seen insomuch as our research results clearly show that LSM is a superior lexical measure for collocativity and, thus, is applicable to any linguistic construction involving phrasal elements as the place to locate syntagmatic attachments. Given this, the acquired sets of collocations may be put in a lexical database[2] in which the entries may be enriched with additional kinds of information. On the one hand, it may be possible to utilize syntagmatic attachments in such a way as to associate particular collocation entries with possible lexico-semantic modifications, along the lines already outlined in subsection 4.3.2. On the other hand, the acquired set of collocations may also be further classified into collocational subclasses using a similar approach as suggested by Lin (1999).[3] Finally, such a collocation database may also be used as input to syntactic parsers and semantic interpreters in order to prevent them from performing superfluous syntactic assignments or semantic interpretations.

As for the extraction of terms, a logical extension to the LPM-based approach to compute the termhood of term candidates would be to place it in the wider context

[2]This is particularly relevant, as it has been shown time and again that existing collocational lexicons are incomplete (Evert & Krenn, 2001; Lin, 1998b).

[3]Being in need of a thesaurus-like lexicon, the German-language Open Office thesaurus (`www.openthesaurus.de`) may serve as an appropriate resource for this task.

of computational terminology, as e.g. outlined by Jacquemin (2001). One issue is certainly that, once a sound basis of terminological data has been acquired by LPM, it is necessary to enrich the respective terms with potential variants. While associating acronyms with their respective full forms is a comparatively easy task in this respect, a more challenging task is to harvest syntactic variants which still denote the same term concept.[4] An approach such as the one taken by Jacquemin's (2001) FASTR formalism, in which a large set of term-specific grammar rules is devised and implemented in a complex feature unification-based formalism, would most likely not only be too laborious and costly, but also be more error-prone and substantially overshoot the target by rendering too many false positives that would need post-editing, as conceded by Jacquemin (2001) himself. Thus, employing a shallower approach to syntactic term variant recognition (e.g. by applying a phrase chunker, some form of stemming and string edit distance matching) would most likely be not only equally viable but also more feasible in terms of effort to invest. Once such an enriched terminological database has been set up, a potential next step could be to associate terms to each other through semantic relations, such as taxonomic relations, in order to create a thesaurus-like structure. Here again, various approaches are conceivable, ranging from full-fledged complex unification formalism in the line of Jacquemin (2001) to shallower procedures based on deleted noun phrase modifiers.[5]

Finally, it should be recalled again that, despite all the promising lines of further research, the crux of the whole enterprise of term and collocation extraction still remains a high-quality lexical association measure. On the one hand, there are technical domains and subject fields which either only possess insufficient terminological resources or even lack them completely. In this sense, not every subject domain is as blessed as the biomedical field with the UMLS resource.[6] In a similar vein, collocational lexicons have shown to be notoriously underspecified. On the other hand, even if resources like the UMLS are on hand for a particular domain, the nature of natural language, i.e. its creativity and productivity, guarantees that new collocations and new terms are constantly being coined.

[4]It should be recalled that the linguistic preprocessing applied in this work already accounts for morphological variation.

[5]In this way, it would e.g. be possible to establish a taxonomic link between "*stem cell*" and "*hematopoietic stem cell*".

[6]In fact, the UMLS resource is probably unique both in terms of its size and dimension.

Chapter 7

Summary

The research presented in this thesis substantiates, defines and evaluates two new linguistically motivated statistical association measures in a language- and domain-independent manner, limited syntagmatic modifiability (LSM) for collocation extraction, and limited paradigmatic modifiability (LPM) for term extraction. The task they are designed for – computing lexical association scores to determine the degree of collocativity and termhood of collocation and term candidates – is the crucial backbone of any approach to collocation and term extraction and, thus, resembles a wide variety of standard frequency-based, statistical and information-theoretic association measures put forth in the computational linguistics research literature. What distinguishes LSM and LPM is that their defining parameters are based on actual linguistic properties of the targeted linguistic constructions, *viz.* collocations and terms.

The central linguistic property which is isolated in the linguistic research literature and which is shared by collocations and terms is denoted by the notion of limited modifiability. This property is parameterized in such a way as to account for the obvious linguistic differences between collocations and terms in that collocations are typically manifested in general language and surface in a variety of syntactic constructions, while terms are typically confined to noun phrases manifested in domain-specific sublanguage. Limited modifiability is embedded within an appropriate linguistic frame of reference – the lexical-collocational layer of Firth (1957)'s contextualist model of language description. With the help of this model, the linguistic differences are realized as limited syntagmatic modifiability, in the case of collocations, and as limited paradigmatic modifiability, in the case of terms. The respective linguistically en-

hanced lexical association measures exploit these properties as observable and quantifiable parameters to their statistical computations in that LSM incorporates the tendency of collocations to limit the number of potential syntagmatic attachments whereas LPM incorporates the tendency of terms to limit the number of potential paradigmatic substitutions. Frequency of co-occurrence is another prominent linguistic property incorporated into both linguistic association measures and is the only linguistic property also exploited by other standard frequency-based, statistical and information-theoretic association measures for collocation and term extraction.

In order to compare the linguistically enhanced lexical association measures LSM and LPM against their standard competitors, a comprehensive performance evaluation setting is established – for collocation extraction on German-language preposition-noun-word collocation candidates and for term extraction on English-language noun phrase term candidates from a biomedical subdomain. In this setting, a wide array of standard quantitative performance metrics is applied as well as, in addition, a new qualitative performance evaluation metric which compares the output rankings of an association measure to the challenging baseline of frequency of co-occurrence. All experimental results show that LSM and LPM outperform the other frequency-based, statistical and information-theoretic lexical association measures by large margins in every aspect of performance evaluation considered. Thus, lexical association measures which base their statistical computations on linguistic parameters instead of standard statistical ones not only exhibit conceptual but also empirical superiority.

Appendix A

Collocation Classification Manual

This appendix contains the annotation manual given to the three human annotators (graduate students of German linguistics) for their classification task of German PNV triples. These guidelines, which are written in German, include the linguistic properties described in subsubsection 2.1.4.1 and a description of the three collocational classes and how they may be distinguished from free word combinations, as outlined in subsubsection 2.1.4.2.

A.1 Definitionen aus der einschlägigen Literatur

"Kollokationen: eine Sequenz aus zwei oder mehreren Wörtern, welche die Eigenschaft einer syntaktischen und semantischen Einheit hat. Ihre exakte und eindeutige Bedeutung kann nicht direkt aus der Bedeutung ihrer Komponenten abgeleitet werden."

"Kollokationen bezeichnen charakteristische, häufig auftretende Wortgruppen. Weitgefasst, subsumieren sie voll idiomatisierte Wendungen ('am Herzen liegen'), Funktionsverbgefüge ('zur Verfügung stellen') und semantisch transparente Gruppen ('Zähne putzen')."

"Kollokationen sind (meist) Paare von lexikalischen Zeichen, die durch häufiges Kovorkommen (innerhalb einer Phrase) eine halbfeste Verbindung eingehen. Die lexikalischen Zeichen werden in ihrer eigentlichen, d.h. im Wörterbuch kodierten Bedeutung verwendet. Die Gesamtbedeutung der Gruppe ist keine direkte Funktion der Einzelbedeutungen. Die Verbindung ist so fest, dass Quasi-Synonyme als Alter-

nativen für eines der Elemente merkwürdig klingen. Eine Variante der Kollokationen sind die Funktionsverbgefüge, die aus einem sehr allgemeinen Verb und einem die eigentliche Bedeutung tragenden, in einer PP eingebetteten Nomen bestehen."

"Kollokationen bestehen aus einer Basis ('Antrag', 'Haar') und einem Kollokator ('stellen', 'schütter')."

"Funktionsverben (wie 'bringen', 'kommen', 'finden', 'stehen', 'nehmen' u.a.) sind eine Teilmenge der Verben, die in bestimmten Kontexten ihre lexikalische Bedeutung als Vollverb fast ganz verloren haben. Die Hauptbedeutung in solch einem Funktionsverbgefüge wird von einem Substantiv, einer Präpositionalphrase oder einem Adjektiv getragen."

in Frage stellen – hinterfragen
instand setzen – reparieren
zu Grunde richten – zerstören
zur Schau stellen – zeigen
Rechnung tragen – berücksichtigen
auf den Weg machen – losgehen
Abschied nehmen – sich verabschieden

Funktionsverb und Ergänzung bilden den Satzrahmen.
Er machte sich sofort nach dem Essen auf den Weg.

A.2 Unsere Richtlinien für die Verwendung des Begriffs 'Kollokation'

Wir werden den Begriff 'Kollokation' in einem **weitgefassten** Sinn verwenden. Obwohl wir uns bei dieser Studie nur mit (P)räposition-(N)ominalgruppe-(V)erb-Verbindungen beschäftigen, gelten die drei hier aufgeführten Kollokationsklassen auch für andere Kombinationen. Um die drei Klassen von einander unterscheiden zu können, werden wir die Bestandteile (Präposition-Nominalgruppe auf der einen Seite, Verb auf der anderen Seite) unter ihrer ***naiven lexikalischen Grundbedeutung*** (NLG) betrachten.

A.2 Unsere Richtlinien für die Verwendung des Begriffs 'Kollokation'209

1. **Idiomatische Phrasen** ("am Herzen liegen"). Betrachtet man die NLGs von "liegen" und "am Herzen" und 'komponiert' die beiden zusammen, so erkennt man schnell, dass keine der beiden semantisch transparent in die Komposition mit einfließt. Die Bedeutung des Ganzen ist figurativ/metaphorisch zu sehen. Weitere Beispiele:

   - auf die Schippe nehmen
   - mit einem blauen Auge davonkommen
   - auf dem falschen Fuß erwischen

2. **Enge Kollokationen** bzw. **Funktionsverbgefüge** ("zur Kenntnis nehmen", "in Ordnung bringen"). Bei diesen Kombinationen ist wenigstens eine Komponente transparent, d.h. ihre NLG trägt zur Gesamtbedeutung bei: daher bildet sie normalerweise den "semantischen Kern" oder die Basis des Ausdrucks. Im Beispiel sind die NLGs von "Kenntnis" und "Ordnung" die semantischen Kerne. Weitere Beispiele:

   - zur Verfügung stehen
   - in Frage stellen
   - in den Hintergrund treten
   - aus eigener Tasche bezahlen

3. **Feste Wendungen** ("im Koma liegen"). Die Bedeutung dieser Kombinationen ist semantisch viel transparenter als die vorigen Beispiele. Das heißt, es ist erkennbar wie die NLGs der einzelnen Komponenten zur Gesamtbedeutung beitragen. Allerdings tun sie das nicht vollständig: man hat das Gefühl, dass dies noch keine freien, willkürlichen Wortkombinationen sind und dass die PNV-Wörter (d.h. alle Komponenten) 'irgendwie' zusammengehören. Man vergleiche "im Koma liegen" (feste Wendung) mit "auf der Strasse liegen" (keine feste Wendung). Weiter Beispiele:

   - ums Überleben kämpfen
   - in die Hand drücken
   - auf der Brust tragen

Bis jetzt sind Kollokationen nur positiv definiert worden (d.h., was sie sein können), aber es ist auch notwendig, ein paar **negative** Kriterien zu bestimmen (d.h. was sind Eigenschaften, die Kollokationen normalerweise nicht besitzen).

- Nicht- oder limitierte **Kompositionalität**. Die Bedeutung einer Kollokation kann nicht direkt aus der Bedeutung ihrer Komponenten abgeleitet werden. Vieles dazu ist schon oben gesagt worden.
- **Nicht- bzw. limitierte Ersetzbarkeit**. Die Komponenten einer Kollokation können nicht oder nur schwer durch gleiche oder ähnliche Wörter ersetzt werden. So kann man "zur Last legen" nicht durch "zur Last stellen/setzen" oder zum "zum Gewicht legen" ersetzen (man beachte auch, dass sich hierbei die Präposition von "zur" nach "zum" geändert hat – auch illegal!). Aber man kann z.B. "um 20 Prozent steigen" zu "auf 20 Prozent klettern" oder zu "um 30 Prozent steigen" ändern. Daher ist dies keine Kollokation!
- **Nicht- oder limitierte Modifizierbarkeit**. Viele Kollokationen sind nicht frei oder nur schwer modifizierbar: "zur Kasse bitten", aber nicht "zur vollen/mit Geld gefüllten/leeren Kasse bitten".

Es ist klar, dass sehr selten alle diese Kriterien zutreffen. Daher muss dies bei der Klassifizierung genau abgewogen werden. Man kann aber ziemlich sicher sein, dass, wenn keines der Kriterien zutrifft, man es nicht mit einer Kollokation zu tun hat.

Daher stehen im **Kontrast** zu diesen drei Klassen freie, willkürliche Wortkombinationen ("auf der Strasse gehen", "sich unter dem Bett verstecken"), die nicht als Kollokation klassifiziert werden.

A.3 Details zur Klassifizierung

Die Klassifizierung der (P)räposition-(N)ominalphrase-(V)erb Kombinationen ist in einem Excel-Spreadsheet vorzunehmen. Es gibt drei Spalten für die Klassifizierung:

- **Grobklassifizierung**: Ist der Ausdruck eine Kollokation im weitgefassten Sinn von 1., 2., oder 3. oben? Wenn ja: 1; wenn nein: 0.

- **Feinklassifizierung**: Wenn es eine Kollokation ist, welche der 3 Klassen von oben trifft am ehesten zu (1., 2., oder 3.)? Wenn es keine ist, wieder eine 0 geben.

- **Ambiguitäts-Klassifizierung**: Manche PNV-Verbindungen haben – je nach Kontext – eine kollokative oder eine wörtliche, NLG-artige Bedeutung. Kann der Ausdruck beides sein? Wenn ja: 1, wenn nein: 0.
  Beispiel: "aus dem Feld schlagen";
  kollokative Bedeutung: einen Gegner/Widersacher/Konkurrent beseitigen/besiegen.
  wörtliche Bedeutung: den Ball (z.B. beim Fußball) aus dem Feld schlagen.
  Weitere Beispiele: "im Abseits stehen"; "auf die Nase fallen"

- **Duden-Klassifizierung**: Findet sich dieser Ausdruck im Duden Band 11, *Redewendungen*? Wenn ja: 1; wenn nein 0;

Ein paar weitere Hinweise zur Klassifizierung:

- Die PNV-Kombinationen wurden automatisch aus einem großen Textkorpus extrahiert. Das bedeutet, dass es auch einige ungrammatische bzw. 'nicht richtig'-klingende Kombinationen gibt. Bitte diese mit 0 markieren.

- Manchmal sind die PNV-Kombinationen eigentlich Teil einer größeren Kollokation (z.B. "mit einer Klappe schlagen" → "zwei Fliegen mit einer Klappe schlagen"; "in den Ring werfen" → "den Hut in den Ring werfen"). Diese auch mit 0 markieren, da es ja nicht die komplette Kollokation ist.

- Achtung vor festen (P)räposition-(V)erb Kombinationen! So ist z.B. bei "um Geld gehen" eine feste Verbindung zu finden – nämlich "gehen um" (es geht um ...). Aber dies ist keine PNV- sondern 'nur' eine PV-Kollokationen/Verbindung. d.h. nur zwei der drei Komponenten gehören zusammen. Deswegen mit 0 markieren!
  Andere Beispiele: "sich **auf** das Spiel **konzentrieren**", "**um** seinen Job **bangen**"

- Ein gute Strategie, um bei schwierigen Kandidaten zur richtigen Klassifizierung zu gelangen (außer die schon genannten Richtlinien oben), ist, die Kombination in einen Kontext einzubinden, d.h. einen oder mehrere Sätze damit zu produzieren, auch mit verschiedenen Wortstellungen; oder versuchen, einen Wörterbucheintrag zu 'imitieren'.

  - Beispiel: "am Herzen liegen" – "Diese Uhr lag mir sehr am Herzen"; "jmdm etw am Herzen liegen"
  - Wichtig!!! Internet: Die Kombination in **Google** in **Anführungszeichen** eintippen. Auch mit Wildcards (*) probieren. Es werden eine Menge von Kontexte zurückgegeben.
    Tipp: dies ist eine wertvolle Strategie! Deshalb beim Klassifizieren unbedingt online sein!

Appendix B

MeSH Terms and UMLS Source Vocabularies

This appendix itemizes the MeSH index terms that were used to obtain our biomedical text corpus (section B.1). It also lists the Umls source vocabularies selected to classify our term candidate sets (section B.2).

B.1 MeSH Terms

The following MeSH index terms describe the domain of Hematopoietic Stem Cell Transplantation and Immunology and were chosen in consultation with a domain expert. They were used to query Medline and download approximately 400,000 abstracts, which amounted to 100 million words of text material (see subsubsection 4.5.3.1).

- Haplotypes
- Immunoglobulins
- Antibodies
- Histocompatibility Antigens
- Antigens
- Leukocytes
- Bone Marrow Cells

- Antibody-Producing Cells
- Antigen-Presenting Cells
- Alleles
- Antigenic Variation
- Cytokines
- Cytokine Receptors
- Bone Marrow Transplantation
- Graft vs Host Disease
- Graft vs Host Reaction
- Graft vs Leukemia Effect
- Graft vs Tumor Effect
- Host vs Graft Reaction
- Hematologic Neoplasms
- Hematopoietic Stem Cell Transplantation
- Hematopoietic Stem Cells
- Histocompatibility
- Histocompatibility Testing
- Immunosuppression
- Leukemia
- Major Histocompatibility Complex
- Minor Histocompatibility Antigens
- Minor Histocompatibility Loci
- Genetic Variation
- Stem Cell Transplantation
- Transplantation Conditioning
- Transplantation Immunology
- Heterologous Transplantation

B.2 UMLS Source Vocabularies

This section contains the list of UMLS source vocabularies included in our experiments, which are deemed to contain relevant terms for the domain of Hematopoietic Stem Cell Transplantation and Immunology. UMLS vocabularies were excluded either if they are basically subsumed by other ones (e.g. the diagnosis vocabulary ICD-9 is subsumed by ICD-10) or if they are completely unrelated to the domain under consideration (e.g. the numerous vocabularies on health care, nursing, billing codes, dental medicine, psychology, and consumer health). We included an additional terminology in this set which is not included in the 2006 edition of the UMLS but which is highly relevant for our domain, *viz.* the OBO CELL ONTOLOGY.[1]

The term candidates in our term candidate sets were assigned the status of being an actual term if they were found in any of these vocabularies it (see subsubsection 4.5.3.3).

- ICD-10: International Statistical Classification of Diseases and Related Health Problems: 10th Revision
- SNOMED-CT: Systematized Nomenclature of Medicine - Clinical Terms
- MESH-2005: Medical Subject Headings, 2005 Edition
- OMIN: Online Mendelian Inheritance in Man
- NCI-2005: NCI Thesaurus, 2005 Edition
- UWDA: University of Washington Digital Anatomist
- GO-2005: Gene Ontology
- NCBI-2005: NCBI Taxonomy
- HL7: Health Level Seven Vocabulary

[1] `http://obofoundry.org/cgi-bin/detail.cgi?id=cell`

Bibliography

Abney, Steven (1991). Parsing by Chunks. In Robert Berwick, Steven Abney & Carol Tenny (Eds.), *Principle-Based Parsing*, pp. 257–278. Dordrecht: Kluwer Academic Publishers.

Abney, Steven (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4):337–344.

Agresti, Alan (1990). *Categorial Data Analysis*. New York: John Wiley & Sons.

Agresti, Alan (1992). A survey of exact inference for contigency tables. *Statistical Science*, 7(1):131–153.

Alshawi, Hiyan & David Carter (1994). Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648.

Baeza-Yates, Ricardo & Berthier Ribeiro-Neto (Eds.) (1999). *Modern Information Retrieval*. Reading, MA: Addison-Wesley & Longman.

Bartsch, Sabine (2004). *Structural and Functional Properties of Collocations in English – a corpus study on lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Gunter Narr Verlag.

Benson, Morton (1989). The structure of the collocational dictionary. *International Journal of Lexicography*, 2(1):1–14.

Benson, Morton (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23–35.

Benson, Morton, Evelyn Benson & Robert Ilson (1986a). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam: John Benjamins.

Benson, Morton, Evelyn Benson & Robert Ilson (1986b). *Lexicographic Description of English*. Studies in Language Companion Series, No 14. Amsterdam: John Benjamins.

Benson, Morton, Evelyn Benson & Robert Ilson (1997). *The BBI dictionary of English word combinations. Revised Edition*. Amsterdam: John Benjamins.

Berry-Rogghe, Godelieve L.M (1973). The computation of collocations and their relevance in lexical studies. In Adam J. Aitken, Richard W. Bailey & Neil Hamilton-Smith (Eds.), *The Computer and Literary Studies*, pp. 103–112. Edinburgh: Edinburgh University Press.

Biber, Douglas (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–241.

Bodenreider, Olivier (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Bourigault, Didier (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *COLING'92 – Proceedings of the 14th International Conference on Computational Linguistics*, pp. 977–981. Nantes, France, August 23–28, 1992, Association for Computational Linguistics Morristown, NJ, USA.

Bourigault, Didier (1995). Lexter, a terminology extraction software for knowledge acquisition from texts. In *KAW'95 – Proceedings of the 9th Knowledge Acquisition for Knowledge Based System Workshop*. Banff, Canada, February 26 – March 2, 1995.

Brants, Thorsten (2000). TnT: A statistical part-of-speech tagger. In *ANLP 2000 – Proceedings of the 6th Conference on Applied Natural Language Processing*, pp. 224–231. Seattle, Washington, USA, April 29 - May 4, 2000. San Francisco, CA: Morgan Kaufmann.

Breidt, Elisabeth (1993). Extraction of v-n-collocations from text corpora: A feasibility study for german. In *VCL'01 – Proceedings of the 1st ACL Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pp. 74–83. Ohio State University, Columbis, OH, USA, June 22, 1993. San Francisco, CA: Morgan Kaufmann.

Browne, Allen C., Guy Divita, Van Nguyen & Vincent C. Cheng (1998). Modular text processing system based on the SPECIALIST lexicon and lexical tools. In C. G. Chute (Ed.), *AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium. A Paradigm Shift in Health Care Information Systems: Clinical Infrastructures for the 21st Century*, p. 982. Orlando, FL, November 7-11, 1998. Philadelphia, PA: Hanley & Belfus.

Burger, Harald (2003). *Phraseologie. Eine Einführung am Beispiel des Deutschen* (2nd ed.). Berlin: Erich Schmidt Verlag.

Cabré Castellví, Teresa M. (2003). Theories of terminology: their description, prescription and explanation. *Terminology*, 9(2):163–199.

Cann, Ronnie (1993). *Formal Semantics – an Introduction*. Cambridge Textbooks in Linguistics. Cambridge, UK: Cambridge University Press.

Carletta, Jean (1996). Assessing agreement on classification tasks: The kappa statistics. *Computational Linguistics*, 22(2):249–254.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.

Chomsky, Noam (1995). *The Minimalist Program*. Cambridge, Mass.: MIT Press.

Choueka, Yaacov (1988). Looking for needles in the haystack or locating interesting collocational expressions in large textual databases. In *RAIO'88 – Proceedings of RIAO*, pp. 38–43.

Church, Kenneth, William A. Gale, Patrick Hanks & Donald Hindle (1991). Using statistics in lexical analysis. In Uri Zernik (Ed.), *Lexical Acquisition. Using Online Resources to Build a Lexicon*, pp. 115–164. Hillsdale, NJ: Lawrence Erlbaum Associates.

Church, Kenneth W (1995). One term or two? In Edward A. Fox, Peter Ingwersen & Raya Fidel (Eds.), *SIGIR'95 – Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 310–318. Seattle, Washington, USA, July 9-13, 1995. ACM Press.

Church, Kenneth W. & Patrick Hanks (1989). Word association norms, mutual information, and lexicography. In *ACL'89 – Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pp. 76–83. Vancouver, B.C., Canada, 26-29 June 1989. San Francisco, CA: Morgan Kaufmann.

Church, Kenneth W. & Patrick Hanks (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cohen, J (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, 20(1):37–46.

Collier, Nigel, Chikashi Nobata & Jun'ichi Tsujii (2002). Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Terminology*, 7(2):239–257.

Collins, Michael J. (1997). Three generative, lexicalized models for statistical parsing. In *ACL'97 – Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 16–23. Madrid, Spain, 7-12 June 1997. San Francisco, CA: Morgan Kaufmann.

Cover, Thomas M. & Joy A. Thomas (1991). *Elements of Information Theory*. New York, NY: John Wiley & Sons.

Cowie, Anthony P. (1981). The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics*, 2(3):223–235.

Dagan, Ido & Kenneth W. Church (1995). Termight: Identifying and translating technical terminology. In *EACL'95 – Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 34–40. University College, Dublin, Ireland, March 27-31, 1995. San Francisco, CA: Morgan Kaufmann.

Daille, Béatrice (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*, (Ph.D. thesis). Université Paris 7.

Daille, Béatrice (1996). Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans & Philip Resnik (Eds.), *The Balancing Act: Combining Statistical and Symbolic Approaches to Language*, pp. 49–66. Cambridge, MA: MIT Press.

Damerau, Fred J. (1993). Generating and evaluating domain-oriented multi-word terms from text. *Information Processing & Management*, 29(4):433–447.

Dennis, Sally (1965). The construction of a thesaurus automatically from a sample of text. In M. E. Stevens, V. E. Guiliano & L. B. Heilprin (Eds.), *Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation*, Vol. 269, pp. 61–148. Washington, DC: National Buerau of Standards Miscellaneous Publication.

Dudenredaktion (Ed.) (2002). *Redewendungen und sprichwörtliche Redensarten*, Vol. 11. Bibliographisches Institut Mannheim, Dudenverlag.

Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Evert, Stefan (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, (Ph.D. thesis). Universiät Stuttgart.

Evert, Stefan & Brigitte Krenn (2001). Methods for the qualitative evaluation of lexical association measures. In *ACL'01/EACL'01 – Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 188–195. Toulouse, France, July 9-11, 2001. San Francisco, CA: Morgan Kaufmann.

Fano, Robert M. (1961). *Transmission of Information; a statistical theory of communications*. New York, NY: MIT Press.

Firth, John Rupert (1957). *Papers in Linguistics 1934 - 1951*. London, U.K.: Oxford University Press.

Firth, John Rupert (1968). A synopsis of linguistic theory, 1930-1955. In F. R. Palmer (Ed.), *Selected Papers of J. R. Firth 1952-59*, pp. 168–205. Harlow, U.K.: Longman.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

Frantzi, Katerina T., Sophia Ananiadou & Hideki Mima (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Funk, Mark E. & Carolyn Anne Reid (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71(2):176–183.

Gene Ontology Consortium, (2006). The Gene Ontology (GO) project in 2006. *Nucleic Acids Research*, 34:D322–D326.

Grefenstette, Gregory (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Halliday, Michael A. K. (1966). Lexis as linguistic level. In Charles E. Bazell, John C. Catford, Michael A. K. Halliday & R. H. Robbins (Eds.), *In Memory of F. R. Firth*, pp. 148–162. Harlow, U.K.: Longman.

Halliday, Michael A. K. (1969). Categories of the theory of grammar. *Word*, 17(3):241–292.

Halliday, Michael A. K., Angus McIntosh & Peter Strevens (1965). *The Linguistic Science and Language Teaching*. London, U.K.: Longman.

Harris, Zelig (1968). *Mathematical Structures of Languages*. New York, NY: Columbia University Press.

Harris, Zelig (1988). *Language and Information*. New York, NY: Columbia University Press.

Hausmann, Franz Josef (1985). Kollokationen im deutschen wörterbuch. ein beitrag zur theorie des lexikographischen beispiels. In Henning Bergenholtz & Joachim Mugdan (Eds.), *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch*, pp. 118–129. Tübingen, Germany, June, 28-30, 1984.

Hirschman, Lynette & Naomi Sager (1982). Automatic information formatting of a medical sublanguage. In Richard Kittredge & John Lehrberger (Eds.), *Sublanguage: Studies of Language in Restricted Domains*, pp. 27–80. Berlin, New York: Walter de Gruyter.

Hodges, Julia, Shiyun Yie, Ray Reighart & Lois Boggess (1996). An automated system that assists in the generation of document indexes. *Natural Language Engineering*, 2(2):137–160.

Hoffmann, Lothar (1985). *Kommunikationsmittel Fachsprache*. Tübingen: Gunter Narr Verlag.

Jacquemin, Christian (1998). Improving automatic indexing through concept combination and term enrichment. In *COLING/ACL'98 – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, Vol. 2, pp. 595–599. Montréal, Quebec, Canada, August 10-14, 1998. San Francisco, CA: Morgan Kaufmann.

Jacquemin, Christian (2001). *Spotting and Discovering Terms through NLP*. Mass.: MIT Press.

Jacquemin, Christian & Didier Bourigault (2003). Term extraction and automatic indexing. In Ruslan Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*, pp. 599–615. Oxford, UK: Oxford University Press.

Jacquemin, Christian, Judith L. Klavans & Evelyne Tzoukermann (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *ACL/EACL'97 – Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 24–31. Madrid, Spain, 7-12 June 1997. San Francisco, CA: Morgan Kaufmann.

Jacquemin, Christian & Evelyne Tzoukermann (1999). NLP for term variant extraction. In Tomek Strzalkowski (Ed.), *Natural Language Information Retrieval*, pp. 25–74. Boston, MA: Kluwer Academic Publishers.

Jurafsky, Daniel & James A. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

Justeson, John S. & Slava M. Katz (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Kim, Jin-Dong & Jun'chi Tsujii (2006). Corpora and their Their Annotation. In Sophia Ananiadou & John McNaught (Eds.), *Text Mining for Biology and Biomedicine*, pp. 179–211. Norwood, MA: Artech House.

Klein, Dan & Christopher D Manning (2003). Accurate unlexicalized parsing. In *ACL'03 – Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 423–430. Sapporo, Japan, 7-12 July 2003. San Francisco, CA: Morgan Kaufmann.

Krenn, Brigitte (1994). Idioms and support verb constructions. In John Nerbonne, Klaus Netter & Carl Pollars (Eds.), *German in Head-Driven Phrase Structure Grammar*, pp. 365–396. Stanford, CA: CSLI Publications.

Krenn, Brigitte & Stefan Evert (2001). Can we do better than frequency? A case study on extracting pp-verb collocations. In *Proceedings of the ACL Workshop on Collocations*. Toulouse, France.

Landis, J. Richard & Gary G. Koch (1977). The measurement of observer agreement on categorial data. *Biometrics*, 33(1):159–174.

Langendoen, D. Terence (1968). *The London School of Linguistics. A Study of the Linguistic Contributions of B. Malinowski and J. R. Firth.* Cambridge, Mass.: MIT Press.

Leech, Geoffrey (1992). 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Leech, Geoffrey (1993). 100 million words of English. *English Today*, 9(1):9.

Lehmann, Erich L. (1997). *Testing Statistical Hypotheses* (2nd ed.). New York, NY: Springer.

Lehr, Andrea (1996). *Kollokationen und maschinenlesbare Korpora. Ein operationales Analysemodell zum Aufbau lexikalischer Netze.* Tübingen: Niemeyer Verlag.

Lehrberger, John (1982). Automatic translation and the concept of sublanguage. In Richard Kittredge & John Lehrberger (Eds.), *Sublanguage: Studies of Language in Restricted Domains*, pp. 81–106. Berlin, New York: Walter de Gruyter.

Lehrberger, John (1988). Sublanguage analysis. In Ralph Grishmam & Richard Kittredge (Eds.), *Analyzing Language in Restricted Domains*, pp. 19–38. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lezius, Wolfgang, Reinhard Rapp & Manfred Wettler (1998). A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for German. In *COLING/ACL'98 – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, Vol. 2, pp. 743–748. Montréal, Quebec, Canada, August 10-14, 1998. San Francisco, CA: Morgan Kaufmann.

Lin, Dekang (1993). Principle-based parsing without overgeneration. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 112–120. Ohio State University, Colombus, OH, USA, 22-26 June 1993. San Francisco, CA: Morgan Kaufmann.

Lin, Dekang (1994). Principar – an efficient, broad-coverage, principle-based parser. In *COLING'94 – Proceedings of the 15th International Conference on Computational Linguistics*, Vol. 1, pp. 482–488. Kyoto, Japan, August 5–9, 1994. San Francisco, CA: Morgan Kaufmann.

Lin, Dekang (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *ACL/EACL'97 – Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 64–71. Madrid, Spain, 7-12 June 1997. San Francisco, CA: Morgan Kaufmann.

Lin, Dekang (1998a). Automatic retrieval and clustering of similar words. In *COLING/ACL'98 – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, Vol. 2, pp. 768–774. Montréal, Quebec, Canada, August 10-14, 1998. San Francisco, CA: Morgan Kaufmann.

Lin, Dekang (1998b). Extracting collocations from text corpora. In COMPUTERM *'98 – Proceedings of the 1st Workshop on Computational Terminology*, pp. 57–63. Montréal, Québec, Canada, August 15, 1998.

Lin, Dekang (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 317–324. College Park, MD, USA, 20-26 June 1999. San Francisco, CA: Morgan Kaufmann.

Long, Thomas H. & Della Summers (1979). *Longman dictionary of English idioms* (2 ed.). Harlow: Longman.

Macleod, C., S. Chen & J.M. Clifford (1987). Parsing unedited medical narrative. In Naomi Sager, Carol Friedman & M.S. Lyman (Eds.), *Medical Language Processing*, pp. 163–174. Reading, MA: Addison-Wesley.

Mandelbrot, Benoit (1954). Structure formelle des textes et communication. *Word*, 10(1):1–27.

Manning, Christopher D. & Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA; London, U.K.: Bradford Book & MIT Press.

McKeown, Kathleen & Dragomir Radev (2000). Collocations. In Robert Dale, Hermann Moisl & Harold Somers (Eds.), *Handbook of Natural Language Processing*, pp. 507–523. New York, NY: Marcel Dekker.

McNemar, Quinn (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

Mel'čuk, Igor (1995a). *Introduction a la lexicolgie explicative et combinatoire*. Paris, France: Ducolot.

Mel'čuk, Igor (1995b). Phrasemes in language and phraseology in linguistics. In Martin Everaert, Eric-Jan van der Linden, André Schenk & Ron Schreuder (Eds.), *Idioms. Structural and Psychological Perspectives*, pp. 167–232. Lawrence Erlbaum Associates.

Mel'čuk, Igor (1996). Lexical functions: A tool for the description of lexical relations in a lexicon. In Leo Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Proecessing*, pp. 23–54. Amsterdam: John Benjamins.

Mel'čuk, Igor (1998). Collocations and lexical functions. In A. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications*, pp. 23–54. Oxford: Clarendon Press.

Miller, George A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Nenadić, Goran, Sophia Ananiadou & John McNaught (2004). Enhancing automatic term recognition through recognition of variation. In *COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, pp. 604–610. Geneva, Switzerland, August 23-27, 2004. Association for Computational Linguistics.

Pearson, Jennifer (1998). *Terms in Context*. Studies in Corpus Linguistics. Amsterdam: John Benjamins Publishing Company.

Pollard, Carl & Ivan Sag (1994). *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Rijsbergen, C. J. van (1979). *Information Retrieval* (2nd ed.). London, U.K.; Boston, MA: Butterworths.

Rosen, Kenneth H. (1999). *Discrete Marthematics and its Applications* (4th ed.). Boston, MA: McGraw-Hill.

Sachs, Lothar (1984). *Applied Statistics: A Handbook of Techniques* (2nd ed.). New York: Springer.

Sager, Juan (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins Publishing Company.

Sager, Naomi (1982). Syntactic formatting of science information. In Richard Kittredge & John Lehrberger (Eds.), *Sublanguage: Studies of Language in Restricted Domains*, pp. 9–26. Berlin, New York: Walter de Gruyter.

ISO 1087 (1990). *Terminology Vocabulary*. Geneva, Switzerland: International Organization for Standardization.

Umls (2004). *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

Umls (2006). *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

Shannon, Claude E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 623–656.

Shannon, Claude E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64.

Sinclair, John (1966). Beginning the study of lexis. In Charles E. Bazell, John C. Catford, Michael A. K. Halliday & R. H. Robbins (Eds.), *In Memory of F. R. Firth*, pp. xxx–xxx. London: Longman.

Sinclair, John (1991). *Corpus Concondance Collocation*. Oxford: Oxford University Press.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants & Hans Uszkoreit (1997). An annotation scheme for free word order languages. In Paul Jacobs (Ed.), *ANLP 1997 – Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 88–95. Washington, D.C., March 31 - April 3, 1997. San Francisco, CA: Morgan Kaufmann.

Smadja, Frank A. (1989). Lexical co-occurrence: The missing link. *Journal for Literary and Linguistic Computing*, 4(3):163–168.

Smadja, Frank A. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Smadja, Frank A. & Kathleen R. McKeown (1990). Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 252–259. Pittsburgh, Pennsylvania, USA, 6-9 June 1990. Association for Computational Linguistics.

Spears, Richard A. & Betty Kirkpatrick (1993). *NTC's English Idioms Dictionary* (2 ed.). McGraw-Hill.

Stevens, Mary E., Vincent E. Giuliano & Laurence E. Heilprin (Eds.) (1965). *Proceedings of the Symposium on Statistical Association Methods For Mechanized*

Documentation, Washington, DC, 1964. National Bureau of Standards Miscellaneous Documentation.

Stickel, Gerhard (1994). *COSMAS Benutzerhandbuch*. Mannheim: Institut für Deutsche Sprache.

Thielen, Christine & Anne Schiller (1996). Ein kleines und erweitertes Tagset fürs Deutsche. In Helmut Feldweg & Erhard W. Hinrichs (Eds.), *Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Vol. 73, Lexicographica Series Maior, pp. 193–204. Tübingen: Niemeyer.

Trimble, Louis (1985). *English for Science and Technology: A Discourse Approach*. Cambridge: Cambridge University Press.

Tsuruoka, Yoshimasa & Jun'ichi Tsujii (2005). Bidirectional inference with the easiest-first strategy for tagging sequence data. In *HLT/EMNLP'05 – Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 467–474. Vancouver, Canada, October 6-8, 2005. Madison, WI: Omnipress Inc.

van der Wouden, Ton (1997). *Negative Contexts. Collocation, polarity and multiple negation*. Routledge Studies in Germanic Linguistics. London, U.K., New York, NY: Routledge.

Wanner, Leo (1996). Introduction. In Leo Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Proecessing*, pp. 1–22. Amsterdam: John Benjamins.

Wasserman, Larry (2005). *All of Nonparametric Statistics* (1st ed.). New York, NY: Springer.

Wermter, Joachim & Udo Hahn (2004). Collocation extraction based on modifiability statistics. In *COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, Vol. 2, pp. 980–986. Geneva, Switzerland, August 23-27, 2004. Association for Computational Linguistics.

Wermter, Joachim & Udo Hahn (2005a). Effective grading of termhood in biomedical literature. In Charles P. Friedman, Joan Ash & Peter Tarczy-Hornoch (Eds.),

AMIA'05 – Proceedings of the 2005 Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations to Applications to Policy, pp. 71–75. Washington, D.C., October 22-26, 2005. Bethesda, MD: Omnipress.

Wermter, Joachim & Udo Hahn (2005b). Finding new terminology in very large corpora. In Peter Clark & Guus Schreiber (Eds.), *KCAP'05 – Proceedings of the Third International Conference on Knowledge Capture*, pp. 137–144. Banff, Canada, October 2-5, 2005. Association for Computing Machinery.

Wermter, Joachim & Udo Hahn (2005c). Paradigmatic modifiability statistics for the extraction of of complex multi-word terms. In *HLT-EMNLP'05 – Proceedings of the 5th Human Language Technology Conference and 2005 Conference on Empirical Methods in Natural Language Processing*, pp. 843–850. Vancouver, Canada, October 6-8, 2005. Association for Computational Linguistics.

Wermter, Joachim & Udo Hahn (2006). You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. In *COLING-ACL'06 – Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 785–792. Sydney, Australia, July 17-21, 2006. Association for Computational Linguistics.

Wüster, Eugen (1974). General terminology theory – fine line between linguistics, logic, ontology, information science and business sciences. *Linguistics*, 119(1):61–106.

Wüster, Eugen (1979). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Vienna/New York: Springer.

Yates, Frank (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1:217–235.

Zipf, George K. (1935). *The Psychobiology of Language*. New York, NY: Houghton-Mifflin.

Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.