Contents

Acknowledgements

First and foremost, I would like to thank my supervisors Hans Uszkoreit and Harald Trost. I must especially thank Hans who patiently worked through several versions of this thesis, and provided fruitful discussion, criticism and support. No less am I indebted to Harald for his patience and support during the final year of working on the thesis. I also owe special thanks to Gregor Erbach, who read the first draft and provided a number of helpful comments, and to Gerald Passath for commenting on the experimental part of the thesis.

I owe a large debt of gratitude to my former colleagues from SFB 378: to Thorsten Brants for kindly letting me use his part-of-speech tagger TnT, and to Wojciech Skut for the phrase chunker Chunkie; these tools are at the heart of the syntactic processing methods used in preparing the corpora for collocation extraction. Stephan Oepen, my third coinhabitant, was equally helpful in allowing me to use TSDB, the database management software behind the collocation database CDB developed in the thesis. Steffan Corley from Sharp UK kindly allowed me to use Corset, which has been very useful for extracting lexical n-grams from corpora with numeric span as the metric. All of these people not only donated their software but gave help and support in customizing it to fit my needs.

I am also indebted to the Department of Computational Linguistics at Saarland University for kind permission to use the Negra corpus, and likewise I owe my thanks to DFKI Saarbrücken and the Seminar für Sprachwissenschaft at Tübingen University for kind permission to employ the newsgroup corpus built as part of the FLAG project; particular thanks go to Berthold Crysmann from DFKI. I also owe a great debt to Christer Samuelsson who was patient enough to teach me at least some statistics. I would also like to extend my thanks to Martine Grice and Martin Corley for helping with the English translation of the collocation examples. I would also like to thank various colleagues within the Saarbrücken computational linguistics community and colleagues at ÖFAI and IMKAI in Vienna for numerous helpful discussions and support. Particular thanks go to Robert Trappl from IMKAI/ÖFAI for his constant care to keep up a stimulating and rich environment for research on AI and LT. Last but not least I wish to thank my family who made it possible to devote large parts of my time to research.

During the work on the thesis, I was financially supported by: Universität des Saarlandes, FR 8.7 Computerlinguistik; Deutsche Forschungsgemeinschaft (DFG) through SFB-378 “Ressourcenadaptive kognitive Prozesse”; Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI); Austrian Science Fund (FWF) through the project C-Perform; Österreichisches Forschungsinstitut für Artificial Intelligence (ÖFAI).

Chapter 1

Introduction

1.1 Topic, Motivation and Main Goals

The work presented in this thesis is a contribution to the integration of generative aspects of collocations, i.e., lexically determined word combinations within particular syntactic structures,[1] and those characteristics of collocations which cannot be covered by existing grammar theories, such as collocation-specific restrictions in morphosyntax, and in structural and modificational flexibility. A corpus-based approach is for the time being the most promising means to account for these seemingly arbitrary and static aspects of collocations. The situation is perfectly reflected in the two major strands of recent computational linguistics research on collocations, namely the competence grammatical approaches to the representation of collocations, and the work on corpus-based collocation identification which strongly relies on statistical models of word co-occurrances. While the former try to account for the nongenerative bit of collocations by enumerating seemingly important variants, the latter still pay far too little attention to grammatical properties of collocations. Even though the positive effect of employing linguistic information in stochastic collocation identification is widely acknowledged.

The main problem of the two lines of research is that the according complementary aspects are not properly treated, i.e., the grammar theoretical approaches account for the fuzziness of collocations mainly by enumerating variants identified by introspection which, however, is doomed to failure, not least because collocations vary with respect to language domain as well as with respect to personal preferences. The statistical approaches, on the other hand, employ linguistic knowledge, if at all, in a fairly rudimentary and unsystematic way.

In this situation, the thesis aims at bridging the gap by, on the one hand, systematically employing linguistic information throughout the whole process

---

[1] The notion of collocation as used in this work is defined in more detail on page 15ff.

of identifying collocations from corpora, and on the other hand by combining standard grammatical descriptions of collocations with large scale corpus evidence. Whereby the grammatical descriptions allow collocations to be linked to the standard generative rules of grammar, and the corpus data give access to the aspects of collocations which are reflected in language usage, but the underlying grammatical principles are not yet understood. The work is thus conceived as an initial step, a precondition for developing an appropriate theory of collocations. Apart from this, the study has a variety of applications including structural ambiguity resolution in parsing, improvement of the naturalness of lexical selection in generation, the construction of new types of lexica combining abstract linguistic description and corpus evidence, identification and representation of collocations for machine translation, and many more.

Competence versus Performance Aspects of Collocations

Grammar Theory: The current situation in grammar theory can be described as follows: Even though language usage is full of lexically motivated word co-occurrences and restrictions to the full generativity of grammar, grammar theories focus on generative aspects only. Lexical restrictions are rather viewed as syntactic anomalies (cf. [Fleischer, 1982]) than as genuine aspects of grammar. Accordingly, grammar theories are typically not well equipped for dealing with collocations. As a consequence, special treatments for collocations have been suggested, see for instance [van der Linden, 1993] for a Categorial Grammar approach to collocations, [Krenn and Erbach, 1993] or [Riehemann, 1997] for representations in Head-Driven Phrase Structure Grammar (HPSG), only to mention a few examples. A controversial issue is also the classification of collocations as lexical or phrasal phenomena, nevertheless a strict distinction of lexicon and grammar has been abandoned in grammar theories prevalent in computational linguistics such as Categorial Grammar, Lexical Functional Grammar, HPSG, or Tree Adjoining Grammar.[2] A view which is also supported by mentalist theories, e.g. [Bolinger, 1976], where it is argued that there is no strict separation between grammar and lexicon within mental reality. Similarly, in the representation model developed in the thesis, lexical and structural properties of collocations are represented in an integrative way. Moreover, the model is designed such that it allows supporting a uniform representation of competence grammatical information and real-world data automatically identified from text corpora, thus accounting for generative properties of collocations as well as peculiarities of their usage.

---

[2] See [Morrill, 1994], [Bresnan, 1982], [Pollard and Sag, 1994], [Joshi and Schabes, 1991].

Corpus-Based Approaches: Current approaches to collocation identification basically rely on the higher recurrence of collocational word combinations in text corpora compared to lower recurrence of noncollocational word combinations. The use of numeric spans[3] is the prevalent strategy for identifying candidate data from text corpora. The lexical closeness of the word combinations identified is then calculated employing statistical measures computing the relation between joint and marginal probabilities of word combinations.

Occurrence frequency, indeed, is a useful indicator for collocativity. This claim is supported by frequency counts from corpora as well as by psycholinguistic experiments, e.g. [Lapata *et al.*, 1999]. However, the sparse data problem remains in a purely frequency-based approach, i.e., a large number of word combinations that are judged as collocations by humans occur only once in a certain corpus or are missing at all. In addition, there is also a number of highly frequent word combinations in each corpus which are collocational just in terms of occurrence frequency within the particular corpus. Moreover, collocational and noncollocational word combinations do not necessarily differ in their frequency distributions. Thus it follows that a frequency-based approach needs to be combined with linguistically-motivated strategies which is widely agreed on in the literature, but not yet consequently pursued. (Cf. [Manning and Schütze, 1999] for a brief overview of methods for collocation identification.) Syntactic information, if at all, has been either used for postprocessing the statistically determined collocation candidates, or for specification of the set of candidate data from which then collocations are extracted, see for instance [Smadja, 1993], [Breidt, 1993], [Daille *et al.*, 1994]. The notion of numeric span has already been infiltrated with syntactic constraints in work on identification of German support-verb constructions where span size and position of the words are defined by linguistic criteria, see [Breidt, 1993] and [Docherty *et al.*, 1997].

Thus one aim is to investigate how the application of linguistic constraints for selecting candidate data from the extraction corpus can improve the set of collocation candidates being the basis to which models for collocation identification are applied. Moreover it is investigated how statistical techniques and linguistics-based strategies can be combined in the identification models in order to improve collocation identification. In order to do so, a broad empirical study on collocation identification is pursued investigating the feasibility of various models for identifying different types of preposition-noun-verb (PNV) collocations from candidate data constructed according to different morphosyntactic and syntactic constraints.

In the following, the main goals of the thesis are summarized.

---

[3] i.e., collocation partners are selected by means of distance expressed by the number of words in between.

Main Goals of the Study

Systematic access to real-world data is required, in order to gain insights into the nature of the interaction between lexicalization and grammatical generativity, and to exploit knowledge on lexicalization processes for linguistic theory and computational linguistics applications. Accordingly, the study focuses on the following goals:

1) Development and implementation of computational linguistics methods and tools that allow automatically identifying collocations from arbitrary text;

2a) Specification of a representation model for collocations that accounts for linguistic regularities of collocations and actual occurrences in various texts;

2b) Development of the data scheme and construction of a collocation database to store abstract, linguistically motivated specifications of collocations, as well as actually occurring instances identified in real text.

While goal 1) aims at flexible and efficient access to collocations in machine-readable corpora of arbitrary size and domains, and thus is essential for acquisition of the basic material required for further investigations, goals 2) aim at providing instruments for analysis and evaluation of the collocation data. A hybrid approach is pursued for both collocation identification and representation by combining linguistic knowledge and statistical information gained from real-world text. The approach is exemplified on German PP-verb collocations.

1.2 Overview of the Thesis: Hypotheses, Methods and Results

The notion of collocation employed in the thesis is presented is section 1.5. The prerequisites for the work are described in chapter 2. Two kinds of prerequisites are distinguished: (i) the state-of-the-art with respect to techniques for corpus-based collocation identification, and the state-of-the-art for representation models for collocations (section 2.1); (ii) techniques and tools for syntactic preprocessing of large corpora (section 2.2), state-of-the-art statistical models for collocation identification which are amongst others investigated in the thesis, (section 2.3.1), methods of inference statistics applied for testing the significance of the experimental results (section 2.3.2) and in section 2.4.2, the database management system behind the collocation database.

As already indicated by the main goals, the thesis thematically divides into two parts,

1. Strategies, methods and tools for corpus-based collocation identification.
2. Specification of a representation scheme for collocations, and implementation of a collocation database.

Part 1. begins with a discussion of the need for syntactically enriched corpora as a starting point for collocation identification (chapter 3). As very large numbers of data are required, an architecture for automatic syntactic preprocessing of arbitrary text is presented in section 3.2. The characteristics of the thus gained extraction corpus are described in section 3.3, and three classes of preposition-noun-main-verb combinations are identified in section 3.4 which are the reference basis for the empirical studies on collocation identification.

In chapter 4, the feasibility of numeric versus syntactic spans for selection of appropriate collocation candidates from the syntactically enriched extraction corpus is discussed (section 4.2). As the investigations clearly show that syntactically motivated candidate selection is superior to a selection based on numerical spans, a number of candidate sets are examined resulting from applying different syntactic constraints for candidate selection from the extraction corpus (section 4.3.1). In addition, implications of a frequency-based candidate selection are discussed in section 4.3.2. In section 4.4, three kinds of models for collocation identification are presented each of which modeling one of the characteristics employed for defining collocations in the thesis. A variety of experiments is presented in chapter 5 providing an empirical background for judging the feasibility of the models for identifying different types of preposition-noun-verb collocations from different kinds of base data. As the empirical study is the first of its kind, it aims at exploring the ground instead of going into depth for a few cases. Thus a number of experiments are conducted varying the test samples from experiment to experiment. The most important results can be found in sections 5.6 and 5.8.

Part 2. of the thesis (chapter 6) is concerned with defining a representation model and relational database for collocations combining competence-based descriptions and real-world occurrences of collocations. In section 6.2 the competence part of the representation model is described, the example base is presented in section 6.3. The relational model of the collocation database is provided in section 6.4, and example queries are given in section 6.5. Facilities for further exploitation of the database output, and for semi-automatic construction of the database entries are sketched in section 6.6.2.

A final summary and outlook of the thesis is given in chapter 7.

In the following, a more detailed overview of the thesis is presented summarizing the underlying hypotheses, the methods employed and the results of each part.

1.2.1 Corpus-Based Collocation Identification

Linguistics-Driven Identification of Collocation Candidates from Text Corpora

Hypothesis

> Syntactically annotated corpora, in contrast to raw text, allow a more accurate set of collocation candidates to be identified.

Realization

Existing computational linguistics tools for shallow syntactic processing are applied for automatically annotating parts-of-speech and rudimentary syntactic structure to an 8 million word sample of the Frankfurter Rundschau Corpus (German newspaper text). Lexical tuples, the collocation candidates, are retrieved from the syntactically preprocessed extraction corpus according to the following basic requirements: preposition and noun need to be constituents of the same PP, and PP and verb need to co-occur within a sentence. In addition, verbal full forms are reduced to base forms, in order to increase frequency counts of morphosyntactically flexible collocates[4]. The resulting set of lexical tuples is manually inspected for occurrences of true collocations which are used as reference data for testing the feasibility of purely statistical and hybrid models for collocation identification.

Results

The vast majority of PNV-combinations occurs only once in the corpus. Thus a very small percentage of word-combinations in texts can be used for statistical identification of collocations, i.e., 3 % of the preposition-noun-verb (PNV) combinations in the extraction corpus occur 3 times or more, 6 % of this small amount of data occur more than 10 times (occurrence frequency $c > 10$). On the other hand, the effort required for a proper treatment of high frequency word combinations, e.g. $c > 10$, is justifiable, as frequent word combinations cover comparably large portions of running text.

Reducing verbs to their base forms leads to an increase in occurrence frequency, but collocation density among the base form data declines compared to full form data.

Comparing PNV-full and -base form triples has revealed that support-verb constructions (SVC) and figurative expressions are reversely distributed in the

---

[4] For a definition of collocate see page 17.

two samples, i.e., the number of SVCs is higher in full form data, whereas the number of figurative expressions is higher in base form data.

Highly recurrent word combinations are more likely to contain collocations than low ranking data. Thus low frequency thresholds, such as $c < 3$, are inappropriate for statistics-based collocation identification. In general, decreasing thresholds lead to a decline in the density of true collocations among the data.

Two major groupings of lexically determined combinations could be identified from the set of PNV-combinations: combinations where two elements (preposition-noun or verb-preposition) are lexically selected, and combinations where preposition, noun and verb are lexically determined. Only the latter are of interest for the present study. Three groups of PNV-collocations are identified: support-verb constructions, figurative expressions and pseudo-collocations which are collocational simply because of their high occurrence frequency in the particular corpus examined.

Numeric versus Syntactic Spans

Hypothesis

> Syntactic spans are more appropriate for collocation identification than numeric spans.

Realization

Three experiments are pursued on the extraction corpus: potential PN- and PNV-tuples are retrieved (1) from the tokenized text, (2) from the part-of-speech tagged text, (3) from the text annotated with rudimentary syntactic structure. The resulting candidate data are examined with respect to the lexical material covered.

Results

The results clearly show that accessibility of syntactic information is important for increasing the proportion of true collocations among the candidates retrieved from the corpus.

Numeric spans are only appropriate if defined in such a way that collocation-specific linguistic units are covered. Spans of size three or four (with the verb as rightmost element), for instance, are well suited for identifying preposition-noun-verb (PNV) collocations from German verb final constructions. The notion of numeric span, however, needs to be replaced by syntactic span, in order to access the full variety of PP-verb combinations without unnecessarily increasing the number of syntactically inappropriate PNV-combinations.

Models for Collocation Identification

Hypotheses

According to the three defining characteristics of collocations – lexical selection, syntactic rigidity and recurrence – employed in this work, the following hypotheses are specified:

> Collocations are recurrent in language usage, and can thus be extracted from large bodies of text applying statistical association models.
>
> As the collocates of a collocation lexically select for each other, employing collocates as key words will lead to an increase of identification accuracy.
>
> Collocations can be reliably identified employing knowledge on collocation-specific grammatical restrictions.

Realization

Three models for collocation identification are defined:

Model 1: Statistical association measures are applied for modeling recurrence of collocations in corpora. Two kinds of statistical measures are tested: (i) Simple association measures that account for the ratio between joint and marginal probabilities of word occurrences. These are mutual information MI as presented in [Church and Hanks, 1989] and the Dice-coefficient $Dice$. (ii) Models that account for the significance or typicality of the individual data with respect to the sample under investigation. These are relative entropy I and the log-likelihood statistics Lgl introduced in [Dunning, 1993]. For comparison, a mere frequency-based approach $freq$ is pursued.

Model 2: Syntactic rigidity of collocations is accounted for by computing the entropy values of the PPs constituted by specific preposition-noun pairs. This way, an information theoretic measure is employed for modeling grammatical regularities that are distinctive for collocations.

Model 3: A kwic-based strategy is utilized to account for lexical selection between the collocates. The model is based on the assumption that the occurrence of a collocate of a particular collocation triggers the occurrence of the partner collocate(s). While statistical association measures account for this characteristic of collocations by comparing probabilities of joint and marginal word occurrences, the kwic-model is purely lexicon-based, and works without reference to occurrence frequency.

Experiments: A broad variety of experiments are conducted for evaluating the models. In the experiments, the following features are varied:

- the thresholds determining the minimal occurrence frequency required for the PNV-combinations to be in the test sample, i.e., samples containing word combinations with occurrence frequency $c \geq 3$, ≥ 5, ≥ 10 are distinguished;
- the syntactic constraints applied for selection of the candidate data, the models for collocation identification are applied to base form and full form PNV-combinations and to sequences comprising a preposition, a noun, and a past participle;
- the extraction corpora, i.e., the major suite of experiments is conducted on the basis of the subset taken from the Frankfurter Rundschau Corpus; for comparison a sequence of key experiments is replicated on the basis of a corpus of German newsgroup contributions. The corpora have been selected as they strongly differ in domain and style.

The experiments are varied broadly, because at this early stage of research an overview of the performance of the different identification models is required as a precondition for more in-depth investigations to be conducted later.

For a summary of the particular hypothsis to be tested see section 5.2.

Results

A very brief summary of the results is given here, for more details see section 5.6 and 5.8.

It could be confirmed that the statistical association measures differ in their suitability for collocation identification depending on the sample employed and on the type of collocation to be identified. MI and $Dice$ are the best association models for identifying SVCs from highly recurrent full form data. I and Lgl, on the other hand, are equally well suited for identifying SVCs from data containing large portions of medium and low frequency PNV-tuples. MI and $Dice$ are better suited for identifying figurative expressions from base form data, whereas I and Lgl are more appropriate for identifying figurative expressions from full form data.

Frequency is a good identifyer for samples including pseudo-collocations, for samples containing large portions of low frequency data and, with restrictions, for samples of medium frequency data.

Accordingly there is no single best measure for identifying different types of collocations from different samples.

The particular strength of the kwic-based approach lies in its ability to improve the identification accuracy for SVCs when combined with a frequency-based or an entropy-based candidate selection.

PP-entropy is a clear alternative to the association measures for identifying SVCs and figurative expressions from high and medium frequency full form data, but also for identifying SVCs from high frequency base form data, and for identifying figurative expressions from medium frequency base form data.

The results achieved from the newsgroup corpus confirm the general findings from examining the newspaper corpus, even though the two corpora differ significantly. This speaks for the general validity of the results. The differences between the results can in the first place be attributed to the differences in the frequency distributions between the corpora. As there is less lexical variation in the newsgroup corpus than in the newspaper corpus, collocation identification becomes harder even from samples comprising highly recurrent word combinations such as set A. As a consequence, methods that have been appropriate for medium occurrence frequencies with $c \geq 5$ in the newspaper corpus are now well suited for collocation identification from high frequency data extracted from the newsgroup corpus.

1.2.2 A Representation Model and Database for Collocations

Hypotheses

> For the time being, collocations, especially the broad range of partially flexible collocations, cannot be appropriately described by a purely competence-based approach. In other words, theoretical understanding of collocations is still insufficient, and thus a means for controlled investigation of collocations is required.
>
> A database that combines a competence-based description of collocations with real-world data is necessary for systematic investigations into collocations.
>
> Identification of collocations from real-world data and construction of database entries needs to be automated, as a corpus-based approach to collocations is very data-intensive.

Realization

Better insights into the grammatical properties of collocations require access to both abstract linguistic descriptions and real-world occurrences of collocations. In order to achieve this, the following methods and techniques are applied:

Feature-based description of collocations: Each collocation (type) is associated with a set of attribute-value pairs, representing on the one hand general features such as morphosyntactic and syntactic properties, and on the other hand collocation-type-specific features.

Collection of real-world occurrences of collocation instances: Sentences containing collocation instances (tokens) are automatically derived from the extraction corpus and described with respect to general and collocation-specific features.

Representation in a relational database: Abstract linguistic descriptions (competence base) and real-world data (example base) are represented in a relational database. Thus flexible access to all kinds of information represented is possible, and a variety of generalizations over the data can be made which are indispensable preconditions for closer investigations of collocation phenomena.

As far as possible, the database entries are automatically generated from the corpus data. Generalizable information is stored in the competence base, whereas highly varying information is represented by means of the example base. Collocation relevant information that cannot be inferred from the data is manually added to the competence base.

Results

The database currently contains descriptions for approximately 1 000 collocations (467 SVCs and 560 figurative expressions). Each collocation is described by a number of corpus examples (sentences containing the collocation) and by a linguistic description which contains information on syntactic structure as well as a collocation-type-specific analysis. This way, linguistic analysis and actually occurring data complement each other. As the representations are stored in a relational database, different views on the data can be generated, and thus, together with the extraction component, a tool is available that allows for systematic studies of collocations, and their usage or function in text. Moreover, the example base can be used for training statistical models of collocations. Methods for automating the database construction have been developed and appropriate tools have been implemented.

1.3 Contributions of the Thesis

The present study provides computational linguistics methods and tools for collocation identification from arbitrary text, and methods and tools for representing collocations in a relational database integrating competence and performance information. The work differs from existing approaches to collocation identification in systematically utilizing collocation type-specific linguistic information for identification of collocations by exchanging numeric with syntactic spans, by employing entropy to model grammatical rigidity, and by using potential collocates as lexical keys. To the knowledge of the author it is the first attempt employing PP-entropy for the distinction of collocational and noncollocational PNV-combinations. The work is also the first providing experimental results on differences between models for automatic collocation identification depending on factors such as sample size, sample type and collocation class. With respect to collocation representation, the work is the first systematically and in a large scale combining competence-based descriptions of collocations with actual occurrences in text. Another novel feature is the automation of both collocation identification and database construction.

For validation, the following strategies have been applied:

Empirical validation: Two text corpora of different origin and style have been used for testing the models for collocation extraction. The corpora are an 8 million word subset of the Frankfurter Rundschau Corpus and a 10 million word subset of newsgroup contributions. The outcome of the extraction models is compared to a list of manually selected word combinations representing the set of true collocations within the test data.

Statistical significance tests: In order to judge the differences between the models, statistical significance tests have been applied, i.e., the χ^2 test for comparing k independent samples, and its special case, the χ^2 test for comparing two independent samples.

1.4 Applications

Linguistic Theory: The methods and tools presented permit a high degree of flexibility in corpus selection, accessing of arbitrary amounts of data, and automatically combining various levels of description such as standard lexica and competence-based as well as performance-based collocation representations. Thus, for the first time, the preconditions are settled for systematic investigation of a principled approach to collocations. This allows developing a theory where collocations are an integrative part of grammar, hopefully shedding more light on the underlying principles that lead to the grammatical rigidity of collocations as it can be seen on the surface.

Parsing: Lexical collocations are valuable indicators for syntactic structure, and thus they are expected to be useful for parse pruning. PP-attachment which is one of the hard problems in parsing is expected to be improved by employing knowledge on preposition-noun, preposition-verb and preposition-noun-verb collocations.

Generation: Data-driven lexical selection in generation is supported by the automatic access to bodies of collocation realizations grouped according to their occurrence in particular corpora and domains. Which leads to an improvement of the naturalness of the utterances generated.

Computational Lexica: The collocation database is the basis for constructing collocation lexica for analysis and generation. The approach can also be applied for the creation of multi-lingual collocation resources.

Dictionary-Cum-Corpus: The collocation database is a prototype of a dictionary-cum-corpus where the dictionary entries consist of generalized linguistic descriptions of collocation types and collections of corpus data (tokens). The representation is such that corpus evidence and linguistic description model two aspects of a coherent whole.

Machine Translation: As word-to-word translation is not possible for a vast majority of collocations, automatic access of relevant mono-lingual collocations is important. Automatic identification of typical word combinations from monolingual texts is thus a useful precondition for the construction of bi- or multi-lingual language resources. High flexibility in collocation identification and representation is particularly important for creating lexical resources for arbitrary domains. The technology developed for identification and storage of collocations may also be employed for building and enhancing translation memories which relief professional translators from repeatedly translating similar segments of text.

Lexicographic Workbenches: The tools for collocation identification presented in this work are well suited for being incorporated into lexicographic workbenches. Because of the modularity of the architecture, the individual tools can be used independently of each other. The tools allow word combinations to be preselected according to a combination of linguistics- and statistics-based criteria. The work presented constitutes a more elaborate approach to collocations than it is the case for current workbenches where selection of collocation candidates is mainly based on numeric spans. The thus resulting lexical tuples are ordered by frequency or in a few exceptions by employing statistical association measures.

Information Retrieval and Document Identification: The possibility to automatically access common word combinations from arbitrary corpora allows common phrases instead of common words to be used as search keys in information retrieval, i.e., the similarity between user query and document is measured in terms of the document-specific commonness of the phrase(s) used in the query. In such an approach, document similarity is modeled by means of phrase similarity instead of similarity at word level. In this vein, it is expected that pseudo-collocations can be utilized for identification of respective domains or topics. This kind of application, however, deserves closer investigation which is beyond the scope of the present study.

1.5 Collocations

1.5.1 Terminology & Definitions

J. R. Firth's Notion of Collocation

The term **collocation** has been introduced in [Firth, 1957] where "meaning by collocation" is distinguished from "contextual meaning". While the latter is defined as functional relation from the sentence to the situative context, collocations are defined at lexical level in order to account for recurrent, lexically determined co-occurrences of words in real text. Firth states:

> "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words." (p.195)

He exemplifies:

> "One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, its collocation with *night*." (p.196).

Even though Firth clearly recognizes the lexical and contextual character of collocations, for him collocability is a feature of word co-occurrences in particular, actually occurring texts, he considers collocations primarily as phenomena of style. Thus Firth neglects conceptual aspects of collocations irrespective of the fact that his example of the collocability of *dark* and *night* represents a relation between the concepts DARKNESS and NIGHT. Conceptual aspects of lexical collocations, on the contrary, are accounted for in [Lakoff and Johnson, 1981] and [Nunberg *et al.*, 1994]. Firth distinguishes "general" or "usual" collocations from "technical" and "personal" collocations (p.195). While general collocations are persistent over time and part of general language, the latter are restricted to domain-specific or personal use, respectively. In this work, we will be concerned with general and domain-specific collocations.

Terminological Uncertainty

In the literature, a variety of terms and definitions is used to address classes of lexically determined word co-occurrences, such as idioms, phraseological units, multi-word lexemes, non-compositional compounds, light-verb constructions, support-verb constructions etc. Phraseological unit (Ge.: Phraseologismus) is a widely used generic term in the German literature, see for instance [Burger *et al.*, 1982; Fleischer, 1982]. Idiom is the term preferably used in the English literature, see for instance [Bar-Hillel, 1955; Hockett, 1958; Katz and Postal, 1963; Healey, 1968; Makkai, 1972]. Terms like multi-word lexemes [Tschichold, 1997] [Breidt *et al.*, 1996], multi-word expressions [Segond and Tapanainen, 1995] and non-compositional compounds [Melamed, 1997] can be found in recent computational literature. The terms light-verb and support-verb construction address a particular class of verb-object collocations which are described in section 3.4.3.

Idiomaticity

Idiomaticity is a frequently mentioned characteristic of lexicalizations. Idiomaticity usually is defined by semantic noncompositionality, i.e., the meaning of an idiomatic word combination is not a function of the semantics of the individual words, but is associated to the word combination as a whole. Semantic opacity, however, is not sufficient for the definition of collocations as there exists a variety of conventionalized word combinations that range from fully compositional ones like *Hut aufsetzen* ('put on a hat'), *Jacke anziehen* ('put on a jacket') to semantically opaque ones like *ins Gras beissen* ('bite into the grass' literal meaning, 'die' idiomatic meaning). For arguments against conflation of conventionality and noncompositionality see [Nunberg *et al.*, 1994].

Defining Characteristics of Collocations

Lexical selection, syntactic rigidity and recurrence are commonly agreed on characteristics of lexicalized word combinations, even though terminology and definitions may vary. These characteristics are also comparable to criteria for the description of phraseological units established in Russian phraseology, a research tradition which has been influential in the field. For influences on German phraseology see [Fleischer, 1982], where a brief survey of the history of research in phraseology is presented, see p. 10ff.

Lexical Selection Word co-occurrence is determined by lexical rather than by semantic criteria. This feature is comparable to Firth's definition of collocation. As a consequence, the lexically selected words cannot be replaced by other semantically and morphosyntactically equivalent ones. A characteristic which is also addressed by the term lexical stability, cf. [Fleischer, 1982].

Restrictions in Syntactic Generativity

A common property of collocations is that they range from completely fixed to syntactically flexible constructions, cf. section 1.5.3. As already mentioned, syntactic restrictions usually coincide with semantic restrictions and thus are indicators for the degree of lexicalization of a particular word combination. Grammatical restrictions have been (mistakenly) considered as useful for subclassification of collocations, see for instance Helbig's criteria for identification of support-verb constructions, [Helbig, 1979] following [Yuan, 1986]. Such an approach, however, easily leads to a wrong account of lexicalization phenomena, as features indicating the degree of lexicalization and collocation-type specific properties are mixed. On the other hand, knowledge of grammatical restrictions is important, particularly in the case of partially restricted collocations, as each particular word combination is associated with specific restrictions that cannot be inferred from standard rules of grammar and thus need to be stored together with the collocation.

Recurrence Within corpora, the proportion of collocations is larger among highly recurrent word combination than among infrequent ones.

Collocations, Collocates and Collocation Phrases

Collocation, as used in the present study, is a word combination that shows at least one of the previous defining characteristics. In addition, the elements of a collocation need to be syntactically dependent. See for instance the adjective-noun combination *blinder Passagier* in examples (1.1)a. and b. Depending on the scope of the adjective in (1.1)a., *blinder* is either syntactically dependent on both *Mann* and *Passagier* (wide scope) or only dependent on *Mann* (narrow scope). In example (1.1)b. *blinder* is a dependent of *Passagier*. For the word combination a collocational reading ('stowaway') as well as a literal reading ('blind passenger') is available. This is not the case in example (1.1)a. where only literal interpretation is possible, either because there is no syntactic dependency as it is the case with narrow scope, or the collocational interpretation is semantically outruled because of the word order, i.e. proximity of *blinder* and *Mann* in the surface string. *Blinder* here is associated with the reading 'blind'. If the nominal co-ordinates are reordered – *ein blinder Passagier und Mann* – the collocational reading becomes prominent, and wide scope is blocked as the reading ‘blind’ is not available. The behaviour of the collocation in co-ordinated structure can be interpreted as an indicator for the tendency of the collocates of a collocation to be syntactically close.

(1.1)a. ein blinder Mann und Passagier der MS Europa
(a blind man and passenger of MS Europa)
b. ein blinder Passagier der MS Europa
(a blind passenger of MS Europa) literal
(a stowaway of MS Europa) collocational

Collocate The individual lexical elements of a collocation are called collocates. Thus in example (1.1)b. *blinder* and *Passagier* are the collocates of the collocation *blinder Passagier*. Similarly to Firth, mutuality is assumed between the collocates of a collocation. Here the fact is in focus that two or more words co-occur more often than by chance. No distinction is made between the importance of individual collocates for the collocation. Open as well as closed class words can be collocates in the approach pursued.[5] The decision which words are collocates depends on the kind of word combinations investigated. In the case of NPs, adjectives and nouns may be collocates, but also determiners or postnominal prepositions. Nouns and verbs are the relevant collocates in object-verb collocations. However, prepositions in the case of PPs may as well be relevant collocates. Summing up, any word in a certain syntactic structure may be used as a collocate. Combinations of lexical and structural collocates are possible as well, see for instance the word combination *im Zuge* (‘during’) which is obligatorily followed by an NP_{gen} or a PP_{von}. Thus the structural element can be considered as a collocate in a wide sense. In word combinations like *von Mann zu Frau* (from man to woman) it is also the scheme *von* X *zu* Y which is collocational while the nouns can be exchanged.

Collocation Phrase Collocations can be word or phrase level phenomena. In the former case, collocations are comparable to words. In the latter, the collocates constitute a phrase that may either consist of the collocates only or contain additional lexically underspecified material. Examples of collocation phrases containing *blind* and *Passagier* are given in (1.2). The examples show that determination and modification of this particular collocation is flexible. For determination see *der, ein, viele*, for modification see *erste, der MS Europa, mit gefälschten Pässen*. *Blind* syntactically is an attributival modifier of *Passagier*. Syntactic variability of the collocation,

---

[5] Typical open class words or content words are nouns, main verbs and adjectives; closed class words or function words comprise determiners, prepositions, auxiliaries, particles and the like.

however, is restricted. A copula construction, for instance, would lead to the loss of the collocational reading – *der Passagier war blind* (the passenger was blind). Similarly, exchange of the adjectives in example (1.2)c. leads to a loss of the collocational reading – *der blinde erste Passagier* (the blind first passenger).

(1.2)a. der blinde Passagier
(the stowaway)
b. ein blinder Passagier
(a stowaway)
c. der erste blinde Passagier der MS Europa
(the first stowaway of MS Europa)
d. viele blinde Passagiere mit gefälschten Pässen
(many stowaways with faked passports)

Summing up, syntactic regularities and restrictions are fairly reliable indicators for collocativity which can be made use of for automatic collocation identification, provided large bodies of syntactically annotated data are available.

1.5.2 Syntactic Properties

In the following, a number of collocations are described with respect to their morphosyntactic and syntactic properties, information which is relevant for a more fine-grained analysis of collocations which will be possible with the availability of collocation databases like the one described in chapter 6.

As the collocates of a collocation are syntactic dependents, they occur in particular structures like NPs (adjective-noun collocations) or PPs (preposition-noun collocations), at clause level (object- or subject-verb collocations), etc. While phrase level collocations constitute collocation phrases which in some cases have full generative potential, and in other cases are grammatically restricted, word level collocations resemble syntactically complex structures, however, they are lexically fixed, no structural transformations are possible, word order is invariant, and internal modification impossible. Morphological properties and syntactic distribution are comparable to single words.

Adjective- and Adverb-Like Collocations Instances of this class of collocations resemble either adverbs or adjectives. See for instance the examples *nichts desto trotz* ('nonetheless') and *fix und fertig* ('exhausted') in (1.3), where the former can be interpreted as adverb, the latter as adjective. The classification is justified by inflectional differences in prenominal position. While *fix und fertig* functions as noun modifier and inflects like an adjective (ADJINFL), see example a., *nichts desto trotz* modifies adjectives (c.) and does not allow for inflection (b.). Inflection in the adjective-like collocation is realized at the rightmost element (a.), whereas in the case of co-ordinated attributive adjectives, inflection is realized at each adjective like in example d.

(1.3) a. der (fix und fertig-e)$_{adj}$ Mann
(the (exhausted-ADJINFL) man)

b. das (*nichts desto trotz-e)$_{adj}$ Konzert
(the (nonetheless-ADJINFL) concert)

c. das (nichts desto trotz)$_{adv}$ gelungene Konzert
(the (nonetheless) successful concert)

d. seine (fix-e)$_{adj}$ und (einfach-e)$_{adj}$ Vorgehensweise
(his (quick) and (simple) procedure)

While the above examples have been constructed, comparable data can be found in corpora, see for instance examples (1.4) for attributive use, and examples (1.5) for predicative use of *fix und fertig*. The examples are extracted from an 8 million word portion of the Frankfurter Rundschau corpus[6]. Two collocational readings of *fix und fertig* exist, one meaning is 'finished', the other is 'exhausted'. The former is represented by the examples in (1.4) and by (1.5)a. and b., the latter by (1.5)c.

(1.4) a. 32 funkelnagelneue, fix und fertige Wohnungen
('32 brand new apartments ready for moving in')

b. eine fix und fertige Disco
('a disco ready for opening')

(1.5) a. den fix und fertig auf dem Tisch liegenden Kompromiß
('the finally worked out compromise')

b. fix und fertig angerührt
(ready mixed)

c. wir sind alle fix und fertig
(we are all completely shattered)

Word level collocations also undergo word formation processes see for instance examples (1.6) and (1.7) which were found in the Rundschau corpus. In example (1.6) the collocation *fix und fertig* merges with the prefix *fertig* of the verb *fertigstellen* (to complete), thus the collocation becomes a verb prefix. The whole sequence *fix und fertiggestellt* functions like a verb. The process is comparable to prefigation of verbs with adjectives like *schön* – *schönfärben* ('whitewash'), *schief*

---

[6] See page 38 for a description.

– *schieftreten* ('wear down (heels) on one side'), *krank* – *kranklachen* ('laugh one's head off'). The fusion in example (1.7) is even more mannered. Here the nouns *Buch* (book) and *Kruzifix* (crucifix) combine to *Buchzifix* which merges with *fix und fertig*. In addition, *and* is replaced by an ampersand.

(1.6) Allerdings soll diese Verfassung im Sinne des Militärs bereits fix und fertiggestellt sein
('Although the constitution should already be fully worked out in the interests of the military')

(1.7) "Buchzifix & fertig" nennt er sein Objekt – die Bibel – ausgeschnitten in Kreuzesform und daneben im gewohnten Format.
(' "Buchzifix & fertig" does he call his object – the bible – cut into the shape of a cross and also in the usual format.')

In the case of *fix und fertig*, linguistic expectations about syntactic function, i.e., attributival and predicative adjective, and corpus data agree. But this is not always the case. *Klipp und klar*, for instance, can also be classified as adjective, and thus attributival and predicative occurrences are expected in the corpus. The corpus examined, however, contains only predicative data like the one in example (1.5.2). This illustrates, on the one hand, that corpus data are incomplete with respect to the occurrence of linguistic phenomena. On the other hand, these restrictions in occurrence provide valuable information on the usage of linguistic entities in a specific context.

(1.8) Eines steht für die Darmstädter Abteilung des ginstergelben Riesen aber klipp und klar fest.
('But one thing is completely clear for the Darmstadt division of the yellow giant.')

Some other examples of adverb-like collocations are *gut und gern(e)* ('easily', 'at least'), *gang und gäbe (sein)* ('be quite usual') which occurs only as copula construction, *an und für sich* ('in itself'), *mit Fug und Recht* ('rightly'), *zu Recht* ('rightly'), *auf gut Glück* ('trusting to luck'), *von Haus aus* ('actually'), *letzten Endes* ('finally'). Even though these collocations structurally resemble phrases, they are more closely related to words. *Zu Recht*, for instance, has already become a single word – *zurecht*. In the corpus, both variants are found, with 95 occurrences of *zu Recht* and 10 occurrences of *zurecht*.

Preposition-Like Collocations Another class of word-level collocations are fixed preposition-noun sequences that syntactically resemble PPs but function more like prepositions, see for instance *im Lauf(e)* ('during'), *im Zuge* ('during'), *an Hand* ('with the help of'). The combinations are followed by a genitive NP

(NP_{gen}) or pseudo-genitive (PP_{von}). A genitive modifier to the right is characteristic for nouns. With respect to the particular PN-combinations, however, the genitive is obligatory. Moreover *im Lauf(e)* and *im Zuge* can be paraphrased by the preposition *während* (during), and *an Hand* is already in a transition from multi-word unit to single word *anhand*. In the Duden dictionary [Drosdowski et al., 1989], *an Hand* is listed as the main variant while *anhand* figures as newer, but nowadays frequently occurring variant. The dictionary information is confirmed by the corpus data. There have been found 70 instances of *anhand* but only 5 instances of *an Hand* in the 8 million word subcorpus of the Frankfurter Rundschau. For comparison, see *zurecht* (right) and *zu Recht* (rightly). According to Duden, the two variants are distributionally distinct. While the multi-word unit functions as adverb, the single word variant only occurs as separable verb prefix. In contrast to Duden, adjectival occurrences of *zurecht* do occur in the corpus. With respect to these examples, it is not clear whether the distributions are due to errors, be it deviations from the spelling conventions of the Frankfurter Rundschau or the style of individual journalists, or whether the examples represent different stages in the transition from multi-word- to single-word-unit.

From a competence grammatical point of view, *an Hand* and *zu Recht* are incomplete PPs because of missing determination. For both word combinations, only collocational interpretations are available. *Im Lauf(e)* and *im Zuge* on the other hand are syntactically complete. For both examples, also literal interpretations exist. The literal meaning of *im Laufe* is 'while running', the literal meaning of *im Zuge* is 'in the {train, draft}'. In addition, *Laufe* and *Zuge* are archaic strong declension forms which is indicated by the e-suffix. Both incomplete structures and archaic forms are marked constituting a bias towards collocational interpretation. This assumption is supported by the corpus, where 181 collocational *im Laufe*-instances and 25 collocational *im Lauf*-instances occur. The distribution of *im Zuge* (134 collocational instances) and *im Zug* (2 collocational instances) is even more distinct. 100% of the *im Zuge*-instances (134 total), but only two of 11 *im Zug*-instances total in the corpus demand collocational interpretation.

Noun and NP-Like Collocations Typical examples of NP-like collocations are adjective-noun combinations like $\textit{blinder}_{adj}$ $\textit{Passagier}_{noun}$ (blind passenger, 'stowaway') or $\textit{kalter}_{adj}$ $\textit{Krieg}_{noun}$ (cold war). The combinations are lexically determined, but constitute NPs which obey the standard rules of grammar except that adjective and noun need to be adjacent to license the collocational reading, see also the discussion at page 17. Further examples of recurrent adjective noun combinations are $\textit{Rotes}_{adj}$ $\textit{Kreuz}_{noun}$ (Red Cross), $\textit{Wiener}_{adj}$ $\textit{Sängerknaben}_{noun}$ (Vienna choir boys), $\textit{Deutsche}_{adj}$ $\textit{Demokratische}_{adj}$ $\textit{Republik}_{noun}$ (German Democrat Republic) which are semantically compositional but function as proper nouns.

Similarly, the collocation *Hinz und Kunz* (1.9) functions more like a noun than a phrasal projection, even though the word combination structurally corresponds to a co-ordinated NP like *Peter and Mary* except that *Hinz und Kunz* is completely fixed, no reordering (b.), determination (c.) or modification (d.) of the conjuncts is possible without losing the collocational meaning.

(1.9) a. Hinz und Kunz
('every Tom, Dick and Harry')

b. *Kunz und Hinz

c. *der Hinz und der Kunz
(the Hinz and the Kunz)

d. *der kleine Hinz und der große Kunz
(the little Hinz and the tall Kunz)

Another special case of nominal collocations are sequences where the nouns are duplicated like *Tonband nach Tonband* (tape after tape), *Schulter an Schulter* (shoulder to shoulder), *Kopf an Kopf* (neck and neck), *von Ort zu Ort* (from place to place), *von Mann zu Mann* (from man to man), or sequences where the nouns contrast each other like *von Mann zu Frau* (from man to woman). As already mentioned, the patterns 'X nach X', 'X an X', 'von X zu Y' are collocational, the nouns inserted may vary.

Collocations Containing Verbal Collocates A variety of combinations exists, some of which will be introduced in the following.

Modal constructions Here modal and main verb are collocational like in *sich (nicht) lumpen lassen* ('to splash out') where the collocation is constituted by the modal *lassen* and the main verb *lumpen*. A special property of the particular word *lumpen* is that it does not occur outside the combination with *lassen*.

Verb-object combinations like

(1.10)a. ins Gras beißen (into the grass bite, 'bite the dust')

b. übers Ohr hauen ('take somebody for a ride')

c. unter die Lupe nehmen ('take a close look at')

d. zum Vorschein bringen ('bring something to the light')

(1.11)a. des Weges kommen ('to approach')

b. eines Besseren belehren ('put someone right')

c. ein Geständnis ablegen ('make a confession')

d. Lügen strafen ('prove somebody a liar')
e. Anzeige erstatten ('report somebody to the police')

In the above examples, verbs and nouns are collocational. The nouns constitute either PPs (1.10) or NPs (1.11) which syntactically can be interpreted as verb arguments.

Copula constructions Another example of noun-verb collocations are predicatives comprising a copula and a lexicalized NP or PP like *guten Glaubens sein* ('be in good faith'), *guter Dinge sein* ('be in good spirits'), *auf Draht sein* ('be on the ball').

Proverbs In proverbs other than in the above examples, more than one argument is lexically determined, see for instance *Morgenstund hat Gold im Mund* (morning hour has gold in the mouth, 'the early bird catches the worm') or *jeder ist seines Glückes Schmied* (everyone is of his luck smith, 'everyone is the architect of his own future'). Here all arguments are determined. An example where the subject is lexically underspecified is *wissen, wo der Barthel den Most holt* (know where the Barthel the cider fetches, 'know every trick in the book') .

Summing up, morphosyntactic and syntactic properties are useful indicators for collocations, such as

- Structural dependency: as shown in this section the collocates of a collocation are syntactic dependents, thus knowledge of syntactic structure is a precondition for accurate collocation identification.
- Syntactic context: may help to discriminate literal and collocational readings, see for instance *im Lauf, im Zug* where a genitive to the right is a strong indicator for collocational reading.
- Markedness: morphologically or syntactically marked constructions like seemingly incomplete syntactic structure or archaic e-suffix are suitable indicators for collocations, see *im Laufe, im Zuge* for e-suffix and *zu Recht, an Hand* for incomplete syntactic structures.
- Single-word versus multi-word units: single-word occurrences of word combinations indicate word-level collocations, see for instance *zu Recht, zurecht*.
- Syntactic rigidity: is an important indicator for collocations see for instance *Hinz und Kunz, an und für sich, fix und fertig, Kopf an Kopf*. Syntactic rigidity will be more closely discussed in the next section.

1.5.3 Restrictions in Generativity

A number of examples have already been given in the previous section, illustrating that grammatical restrictions are useful indicators for collocations, and thus can be employed for the distinction of collocations from noncollocational word combinations. In this section, examples are presented for collocations with different degrees of grammatical rigidity. The examples also show that similar grammatical restrictions occur at different classes of collocations, and therefore are only restrictedly applicable for distinguishing between individual classes of collocations.

Rigid Word Sequences

Rigid Word Sequences cannot be interrupted, broken into smaller pieces or ordered in different ways without losing their meaning. Their semantics is non-compositional, i.e. the collocation as a whole is assigned a particular meaning. Typical instances are word level collocations such as *hin und wieder* (‘now and again’), *je nachdem* (‘depending on’), *ab und zu* (‘occasionally’), and rigid noun phrases like *Hinz und Kunz* (‘every Tom, Dick and Harry’) or rigid PPs like *auf jeden Fall* (‘in any case’).

Phrasal Templates

Phrasal templates are comparable to rigid word sequences as their word order and lexical material is fixed. In contrast to rigid word sequences, phrasal templates have compositional semantics, and may contain one or more positionally fixed slots that can be filled with lexically flexible material. The term has been used in [Smadja, 1993] where the following example has been given: *The average finished the week with a net loss of* *NUMBER*. Here only the variable *NUMBER* can be flexibly instantiated. In principle, phrasal templates are fully generative, their occurrence as rigid word strings with positionally and semantically fixed but lexically flexible slots, however, is characteristic for domain-specific usage.

Collocations with Syntactically Restricted Complements

Typical examples are verb-noun collocations. Here the phrase containing the nominal collocate functions as syntactic argument of the verb. The nominal collocate is restricted with respect to morphosyntax, syntax, and modification, see for instance idioms such as *jemandem (schöne Augen) machen* (‘make eyes at somebody’), *jemandem (die Leviten) lesen* (‘lecture somebody’), noun-copula constructions like *(guter Dinge) sein* ('be in good spirits'), or support-verb constructions like *(in Frage) kommen* ('be possible'), *(zu Fall) bringen* ('to ruin'), *(ins Rollen) bringen* ('get something going'), *(in Frage) stellen* ('to doubt'). The components surrounded by brackets are morphosyntactically and syntactically rigid, e.g.: change in number would lead to the loss of the collocational reading. Similarly, separating fused preposition and article would require literal interpretation. See for instance *ins Rollen kommen* ('get under way') versus *in* {*das, ein*} *Rollen kommen* (into {the, a} rolling come, 'start rolling'). Internal modification is either impossible as with *in Frage, zu Fall, die Leviten*, rigid as with *guter Dinge*, e.g. *er ist* {*sehr, besonders*} *guter Dinge* (he is in {very, particularly} good spirits) or destroys the collocational reading as in *er macht ihr schöne blaue Augen*. In this case, only compositional interpretation is possible like 'he makes beautiful blue eyes for her'. Usually modification by metalinguistic comments is possible, see for instance *ins sprichwörtliche Rollen bringen* ('to get something going in its proverbial meaning').

Other Restrictions

Apart from syntactic restrictions within NP- or PP-complements, syntactic restrictions also occur with respect to verb transformations. Idioms like *den Löffel abgeben* (the spoon give-away, 'to kick the bucket') or *ins Gras beissen* (into the grass bite, 'to bite the dust') cannot be passivized without losing idiomaticity, although the verbal collocates *abgeben* and *beißen* can be passivized. A similar behaviour is also shown by support-verb constructions. See for instance *die Fassung verlieren* ('to lose composure') where the collocational reading is lost under passivization, and only the literal interpretation is available – *die Fassung ist verloren worden* (the {frame, socket, version, ...} has been lost).

Syntactically Fully Flexible Collocations

Syntactically fully flexible collocations are collocations where the rules of grammar apply without restriction except for lexical selection between the collocates, see for instance the examples of *Frage* and *stellen* in 1.12, where variation in number (singular, plural) and mode (active, passive) is illustrated, as well as pronominalization in relative clause (1.12).c or anaphoric reference (1.12).d. Two other examples of fully flexible collocations are constituted by *Hut* and *aufsetzen*, *Jacke* and *anziehen*.

(1.12)a. eine Frage stellen (singular)
(to pose a question)

b. viele Fragen stellen (plural)
(to pose many questions)

c. die Fragen, die nie gestellt wurden (plural, relative clause, passive)
(the questions which never have been posed)

d. ich habe noch eine Frage, sie zu stellen wäre aber unfair (active, anaphoric reference)
(I still have a question, to pose it would be unfair)

1.5.4 Summary of the Characteristics of Collocations Relevant for the Current Study

The notion of collocation as it is used in the study combines contextual, grammatical and phraseological aspects. The aspect relating to a contextual approach as suggested in [Firth, 1957] is that corpus data are used for collocation extraction, i.e. actual occurrences of words in context are examined. In contrast to Firth, where a whole text is the potential span to contain the collocates, spans in the present study are constrained by syntactic structure. The span is reduced to certain grammatical relations depending on the kind of collocations examined. As the present study focuses on PP-verb collocations, preposition-noun and preposition-noun-verb combinations are looked at, where preposition and noun need to be constituents of a single PP and co-occur with the verb in the same sentence. Thus the maximal span for a PN-combination is a PP, and for a PNV-combination it is a sentence. With respect to a phraseological approach again syntactic aspects are considered, i.e., grammatical restrictions related to the phrases constituted by the potential collocates are used as additional indicators for collocativity.

The approach presented focuses on computational tractability. Thus collocation identification centers around lexical occurrence frequencies and recurrent collocation-type-specific syntactic properties. More precisely, recurrent preposition-noun-verb combinations and recurrent restrictions in grammatical variability of the PP-instances constituted by a particular preposition-noun combination (a potential collocate) will be used as input to statistical models. Association strength between preposition, noun and verb will be calculated, as well as the entropy of the PP-instances constituted by potential preposition-noun collocates. In addition, a kwic-based approach is pursued, accounting for the mutual lexical determination of the collocates of a collocation, in particular, typical support-verbs are employed for differentiating between support-verb constructions and other kinds of PNV-combinations. Similar to Firth, habituality of a collocation depends on its occurrence frequency. As collocations are derived from corpora, statements on the habituality of a word combination clearly depend on the text base under investigation.

Chapter 2

Prerequisites

In this chapter, first of all the scientific background of the current study is given, including

1. a survey of computational approaches to corpus-based collocation identification (section 2.1.1), and
2. a discussion of representation models for collocations, as well as a presentation of linguistic databases related to the collocation database developed in this work (section 2.1.2).

In the remaining sections, a brief motivation and description of the techniques and tools is given which are applied in the study:

1. A short introduction to the Markov Model technology employed in the part-of-speech tagger and phrase chunker is presented. These tools are at the heart of the preprocessing component described in section 3.2. The treebank applied for training the tools is described in section 2.2.2.
2. Section 2.3 describes, on the one hand, the statistics employed for identifying collocations from a set of candidate word combinations (section 2.3.1). On the other hand, the statistics are presented which are applied for testing the significance of the differences between the identification models (section 2.3.2).
3. In section 2.4, a brief introduction to the concept of a relational database is given (section 2.4.1), and it is motivated why the database management system TSDB has been chosen as core engine of CDB, the collocation database developed in the thesis.

2.1 State of the Art

2.1.1 Techniques for Corpus-Based Collocation Extraction

There is an increasing interest in automatic retrieval of collocations from text corpora, because accessing an arbitrary number of real-world collocations from various domains leads to better insights into the phenomenon, especially with respect to actually occurring syntactic variation of collocations, common modification, typicality of certain lexical collocations for particular domains, etc. Large machine-readable text corpora are available, and processing of huge bodies of text has become feasible as appropriate processing methods and tools have been developed during the last few years, and from an economic point of view, memory cost is negligible. Thus corpus-based collocation identification and retrieval is becoming an important factor towards a more appropriate theory of collocations, which for the time being is still lacking. A more appropriate theory of collocations is desired for a wide range of computational linguistics applications, such as machine translation and machine aided translation, natural language generation, information retrieval and topic identification, sublanguage applications, dictionary construction for computational linguistics applications and in lexicography, second language learning, etc.

Statistics-Based Approaches

Collocations are identified by the frequency of word co-occurrences in corpora. Basically, word n-grams (mostly bi-grams) are collected from varying spans. [Smadja, 1993] for instance looks for a collocate within a span of five words to the left and to the right of a word in English. [Breidt, 1993] reports on an optimal span of two words to the left of the verb for identification of German noun-verb collocations from untagged text.[1] The size of the n-grams is varied as well. [Church and Hanks, 1989] consider only bi-grams. [Smadja, 1993] uses statistically significant bi-grams as basis for extraction of larger n-grams. [Frantzi and Ananiadou, 1996] consider n-grams up to $n = 10$, [Ikehara *et al.*, 1996] look at n-grams of arbitrary length. N-grams consist either of sequences of adjacent words (see for instance [Frantzi and Ananiadou, 1996], [Ikehara *et al.*, 1996], [Shimohata *et al.*, 1997]), or of word tuples selected from certain span sizes, cf. [Smadja, 1993], [Breidt, 1993]

As relative n-gram frequency is only a coarse indicator for collocations, the lexical closeness between words is measured. A frequently applied measure is

---

[1] Note while identification of the initial material for candidate collocations is based on numeric spans in [Smadja, 1993] and [Breidt, 1993], their approaches make use of linguistic information for further reduction of the collocation candidates.

so called mutual information MI. In most cases, MI addresses a logarithmic ratio between the probabilities of joint and marginal word co-occurrences. Thus MI differs from the information theoretic measure called mutual information which determines the relative entropy between two probability distributions. Mutual information has been proposed for bi-grams, see for instance [Church and Hanks, 1989], [Smadja *et al.*, 1996], trigrams [Kim and Cho, 1993], or an arbitrary number of n [Magerman and Marcus, 1990]. Even though MI strongly overestimates with respect to low frequencies, the measure is quite persistent in the literature on corpus-based collocation or term identification. A proposal for an alternative measure is given in [Dunning, 1993], where a log-likelihood statistics is proposed. Another alternative to MI is presented in [Smadja *et al.*, 1996], the Dice coefficient. The measure, however, shows similar problems as MI. While log-likelihood already accounts for the significance of the data, MI and $Dice$ do not. Thus additional significance tests need to be performed to indicate whether the difference between the occurrence of a word combination and the occurrences of the individual words is significant. The most commonly applied strategies are calculation of z- and t-scores.[2] MI, $Dice$ and log-likelihood statistics will be discussed in more detail in section 2.3.1.

Linguistics-Based and Hybrid Approaches

The statistics-based approaches proposed for collocation extraction typically start with little linguistic information. They usually operate on n-grams over part-of-speech tagged word forms, see for instance [Smadja, 1993], [Frantzi and Ananiadou, 1996], [Haruno *et al.*, 1996], [Docherty *et al.*, 1997]. In [Breidt, 1993], unannotated text is used because of a lack of German part-of-speech tagged text at the time of the research. In a number of approaches, different kinds of linguistic information are subsequently used to reduce the number of false collocation candidates. In the following, three approaches will be briefly discussed that make use of linguistic properties of collocations, that is: the work presented in [Smadja, 1993], because it is the first extensive computational linguistics approach to collocation identification (the approach is designed for English); the approaches described in [Breidt, 1993] and [Docherty *et al.*, 1997], because they provide methods for corpus-based identification of German noun-verb collocations, and thus are directly related to the work presented in the present study. While [Smadja, 1993] and [Breidt, 1993] make use of both statistical models and linguistic information, [Docherty *et al.*, 1997] pursue a purely linguistics-based approach, where the resulting word combinations are sorted according to frequency.

---

[2] For a description of the z- and t-distribution, see any standard book on test statistics, for instance [Bortz, 1985].

In [Smadja, 1993] bi-gram collocation candidates are first extracted from the corpus employing statistics-based methods. Then all sentences containing candidate collocations are retrieved from a part-of-speech tagged corpus. In the next step, syntactic relations are added to the part-of-speech tagged sentences, in order to distinguish subject-verb or object-verb collocations. Candidates with inappropriate syntactic structure are discarded. This way precision of the extraction component is increased, and the number of data for final hand-selection is reduced. A similar approach is applied for translating collocations, cf. [Smadja *et al.*, 1996].

Approaches for identification of noun-verb (NV) collocations, mainly support-verb constructions, from German text corpora are presented in [Breidt, 1993][3] and [Docherty *et al.*, 1997]. [Breidt, 1993] is a feasibility study on identification of NV-collocations from corpora. [Docherty *et al.*, 1997] make use of insights from [Breidt, 1993] for a corpus-based dictionary update. In order to compensate the lack of part-of-speech tagged and syntactically annotated corpora, Breidt makes use of typographic and lexical information as well as word order regularities, i.e., nouns are identified by an initial capital letter, only infinitive forms of pre-specified verbs are searched for which are typical support-verbs and which occur at the right periphery of a sentence. The related nouns are looked up within a span of two to five words to the left of the verb, whereby specification of the span size has been influenced by the literature on collocation identification from English text, cf. [Smadja, 1993]. Similarly [Docherty *et al.*, 1997] extract NV-collocations from sentences with a verb complex at the right sentence boundary which is immediately preceded by a noun complex. The corpus is part-of-speech tagged and lemmatized. While Breidt applies association statistics (MI as defined in [Church and Hanks, 1989] using t-scores to distinguish significant from insignificant word combinations), Doherty et al. strictly rely on linguistics-based corpus queries and frequency counts. Employing appropriate corpus queries, the authors account for reflexivity and nonreflexivity of the main verb, morphosyntactic properties of the noun group like occurrence of a preposition, accusative and dative NPs; they also make use of determination, adjectival modification and genitives or PPs to the right of the head noun. Breidt, on the other hand, experiments with span size, lemmatization of the verb, variation of corpus size, and introduction of syntactic relations which were manually added. All in all, the work of Breidt can be viewed as a basis for identification of noun-verb collocations from German.

Conclusions from Breidt are:

- A part-of-speech tagged corpus is the basic requirement for identification

---

[3] A revised version of the article is available from cmp-lg/9603006 (`http://xxx.lanl.gov/find/cmp-lg/1/Breidt/0/1/0/all/1/0`).

of collocation candidates if no verbal or adjectival keys are given.

- In an unparsed corpus, identification of the noun within a span of two to the left of an infinitive or past participle leads to the best accuracy[4] results.
- Lemmatization of the verb is not useful with unparsed corpora as in this case a gain in recall[5] is paired with a loss in precision.
- Increase of corpus size leads to improvement of recall but to a decline in precision which however is not dramatical.
- Raising the cut-off threshold for occurrence frequency from 3 to 5 improves precision, but leads to a serious decline in recall.
- Access to syntactic relations drastically raises precision. A finding which has been also made in Smadja.

A major drawback of both approaches to German is that only a very restricted set of collocations is accessed, namely verb-object collocations, where verb complex and object are adjacent in the surface string. This leads on the one hand to restrictions in occurrence frequency, and on the other hand to an overproportional number of SVCs among the data, whereas other noun-verb collocations are underrepresented. Huge corpora need to be processed in order to compensate for the low occurrence frequencies.

The main disadvantage of the approach described in [Smadja, 1993] is that in the first step a number of syntactically false candidate collocations are specified which need to be discarded in a second step. Furthermore, a span of five words to the right and to the left of a lexical key is not optimal for identification of German verb-object collocations, as has been shown in [Breidt, 1993]. In general, due to word order variation in German, numeric spans are inappropriate to cover the collocates of syntactically flexible collocations. In order to achieve high recall of collocation realizations, the span needs to be enlarged which also leads to an increase of noise in the set of collocation candidates. On the other hand, narrowing the span size allows to reduce the number of noisy data, but also leads to a decline in recall.

In contrast to the approaches described, the approach presented in the current work leads to a more accurate access of collocation data from the beginning, as the extraction corpus is part-of-speech tagged and annotated with basic syntactic structure. Retrieval of NV-combinations is not restricted to verb final

---

[4] Accuracy or precision is the ratio between actually true collocations and the sum of collocations classified by the system as true collocations.

[5] Recall is the number of true collocations identified by the system.

constructions, and no adjacency requirements for noun and verb are given. Collocation data are accessed within arbitrary structures. The only requirement is that noun and main verb co-occur within a sentence. Thus verb second data, and data where the noun is head of a nominal projection and the verb occurs in a dependent relative clause can also be accessed. This degree of flexibility, on the other hand, leads to an increase of noncollocational combinations among the candidate data, a large number of which is expected to fall below the co-occurrence threshold determined for the data under consideration. Syntactically flexible data may, to a certain extent, as well be accessed by means of corpus queries based on a regular language as it is used in [Docherty *et al.*, 1997]. Another difference between the present study and [Breidt, 1993] or [Docherty *et al.*, 1997] is that prepositions are also considered as collocates, and thus the candidate set consists of preposition-noun-verb triples instead of noun-verb combinations. While [Docherty *et al.*, 1997] do not apply statistical measures at all, and [Breidt, 1993] restricts herself to MI, different kinds of statistical measures are examined in this work with respect to their suitability for the identification of particular classes of collocations.

2.1.2 Representation Models for Collocations

In phraseological dictionaries or databases, collocations (usually termed multi-word-units MWUs) are typically described at morphological and syntactic level. In the case of partially (in)variable collocations, both variable and invariable aspects are stated explicitly, see for instance [Keil, 1997], [Tschichold and Hacken, 1998], [Segond and Tapanainen, 1995], [Breidt *et al.*, 1996]. In all approaches, a hand crafted local grammar is specified for each collocation (MWU) representing location, morphological and syntactic properties of the components, position and type of external modifications, and permissible syntactic transformations. A serious drawback of this kind of approach is that explicit descriptions of collocations do not meet the tendency of collocations to vary with respect to domain and speaker. Thus these approaches are most likely to over- or undergenerate when used in analysis.

An attempt to overcome these shortcomings is presented in [Dufour, 1998]. Dictionary entries of MWUs are represented by linguistic descriptions containing features which are associated with empirically motivated weights. The dictionary representations are matched against representations of parsed sentences. Based on the weighted features, matching dictionary entries are presented to the user in order of closeness of match. The approach is a first step in dealing with frequency-based aspects of collocations. Using natural numbers as weights, however, is a drawback as they are hard to interpret, and the tools of statistics are not applicable. In addition, assignment of weights is fairly arbitrary as it is most likely to reflect the intuitions of the human annotator instead of the facts of language usage. In this case, statistical learning techniques lead to more appropriate results, and a database as described in chapter 6 provides the training data, i.e. large bodies of real-world collocations enriched with competence-based linguistic descriptions.

An attempt for a syntactic as well as a semantic and pragmatic description of idioms is made in [Keil, 1997]. Idioms are distinguished with respect to non-compositionality and figurativity. Semantic structure is represented by means of predicate-argument structure and theta roles. Semantic features like HUMAN, or ABSTRACT, etc. are assigned to parts of idioms. Synonyms and antonyms are specified, if possible. At pragmatic level connotations of the idiom are stated. The merits of the work are the it exceeds a purely syntactic view on collocations. Apart from that, the work classifies as a standard competence-based approach where all information is designed by the linguist, which makes it impossible to account for the subtle variations in the usage of collocations.

In the present study, an alternative approach to the description of collocations is made. The descriptions combine competence-grammatical knowledge and realizations of collocations derived from corpora. Each collocation is associated with

- an abstract, potentially over-generating competence-grammatical description, and
- a corpus consisting of real-world occurrences of the particular collocations, where real-world data and competence-based descriptions are linked.

On the one hand, the competence-grammatical description allows specifying basic linguistic information like parts-of-speech, inflectional features, and syntactic structure, and collocation-specific information like argument structure in the case of support-verb-constructions or Aktionsart in the case of support verbs, as well as pragmatic information, information on connotations and the such. On the other hand, the linguistic abstractions are supported but also relativized by the data collections extracted from real-world corpora. Thus information is provided on the usage of a particular collocation in certain contexts. Storage of the representations in a relational database is a powerful means for theoretical investigations of collocations as the database allows for flexible views on the data, and as already mentioned, the linguistically refined corpus data are a valuable training material for statistical learners.

The work presented in chapter 6 relates to work on linguistic databases as developed in the projects TSNLP (http://cl-www.dfki.uni-sb.de/tsnlp/) [Oepen *et al.*, 1998], and DiET (http://www.dfki.de/pas/f2w.cgi?ltp/diet-g) [Netter et al., 1998]. TSNLP and its successor DiET provide reference data for evaluation of natural language processing systems. While TSNLP is a suite of test items constructed by linguists for the evaluation of syntactic processors, DiET is designed for testing a broader range of applications. In DiET, constructed test items are combined with application specific corpora. Corpus data are amongst others used for harmonizing the vocabulary of the test suite. The frequency of a certain phenomenon in a particular corpus is interpreted as the relevance of the phenomenon for the particular application. Similarly, corpus data in the collocation database created in this work are used to determine the importance and particular usage of a certain collocation within a specific corpus or domain. TSDB – the database constructed in the TSNLP project – is used as database management system for the collocation database.

The work presented also relates to lexicographic workbenches like the IMS Corpus Workbench developed at the University of Stuttgart, [6] Qwick developed by Oliver Mason and John Sinclair at Birmingham University,[7] or System Quirk developed at the University of Surrey.[8] All three systems implement a kwic-based[9] approach, i.e., a search string (key word) needs to be specified by the user, and the system identifies lines from a corpus containing the key and n words to the left and/or right. Search patterns are specified by means of regular expressions. Cut-off thresholds for co-occurrence frequencies and span sizes can be defined by the user. Qwick, in contrast to the other workbenches, also allows for statistical evaluation of collocations by providing calculation facilities for a number of association measures.

While the above workbenches rely on kwic-based collocation identification, the current work offers more flexibility as potential collocations are identified employing various aspects of collocations such as lexical co-occurrence frequency, grammatical rigidity and lexical keys. In addition, the notion of collocation is restricted to syntactically meaningful units like PP or PP-verb combination. This is possible because of large scale syntactic preprocessing. For candidate selection, on the one hand, existing tools such as Corset and Gsearch [Keller *et al.*, 1999] can be used. On the other hand, a tool has been implemented which is designed for collecting the specific data (cf. section 3.3.1) required in this study.

Corset and Gsearch are parameterizable with respect to corpus and search fields, and thus allow for search on texts associated with arbitrary tagsets. While Corset enables search over n-grams applying numerical spans, its successor Gsearch allows for specification of context-free grammars whereby the terminals can be expressed by means of regular expressions. Corset has been

---

[6] [Christ, 1994], http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/index.html

[7] http://www.clg.bham.ac.uk/QWICK/doc/

[8] http://www.mcs.surrey.ac.uk/SystemQ/

[9] Kwic means key word in context

used for retrieval of the data presented in sections 4.2.2 and 4.2.2. It is envisaged to reformulate the grammar used for extraction of PP-verb combinations in Gsearch, as the tool is particularly optimized for search in very large corpora which is an important feature for corpus-based retrieval of collocations. Gsearch has not been used from the beginning, as it has not been available at the time of the study. It is also not yet clear, whether Gearch is applicable to the task at hand without changes to the program.

2.2 Corpus Tools and Training Data

2.2.1 Markov Model Technology for Shallow and Robust Syntactic Processing

The need for processing real-world language data has increased with the development of computational linguistics applications. As a consequence, development and implementation of efficient and robust processing techniques has become an important area of research in computational linguistics. Robustness of processing is essential for handling incomplete and partially incorrect input as well as for dealing with unknown words. Robustness is a particular characteristic of statistics-based approaches as statistical models inherently do not distinguish between correct or incorrect input but between more or less probable one. This is a major advantage for processing free input.

Computational linguistics approaches to parsing[10] have for a long time relied on correct and complete input, which made their application to real language impossible. Processing efficiency, on the contrary, results from low level grammatical analysis, i.e. only partial linguistic information is used, such as word level syntactic category (part-of-speech), and instead of full syntactic structure, only syntactic chunks are built such as NP, PP, ADJP. Lexical and semantic information is usually omitted.[11] Due to the reduced amount of linguistic information available, and the absence of phrasal attachments, a crucial amount of ambiguity is eliminated. The price payed for processing efficiency and robustness is lack of deep analysis. However, availability of partial (shallow) information is valuable for a variety of tasks. Part-of-speech tagging for instance is a useful preprocessing step for parsing, because it disambiguates word level syntactic categories. Shallow parsing like phrase chunking is sufficient for identification of NP and PP structures, which are of interest for identification of verb-noun

---

[10] See for instance [Uszkoreit *et al.*, 1994], [Flickinger *et al.*, 1998], [Müller, 1999] for examples of state-of-the-art grammar and parsing systems with high coverage.

[11] Recent attempts to include lexical information into stochastic parsing are the lexical dependency parsers proposed in [Eisner, 1996] and [Collins, 1997]

collocations.

Both, tagger and chunker applied for preprocessing the extraction corpus are based on Hidden Markov Models HMMs and Viterbi search, a method of dynamic programming, techniques that are widely used in statistical speech and language processing. A standard tutorial on HMMs is [Rabiner, 1989]. HMMs and Viterbi search are a standard approach to part-of-speech tagging. Basics of stochastic taggers and parsers are described in [Krenn and Samuelsson, 1996] where also an extended list of literature is given.

Markov Models MMs basically are finite state automata. They consist of

- a finite set of states $\Omega = s_i, \cdots, s_n$;
- a signal alphabet $\Sigma = \sigma_1, \cdots, \sigma_m$,
- a $n \times n$ state transition matrix $\mathbf{P} = [p_{ij}]$ where a transition $p_{ij} = P(s_j|s_i)$ is the conditional probability of state s_j given state s_i, and $\sum_{j=1}^{n} p_{ij} = 1$;
- a $n \times m$ signal matrix $\mathbf{A} = [a_{ij}]$ where $a_{ij} = P(\sigma_j|s_i)$ is the conditional probability that signal σ_j is emitted at state s_i, and $\sum_{j=1}^{m} a_{ij} = 1$;
- an initial vector $\mathbf{v} = [v_i, \cdots, v_n]$ where $v_i = P(s_i)$ is the probability to be in state s_i;

with $s_i = q_t$ and $s_j = q_{t+1}$ i.e. s_i is the actual state q at time t and s_j is the actual state q at time $t+1$. There are two particular states, the start and the end state. A particular characteristic of Markov processes is the Markov property, i.e. the current state s_k only depends on the previous state $s_j = q_{t-1}$. A MM of this kind is called first order model. In analogy, a model where the probability of state s_k is conditioned on the two previous states $s_j = q_{t-1}$ and $s_i = q_{t-2}$ is called trigram- or second order model. In a Hidden Markov model HMM only the emissions can be observed while the states remain unseen.

In the case of part-of-speech tagging, the tags t_i, where $i = 1, \cdots, n$, are the states and the words w_j, where $w = 1, \cdots, m$, are the signals emitted. The transitions are the probabilities that a particular tag is followed by another particular tag or in terms of a trigram model by two particular tags, i.e.

$$P(t_i) = P(t_i|t_{i-2}, t_{i-1})$$

Here the probability of tag t_i is defined by the conditional probability that t_i occurs after t_{i-2} and t_{i-1}. $P(t_i|t_{i-2}, t_{i-1})$ is also called context probability. In addition to context probabilities, lexical probabilities are defined. A lexical probability is the probability that a particular word w_j occurs given a certain tag t_i, i.e.

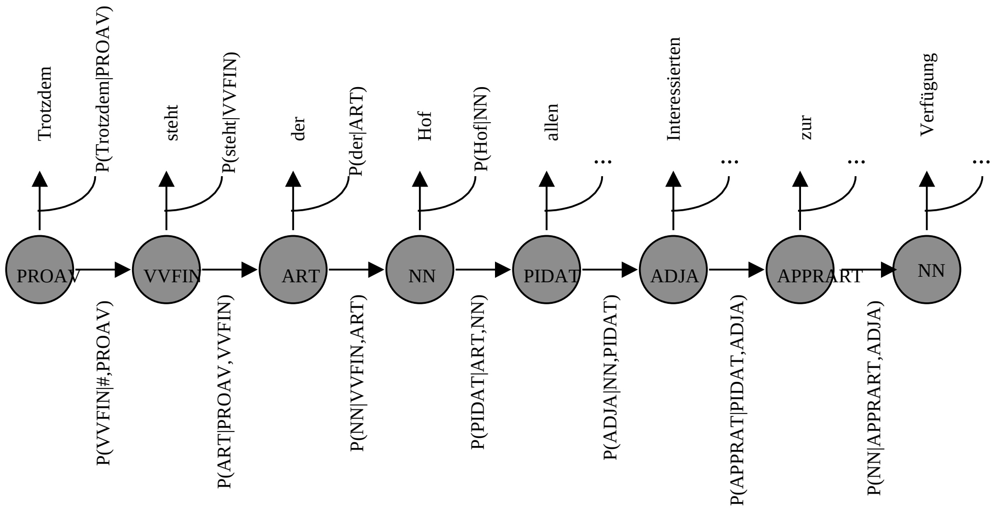


Figure 2.1: Part-of-speech tagging: second order HMM for the sequence *Trotzdem steht der Hof allen Interessierten zur Verfügung.*

$$P(w_j) = P(w_j|t_i)$$

Figure 2.1 illustrates a second order HMM for the sentence (2.4). Start and end state have been omitted in the picture. The context probability assigned to the transition from the start state to the state representing the part-of-speech PROAV can be written in analogy to the sentence internal transitions as $P(PROAV|\#,\#)$ which is the conditional probability that the tag PROAV occurs at the beginning of a sentence. Because of the trigram model, information on the beginning (#) is still available in the condition of the following context probability $P(VVFIN|\#, PROAV)$. Similarly the transition from the prefinal to the end state is $P(\#|APPRART, NN)$. $P(Trotzdem|PROAV)$ is an example for the lexical probability that the form *Trotzdem* occurs given the part-of-speech PROAV.[12]

In order to find the best tag sequence for a given word sequence the following model is calculated.

$$argmax_t \prod_{i=1}^{n} P(t_i|t_{i-2}, t_{i-1})P(w_j|t_i)$$

---

[12] PROAV stands for pronominal adverb which is a proform replacing a PP. VVFIN stands for finite main verb, ART for article, NN for noun, PIDAT is an indefinite pronoun that functions as a determiner, ADJA is an attributive adjective, APPRART stands for a fusion of preposition and article. The full part-of-speech tagset is described in [Thielen and Schiller, 1995].

i.e., the product of context probabilities and lexical probabilities is calculated for each word i, and the model is maximized over the part-of-speech tags t ($argmax_t$). The Viterbi algorithm is a widely used technique to find the single best state (tag) sequence for a given sequence of observations (words) in time complexity $O(n^2T)$. The following set of variables needs to be established.

$$\delta_t(i) = max_{s_{i_1},\cdots,s_{i_{t-1}}} P(s_{i_1}, \cdots, s_{i_{t-1}}, s_{i_t}; \sigma_{j_1}, \cdots, \sigma_{j_t})$$

This is the joint probability of the best sequence of states from time t_1 to time t_t and the sequence of observations $\sigma_{j_1}, \cdots, \sigma_{j_t}$ from time t_1 to time t_t. The variables $\delta_t(i)$ can be defined recursively as follows with the initial equation (2.1) representing the initial states v_i and its related emissions a_{ik_i}, and the recursion (2.2) where the most likely state and emission sequence up to time $t-1$ $\delta_{t-1}(i)$ is combined with the most likely current transition $max_i p_{ij}$.

$$\delta_1(i) = v_i \cdot a_{ik_i} \tag{2.1}$$

$$\delta_t(j) = [max_i \delta_{t-1}(i) \cdot p_{ij}] \cdot a_{jk_t} \tag{2.2}$$

For implementation, a trellis structure is suitable, i.e. each state s_i at time t_t is represented as a node in a lattice where for each state s_i the predecessor states s_{i-1} and the successor states s_{i+1} are represented.

2.2.2 Negra – A Syntactically Annotated German Newspaper Corpus

For training of the tools used for syntactic preprocessing, a syntactically annotated corpus of German newspaper text is applied. The text is taken from the Frankfurter Rundschau (FR) Corpus which is part of the ECI Multilingual Corpus 1 distributed by ELSNET.[13] At the time of this work, approximately 12 000 sentences from FR have been available structurally annotated, and hand corrected.[14] Annotation has been carried out under the projects LINC and NEGRA at the University of the Saarland. The 12 000 sentence corpus was a reasonable basis for training a stochastic part-of-speech tagger and just large enough to train a phrase chunker.

The sentences are annotated with parts-of-speech, phrasal category (node labels), grammatical function (edge labels) and syntactic structure. Structure

---

[13] ECI European Corpus Initiative, ELSNET European Network in Language and Speech

[14] In the meanwhile the corpus has increased to 20 000 sentences. The corpus is available free of charge for noncommercial purposes. For information see `http://www.coli.uni-sb.de/sfb378/negra-corpus/`.

is represented by unordered trees with crossing branches to represent non-local dependencies. For illustration see the analysis of sentence (2.3) in figure 2.2.[15]

(2.3) Das schmucke Aushängeschild löst mehr Fragen aus, als es Antworten gibt.
the smart advertisement causes more questions PRE than it answers gives
'the smart advertisement asks more questions than it answers'

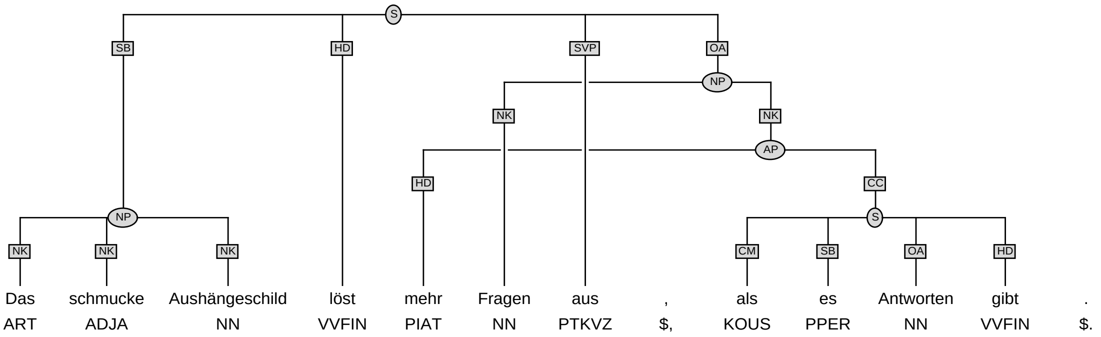


Figure 2.2: Syntactic information as annotated in the Negra Treebank

The sentence (S) has a verbal head (HD) with separable prefix (SVP) and two NPs functioning as subject (SB) and direct object (OA), respectively. The object NP consist of a noun (NN) modified by an adjective phrase (AP) comprising the comparative *mehr* and the comparation clause (CC) *als es Antworten gibt*. Non-local dependencies are indicated by crossing branches, see the separable verb prefix *aus* which occurs in the surface string among the elements of the object NP, see also the AP *mehr als es Antworten gibt* which is interrupted by the noun *Fragen*. The example also shows another characteristic feature of the annotation scheme, namely the representation of NPs. NPs consist of nominal kernel elements (NK) and left and right modifiers. The nominal kernel covers all lexical elements or phrases constituted by lexical elements that can constitute an NP on their own. These are articles (ART), attributive adjectives (ADJA), and nouns (NN). In contrast to standard approaches, no head is specified for NPs and PPs, the latter are analyzed as NPs with additional morphological marker (AC). Thus, commitment to DP- or NP-analysis is avoided, because neither is justified from a descriptive point of view.

Collocations are not specifically marked in the annotation scheme. See for instance the treebank representation (figure 2.3) for the example sentence (2.4) which contains the support-verb construction *steht zur Verfügung* (stands at the disposal, 'is at the disposal'). The verbal collocate *steht* is annotated like any other head of a finite sentence. Similarly, the nominal collocate *Verfügung* is

---

[15] For a detailed discussion of the annotation scheme see [Skut *et al.*, 1997].

annotated according to its syntactic function as element of the nominal kernel (NK) of a PP. The example shows another peculiarity of the annotation scheme. PPs a are underspecified with respect to their grammatical function as arguments or adjuncts which is expressed by the label MO. Underspecification has proven to be useful for quick annotation of basic information. Adjunct-argument distinction, on the contrary, requires specialists' knowledge, and stronger commitment to theoretical assumptions with respect to adjunct- and argumenthood of PPs and datives.

(2.4) Trotzdem steht der Hof allen Interessierten zur Verfügung
nevertheless stands the yard all interested parties at the disposal
'nevertheless the yard is at the disposal of all interested parties'

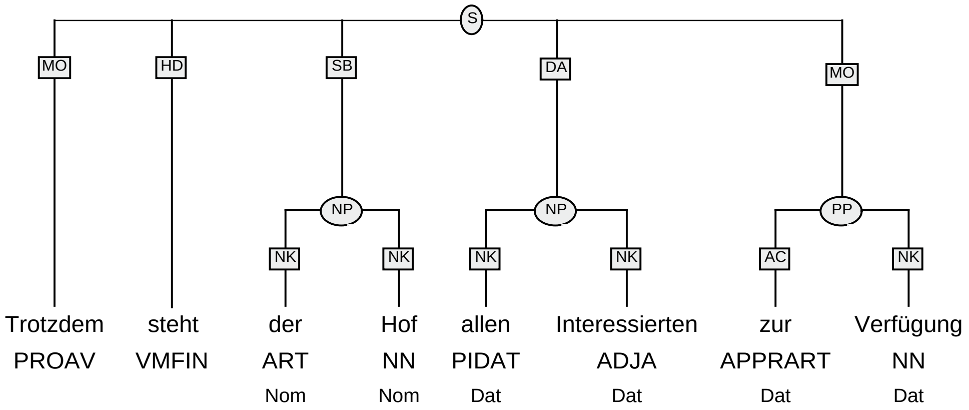


Figure 2.3: Syntactically annotated sentence containing the SVC *steht zur Verfügung*

For training of part-of-speech tagger and phrase chunker, only parts of the above information are used. The tagger is trained on the part-of-speech aligned word string, see the example below.

Trotzdem	steht	der	Hof	allen	Interessierten	zur	Verfügung
PROAV	VVFIN	ART	NN	PIDAT	ADJA	APPRART	NN

For training the chunker, part-of-speech information and partial information on syntactic structure and phrasal category is used as shown in figure 2.4. In contrast to the original treebank annotation (figure 2.3), function labels are omitted, and phrase structure is considered only at subsentential level. This strategy has been chosen because of the small amount of training material available. Accounting for complete syntactic structure would lead to an increase of the variety of structural patterns, and accounting for grammatical functions would lead to a drastic increase of the tagset, but would result only in a little gain of information, because of the functional underspecification of PPs and datives. In combination with a small training set, this is an unreliable basis for statistical learning. More importantly, PP-attachment is assumed to be guided by lexical information, thus an attachment model induced on the basis of parts-of-speech and phrasal category is considered to be inappropriate anyway.

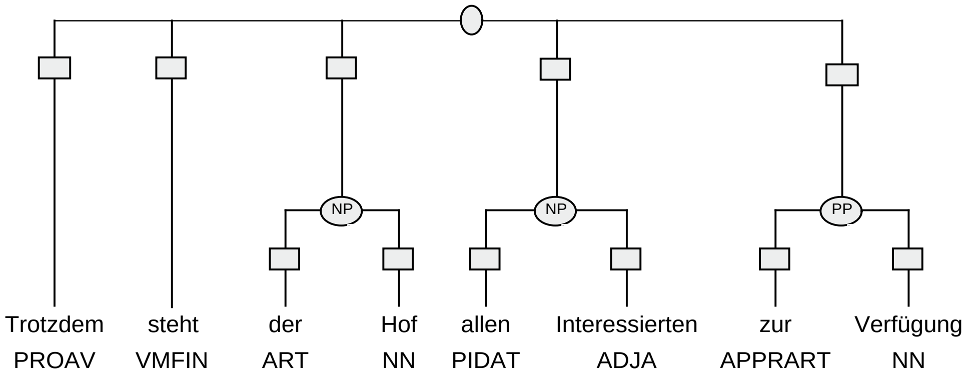


Figure 2.4: Information available for collocation identification

2.3 Statistics

2.3.1 Measures Applied for Collocation Identification

Frequency counts of word co-occurrences are the simplest estimates for lexical association between two or more words. The method, however, is fairly poor as only positive co-occurrences are taken into account, and occurrence frequencies of the single words are ignored. A variety of statistics-based lexical association measures have been proposed in the literature as an alternative to mere frequency-based collocation identification. Four of which will be presented in the following, namely mutual information MI as presented in [Church and Hanks, 1989], Dice coefficient [Smadja *et al.*, 1996], relative entropy I [Cover and Thomas, 1991], and a log-likelihood statistics Lgl [Dunning, 1993]. A general drawback of statistical measures is that they overgenerate in the case of low frequency data. MI, $Dice$, I and Lgl have been chosen as association measures, because they stand for two types of statistical measures. MI and $Dice$ are simple association ratios where the significance of the data is not accounted for. The measures differ with respect to the kind of association they model, i.e., MI models the ratio between the conditional probability $p(X|Y)$ and the marginal probability $p(X)$ or $p(X|Y)$ and $p(Y)$, respectively, while *Dice* sums the conditional probabilities $p(X|Y)$ and $p(Y|X)$. *Lgl* and I, in contrast to *MI* and *Dice*, take the significance of the individual word combinations into account. Thus they are less biased towards low frequency data. Both measures compare the informativity of frequency distributions of joint and marginal events. *Lgl* assigns extra weight to the joint probability.

Contingency tables are the standard means for representing positive and negative word co-occurrences. An example for a contingency table representing a collocation $c1c2$ with two collocates, $c1$ and $c2$, is given in table 2.1. $\neg$ indicates that the particular collocate is missing. Thus $c_1\neg c_2$ represents a pair consisting of $c1$ and any other word or word combination but $c2$.

	c_2	$\neg c_2$
c_1	c_1c_2	$c_1\neg c_2$
$\neg c_1$	$\neg c_1c_2$	$\neg c_1\neg c_2$

Table 2.1: Contingency table for collocations with two collocates

Simple Association Ratios

(Specific) Mutual Information

Mutual information *MI* as it has been introduced in [Church and Hanks, 1989] is a popular measure in computational linguistics to determine the strength of lexical association, see for instance [Smadja, 1993; Breidt, 1993; Daille *et al.*, 1994; Shimohata *et al.*, 1997]. Referring to [Fano, 1961], [Church and Hanks, 1989] present the following formula

$$MI = \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2.5}$$

In terms of word association, $p(x,y)$ represents the joint probability of a word combination c_1c_2, and $p(x)$, $p(y)$ represent the marginal probabilities of the potential collocates c_1 and c_2. Computing the logarithmic association ratio between joint and marginal probabilities, *MI* models the degree of association between c_1 and c_2 as follows:

$$\log(m) = \begin{cases} 0 & m = 1 \\ positive & m > 1 \\ negative & 0 < m < 1 \\ undefined & otherwise \end{cases}$$

with $m = \frac{p(c_1,c_2)}{p(c_1)p(c_2)}$.

Following [Church and Hanks, 1989], the cases can be interpreted as stated below:

$MI(c_1;c_2) = 0$, there is no particular relationship between c_1 and c_2.

$MI(c_1;c_2) < 0$, c_1, and c_2 are complementarily distributed.

$MI(c_1;c_2) > 0$, a genuine association between c_1 and c_2 exists.

Formula (2.5) is referred to by the term specific mutual information in [Smadja *et al.*, 1996]. The authors criticize that only positive occurrences can be accounted for, see formula (2.6).

$$\log \frac{p(X=1,Y=1)}{p(X=1)p(Y=1)} \tag{2.6}$$

This is a weak estimate for lexical association, as the association strength of low frequency occurrences is overestimated. The weakness of MI has already been pointed out in [Church and Hanks, 1989], where a threshold of 5 has been suggested as a remedy, i.e., an MI-value is only computed for word combinations that occur at least five times in the corpus used for collocation identification. Another strategy to achieve more reliable results is the application of significance tests to the co-occurrence data. For this task, the use of the t-test has been suggested in [Church and Hanks, 1989]. A drawback of the t-test is that it is valid only for normally distributed data, but normal distribution is fairly unlikely for language data. An alternative is the nonparametric χ^2 test. A concise presentation of the application of the t-test and the χ^2 test to collocation data is given in [Manning and Schütze, 1999]. The authors, however, claim that the differences between the t- and the χ^2 test in practice are rather small. They report that both tests lead to the same results for the 20 highest scoring bi-grams. Details on the test statistics can be found for the t-test in any standard book on parametric statistics, and for the χ^2 test in according books on nonparametric statistics. The χ^2 test is also discussed in section 2.3.2 of this work, as it is used for testing the significance of differences in accuracy between the models applied for collocation identification.

In [Smadja *et al.*, 1996], specific mutual information is opposed to average mutual information which is commonly known as mutual information in current information theory. In the following the logarithmic ratio presented in [Church and Hanks, 1989] will be termed MI, while the information theoretic measure will be addressed as I in accordance with newer information theoretic literature. I is more closely discussed in section 2.3.1.

Dice Coefficient

The *Dice* coefficient has been introduced in [Smadja *et al.*, 1996] as an alternative to *MI*. Comparably to *MI*, only positive occurrences are taken into consideration. Another similarity to *MI* is that large *Dice* values indicate strong lexical association. The formula for the *Dice* coefficient is

$$Dice(X,Y) = \frac{2 * p(X=1, Y=1)}{p(X=1) + p(Y=1)} \tag{2.7}$$

Unlike *MI* where the difference between conditional and marginal probabilities is calculated, word combinations are sorted according to the conditional probabilities with $p(X|Y)$ and $p(Y|X)$ having equal weight when *Dice* is applied (cf. the last equation in formula 2.8). Thus *Dice* is a means to account for the mutuality of the collocates. The formal differences between the measures are shown in equations 2.8 and 2.9.[16]

$$\begin{aligned} Dice(X,Y) &= \frac{2 * p(X,Y)}{p(X)+p(Y)} \\ &= \frac{2}{\frac{p(X)}{p(X,Y)} + \frac{p(Y)}{p(X,Y)}} \\ &= \frac{2}{\frac{p(X)}{p(Y|X)p(X)} + \frac{p(Y)}{p(X|Y)p(Y)}} \\ &= \frac{2}{\frac{1}{p(Y|X)} + \frac{1}{p(X|Y)}} \\ &= \frac{1}{2} * [p(Y|X) + p(X|Y)] \end{aligned} \tag{2.8}$$

$$\begin{aligned} MI(X,Y) &= log\frac{p(X,Y)}{p(X)p(Y)} \\ &= log\frac{p(X|Y)}{p(X)} \\ &= log\frac{p(Y|X)}{p(Y)} \\ &= log\,p(X|Y) - log\,p(X) \\ &= log\,p(Y|X) - log\,p(Y) \end{aligned} \tag{2.9}$$

---

[16] Note: $p(X|Y) = \frac{p(X \cap Y)}{p(Y)}$, $p(X \cap Y) \equiv p(X,Y)$.

Significance-Oriented Association Measures

Log-Likelihood

In [Dunning, 1993], a log-likelihood statistics (henceforth *Lgl*) is introduced as an alternative to simple association ratios like MI. Other than in MI and *Dice*, positive and negative word co-occurrences are accounted for. The measure is sensitive to the significance of a word co-occurrence. Dunning presents different formulations of the statistics amongst others the one given in formula 2.10.[17]

$$-2 \log \lambda = 2 \sum_{ij} k_{ij} \log \frac{k_{ij} N}{C_j R_i} \tag{2.10}$$

with

$$N = \sum_{ij} k_{ij}$$

$$C_j = \sum_{i} k_{ij}$$

$$R_i = \sum_{j} k_{ij}$$

where N is the total number of positive and negative occurrences in the table, k_{ij} is the frequency count in table cell ij. C_j is the sum over table column j, and R_i is the sum over table row i.

Relative Entropy

Formula (2.11) defines the relative entropy between two probability distributions. Given two random variables X and Y with a joint probability mass function $p(x, y)$, and marginal probability mass functions $p(x)$, $p(y)$, I is the relative entropy (Kullback-Leibler distance D, cross entropy, information divergence, mutual information) of the joint distribution $p(x, y)$ and the product distribution $p(x)p(y)$; it can also be expressed as expectation value E, cf. [Cover and Thomas, 1991].

---

[17] Cf. corpora list, 20 July 1997. The original formula is

$$-2 \log \lambda = 2 \sum_{ij} k_{ij} \log \frac{k_{ij} N}{R_j C_i}$$

The labels R and C have been exchanged in formula 2.10 for intuitivity.

$$
\begin{aligned}
I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= D(p(x,y)||p(x)p(y)) \\
&= E_{p(x,y)} log \frac{p(X,Y)}{p(X)p(Y)}
\end{aligned} \tag{2.11}
$$

The relation between I and entropy $H = -\sum_{i=1}^{n} x_i \log x_i$ can be described as follows:

$$
\begin{aligned}
I(X;Y) &= H[X] - H[X|Y] = H[Y] - H[Y|X] \\
&= H[X] + H[Y] - H[X,Y] \\
I(X;X) &= H(X)
\end{aligned} \tag{2.12}
$$

Similarly to Lgl, I-values for word combinations are calculated from positive and negative word co-occurrences of the potential collocates. In general the formulas for I and Lgl are largely comparable which is best seen in the next section. The major difference is that in Lgl the joint probability gets extra weighted by multiplying with N the number of PPs in the corpus. Thus interpretation of Lgl-values is similar to interpretation of I-values; i.e., the smaller the value, the higher the lexical association.

2.3.2 Statistics Employed for Significance Testing

"The procedures of statistical inference enable us to determine, in terms of probability, whether the observed difference is within the range which could easily occur by chance or whether it is so large that it signifies that the ... samples are probably from ... different populations." This sentence quoted from [Siegel, 1956], p. 2 is a good summary what statistical inferencing is about, namely to determine the probability of differences between two or more observed samples, by combining the actually identified sample distribution to hypothetic distribution by means of particular inference statistics. In the thesis, we are interested whether the differences in the number of true collocations identified from different sets of PNV-combinations are due to chance or result from a general difference of the goodness of the identification models applied.

In statistical inferencing, parametric and nonparametric tests are distinguished. **Parametric tests** require measurements at strength of at least interval scale. **Interval scale** means that the data of research are ordered according to a scale, and that the distances between any two numbers on the scale are known. Admissible operators are $=$, $\neq$, $>$, $<$, $+$, $-$. **Nonparametric tests**, on the other hand, are much weaker in their assumptions about the applicability of the test statistics to the data of the research. Nonparametric tests can be used with measurements at strength of ordinal scale, and some tests are even valid for **nominal scale** which is the weakest level of measurement. It is also called classificatory scale, as numbers or symbols are used to identify groups of entities. The admissible operators are $=$, $\neq$. If the groups can be ordered according to the relation $>$, and the relation holds for all pairs of groups, the level of measurement is said to be at strength of **ordinal scale**. The following operators are admissible: $=$, $\neq$, $>$, $<$. Further advantages of nonparametric tests are: the power of any such test may be increased by simply increasing the sample size; most nonparametric tests lead to exact probabilities; and if the sample size is very small, there are no alternatives to nonparametric tests except the properties of the underlying distribution, the population from which the test sample has been drawn, is known exactly, which is rarely the case.

As already mentioned, observed and theoretical values are compared in statistical significance testing. The observed values provide information about the actually occurring differences between the test samples under investigation, whereas the theoretical values provide the underlying hypothetic distribution against which the **null hypothesis** H_0 and the **alternative hypothesis** H_1 are tested. H_1 is also called research hypothesis. As testing the differences between the models for collocation identification is the task of the current work, H_0 and H_1 are defined as follows:

- H_0: there is no difference between the identification models being compared.
- H_1: differences between the models exist.

There are two common **significance levels**: $\alpha = .05$ and $\alpha = .01$. The levels indicate that the possibility is very small that the **null hypothesis** H_0 is true. In other words, α determines the size of the **region of rejection.** If $\alpha = .05$, the size of the region of rejection is 5 % of the total space determined by the curve of the sampling distribution. H_0 will be rejected in favor of H_1, the **alternative hypothesis** which is the research hypothesis. It is then said that the observed sample is in the region of rejection. Regions of rejection are illustrated in figure 2.5. Part a) shows the one-tailed case, part b) the two-tailed case. In the one-tailed case, the region of rejection is located at one side of the curve, whereas in the two-tailed case the region is divided into two equal parts which are located at the left and right end of the curve. Whereby the regions in a) and b) differ in location but not in total size. If no statement about the direction of the difference is made, a two tailed test is called for. This is the case with respect to the present study, as there are no a priori assumptions which would justify the one-tailed case. Employing the one-tailed case is only justified if there is strong theoretical or empirical a priory evidence that one of the models tested will be better than the other one(s). In either case, the significance of the observed value ought to be looked up in a table specifying the values of the theoretical distribution.

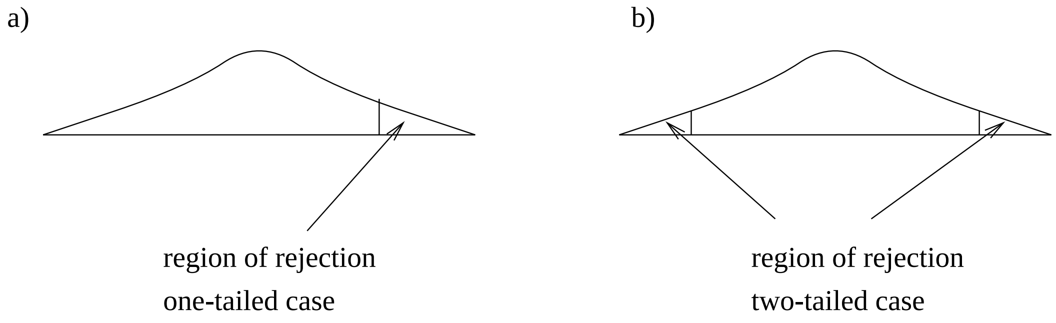


Figure 2.5: Schematic representation of regions of rejection for one- and two-tailed tests

The following tests are employed for comparing the differences between the collocation identification models: the χ^2 test for k independent samples and its special variant for the 2-sample case. The tests have been chosen as they are nonparametric and allow applying to data at nominal scale. In terms of our test data, independent samples means that each model for collocation identification selects a different subset (sample) from the initial data. The data are at nominal scale, as PNV-combinations are grouped together according to their occurrence frequency in the extraction corpus under investigation, with occurrence frequencies being used as labels but not for ranking the combinations.

The χ^2 Test for k Independent Samples

The procedure for the χ^2 test is defined as quoted from [Siegel, 1956], see p. 178:

> 1. Cast the observed frequencies in $k \times r$ contingency table, using the k columns for the groups.
> 2. Determine the expected frequency under H_0 for each cell by finding the product of the marginal totals common to the cell and dividing this product by N. (N is the sum of each group of marginal totals. It presents the total number of *independent* observations. Inflated N's invalidate the test.)
> 3. Compute χ^2 by using Formula 2.13. Determine
>
> $$df = (k-1)(r-1)$$

4. Determine the significance of the observed value of χ^2 by reference to Table 2.2.[18] If the probability given for the observed value of χ^2 for the observed value of *df* is equal to or smaller than α, reject H_0 in favor of H_1.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.13}$$

O_{ij} is the observed value in each cell of the $k \times r$ contingency table representing the data of research. E_{ij} is the expected value related to each observed value. E_{ij} is computed as follows: "To find the expected frequency for each cell (E_{ij}), multiply the two marginal totals common to a particular cell, and divide this product by the total number of cases, N." (Cf. [Siegel, 1956], p. 105.)

df	probability under H_0 that $\chi^2 \geq$ *chi square*						
	.99	.98	.95	.90	.80	.70	.50
1	.00016	.00063	.0039	.016	.064	.15	.46
3	.12	.18	.35	.58	1.00	1.42	2.37
4	.30	.43	.71	1.06	1.65	2.20	3.36

df	probability under H_0 that $\chi^2 \geq$ *chi square*						
	.30	.20	.10	.05	.02	.01	.001
1	1.07	1.64	2.71	3.84	5.41	6.64	10.83
3	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	4.88	5.99	7.78	9.49	11.67	13.28	18.46

Table 2.2: Critical values of chi square when *df* = 1, 3, 4, with χ_2 representing the observed values and *chi square* standing for the theoretical value

In the following example, it is tested whether *MI*, *Dice*, *I* and *Lgl* differ significantly given the data in the contingency table 2.3. E_{ij} values are set in brackets.

The research hypothesis to be tested is

H_1: The lexical association models differ with respect to their goodness for collocation identification.

The related null hypothesis is

H_0: The lexical association models do not differ in their ability for collocation identification.

	MI	Dice	I	Lgl	$\sum$
true collocations	214	189	180	180	763
	(190.75)	(190.75)	(190.75)	(190.75)	
false collocations	286	311	320	320	1237
	(309.25)	(309.25)	(309.25)	(309.25)	
	500	500	500	500	2000 N

Table 2.3: Contingency table containing data gained by applying *MI*, *Dice*, *I* and *Lgl*

The E_{ij} values for positive and negative hits are calculated as follows:

$$E_{ij_{true\,coll}} = \frac{763 * 500}{2000} = 190.75$$

$$E_{ij_{false\,coll}} = \frac{1237 * 500}{2000} = 309.25$$

Applying formula 2.13 to the data we get

$$\begin{aligned}
\chi^2 &= \frac{(214-190.75)^2}{190.75} + \frac{(189-190.75)^2}{190.75} + \frac{(180-190.75)^2}{190.75} + \\
&\quad \frac{(180-190.75)^2}{190.75} + \frac{(286-309.25)^2}{309.25} + \frac{(311-309.25)^2}{309.25} + \\
&\quad \frac{(320-309.25)^2}{309.25} + \frac{(320-309.25)^2}{309.25} \\
&= 6.56685
\end{aligned}$$

Comparison of the observed value $\chi^2 = 6.56685$ with the table 2.2 of critical values reveals that χ^2 has probability of occurrence under H_0 of $p > .05$, with level of freedom $df = (4-1)(2-1) = 3$. As significance level $p > .05$ is larger than the upper limit for the critical value for rejection $\alpha = .05$, H_0, cannot be rejected. This means, there is no significant difference between the measures.

The χ^2 Test for Two Independent Samples

The procedure for the χ^2 test is defined as quoted from [Siegel, 1956], see p. 109:

1. Cast the observed frequencies in a $k \times r$ contingency table, using the k columns for the groups and the r rows for the conditions. Thus

---

[18] Table 2.2 refers to table C in [Siegel, 1956], p. 249.

> for this test $k = 2$.
> 2. Determine the expected frequency for each cell by finding the product of the marginal totals common to it and dividing this by N. (N is the sum of each group of marginal totals. It represents the total number of *independent* observations. Inflated N's invalidate the test.) Step 2 is unnecessary if the data are in a 2×2 table and this formula 2.14 is to be used.
> 3. ...
>
> 4. Determine the significance of the observed χ^2 by reference to Table 2.2.[19] ... If the probability given by Table 2.2 is equal to or smaller than α, reject H_0 in favor of H_1.

The χ^2 test is exemplified by comparing the best association model and the entropy model. The test for independent samples is suitable as the data of research differ with respect to sample and size. See table 2.4 for illustration. Here freq, the best association model for A is compared with the entropy model with respect to identification of collocations$_{all}$. There are two different samples: the 500 highest ranked PNV-combinations identified by co-occurrence frequency, and the set of 235 PNV-combinations with entropy values < 0.7 of the PP-collocates. Note the sum of the row sums equals the sum of the column sums equals the set size $N = 735$.

	true colloc.	noncolloc.	sample size
freq	353 A	147 B	500
entropy	182 C	53 D	235
$\sum$	535	200	735 N

Table 2.4: Number of true collocations and noncollocations identified by (i) frequency *freq* from the 500 highest ranked PNV-combinations in set A, collocations$_{all}$, and (ii) applying the entropy model to set A, collocations$_{all}$

Entering the values in formula (2.14) results in

$$\begin{aligned}(2.14)\qquad \chi^2 &= \frac{N(|AD - BC| - \frac{N}{2})^2}{(A+B)(C+D)(A+C)(B+D)} \\ &= \frac{735(|353 * 53 - 147 * 182| - \frac{735}{2})^2}{(353+147)(182+53)(353+182)(147+53)} \\ &= 3.20\end{aligned}$$

The hypotheses used in our example are:

---

[19] Table 2.2 refers to table C in [Siegel, 1956], p. 249.

H_1**:** Frequency and the entropy model differ with respect to their goodness for collocation identification.

H_0**:** There is no difference between mere co-occurrence frequency and the entropy model.

Since H_1 predicts no direction of the difference between the models, the region of rejection is two-tailed. The observed probabilities are compared with the theoretical values by looking up the table of critical values (table 2.2).

Comparing the observed value $\chi^2 = 3.2$ with the critical values reveals that χ^2 has probability of occurrence under H_0 of $p > .10$, which is above the critical value $\alpha = .05$ for $df = 1$. Thus H_0 is not in the region of rejection, and cannot be rejected. In other words, for set A, collocations$_{all}$ the entropy model and a merely frequency-based approach do not significantly differ.

2.4 Database Technology

2.4.1 The Concept of a Relational Database

The key idea of a relational database is to think of information as being grouped in **tables,** also called **relations,** and the tables having the properties of sets. Two kinds of tables are of interest for the work presented here: base relations and query results. **Base relations** are named relations which are important enough to be a direct part of the database. They are defined by the database designer. A **query result** is an unnamed derived table which results from executing a query. A table or relation consists of a heading and a body. The **heading** is defined as a set of attributes, with an attribute or field occupying a column in the table. All attribute values are atomic. The pool of legal values for an attribute is called **domain**. The number of attributes is the **degree** of the relation. The **body** consists of a set of tuples, with a tuple or record occupying a row in the table. The number of tuples constitutes the **cardinality** of the relation. Each table has a **primary key**, i.e., at least one attribute which has different values in each row of the table. SQL is the standard language for interacting with a relational database. More details on relational databases can be found in [Date, 1995].

2.4.2 The Core Machinery

The database management system TSDB [Oepen *et al.*, 1998] is used for storing the descriptions and corpus examples related to collocations. TSDB has been developed in the TSNLP[20] project at the German Research Institute for Arti-

---

[20] See `http://cl-www.dfki.uni-sb.de/tsnlp/` for a comprehensive presentation of the project.

ficial Intelligence (DFKI), Saarbrücken. The core engine written in ANSI C is highly flexible in its interfaces and portable.

The database has been chosen for the following reasons:

Adequacy TSDB has been designed with the aim of developing data for natural language research and applications, thus the database kernel is small and flexible. Retrieval by string manipulation (regular expression matching) is supported. The interface allows connection to arbitrary applications.

Flexibility The database consists of

- a binary file comprising the engine and a library of interface functions;
- the relations file storing the names of the base relations and the headings, i.e., the names of the permissible attributes and the types of their values;
- a data file for each base relation comprising the body of the relation.

The relations file and the data files are plain ASCII. The user is free to define the data format. Thus new relations and databases can be easily set up. Headings and bodies can be easily changed or extended by manipulation of the relations file and string operations on the data files.

Availability and Compatibility The database is non-commercial and runs on different platforms, such as Unix, Macintosh and Intel-based personal computers. Thus exchange with other research institutions is facilitated.

In contrast to the original use of the database for storing a restricted amount of manually constructed data, the database is now used for handling large amounts of data derived from corpora, leading to relations with high cardinality.

Chapter 3

Construction and Characteristics of the Extraction Corpus

3.1 Introduction

In the current chapter an architecture for shallow syntactic processing of arbitrary text is presented (section 3.2). Characteristics of the resulting extraction corpus are discussed in section 3.3, and a classification of the preposition-noun-verb combinations found in the extraction corpus is given in section 3.4.

Syntactically annotated corpora are a suitable basis for collocation extraction. From a statistics-based point of view, collocations are recurrent events in natural language. Lexical selection between the collocates is reflected by overproportionally large occurrence frequencies of collocations in corpora. Thus, frequency-based approaches are expected to be well suited for retrieval of collocations from corpora. Corpora for collocation extraction are required to be large, as the lexical material of a corpus is distributed comparable to Zipf's law, see section 3.3.2. Thus there is, on the one hand, a small number of frequently occurring words, and on the other hand there is a large number of infrequently occurring words, with content words usually being infrequent, and function words being frequent, apart from a few exceptions like rare prepositions, adverbs or particles. As the majority of collocates are content words, automatic preprocessing of extraction corpora without hand correction is an important precondition for accessing sufficiently large amounts of data for collocation extraction from various domains.

In an approach, where candidate collocations are derived from syntactically preprocessed text collocation identification is guided by explicit linguistic information. The notion of numeric span is replaced by syntactic span, i.e., instead of looking at word sequences of particular length, specific syntactic structures are examined, which leads to the following advantages:

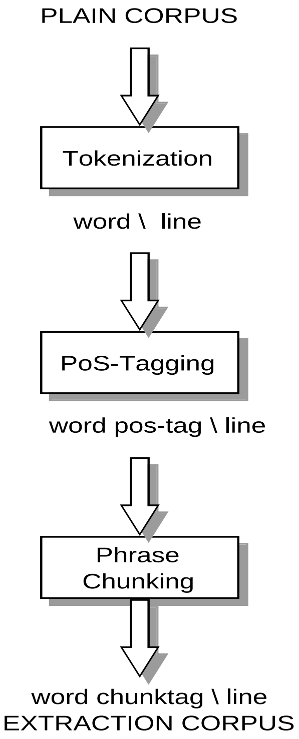
- Collocation candidates with inappropriate syntactic structure are avoided.
- Due to part-of-speech information, a distinction is possible between collocations and purely syntactically motivated co-occurrences of lexical items like article-noun, auxiliary-participle, or auxiliary-infinitive co-occurrences.
- Syntactically flexible collocations can easily be identified, which is important for PP-verb collocations, as many of them are flexible with respect to word order and syntactic transformation.
- Syntactic rigidity in the collocation phrase can be utilized as extra evidence for the collocativity of a word combination.
- Better insights into the interaction of lexical and structural processes are possible which is important for further development of grammar theory.

Three classes of collocations – figurative expressions including idioms, support-verb constructions, and pseudo-collocations – are manually identified from the subset of PNV-combinations which occur three times or more in the set of PNV-triples selected from the syntactically preprocessed corpus. While figurative expressions cover uninterpretable constructions as well as a broad range of constructions that require figurative or metaphoric interpretation, the group of support-verb constructions consists of noun-verb collocations that are comparable to verbal predicates with respect to their grammatical function. The major characteristics of the third group of collocations is their high occurrence frequency in the particular corpus under investigation. See sections 3.4.2, 3.4.3 and 3.4.4 for a discussion of the respective collocation classes. Even though the groups differ in their cores, no sharp borderlines can be drawn. Figurative expressions for instance are closely related to idioms as in both cases the lexical material is reinterpreted, and literal interpretation is available but unlikely. Another group of PP-verb combinations on the one hand require figurative interpretation, and on the other hand are comparable to SVCs insofar as a particular noun combines with more than one verb to express different aspects of the meaning of the underlying predicate. The manually identified word combinations are the reference material against which the collocation identification methods described in section 4 are tested. Collocation-class-specific frequency distributions are briefly summarized in section 3.4.5.

3.2 An Architecture for Automatic Syntactic Preprocessing of Large Text Corpora

Figure 3.1 shows the architecture used for automatic annotation of basic syntactic information to arbitrary text. The following processing steps are applied.

Figure 3.1: Architecture for preprocessing of the extraction corpus



First, the plain corpus is tokenized. Any standard tokenizer will be sufficient, i.e., the text shall be split on white space. Punctuation, apostrophes, brackets etc. shall be analyzed as tokens, and hyphenation shall be taken care of. Some effort is required for identification of sentence boundaries, as sentences, on the one hand, constitute the maximal spans from which PNV-combinations are selected, and on the other hand, sentences are the units which are stored in the collocation database. Therefore the distinction is important between dots functioning as full stops and dots functioning as abbreviation markers or as parts of numbers and dates.

The tokenized text is used as input to a part-of-speech tagger. The particular tagger used in this work is described in [Brants, 1996; Brants, 1999]. Part-of-speech tagging is a first step in reducing the amount of syntactically implausible collocation candidates. Shallow parsing like phrase chunking is a further step to increase the accuracy of collocation identification. Phrase chunking instead of full parsing is applied because it allows for robust and efficient processing of arbitrary text, and it provides sufficient information for the task at hand. Ideally, full parsing is the best means to avoid retrieval of false collocation candidates. In practice, however, full parsing is not feasible as the coverage of existing parse grammars is insufficient, and structural alternatives lead to hardly resolvable ambiguity. Thus statistical methods are used in the study, because they are for the time being best suited for efficient and robust processing of large amounts of arbitrary text. The particular chunker applied is described in [Skut and Brants, 1998; Skut, forthcoming].

In addition, morphosyntactically flexible collocates are reduced to base forms, which helps to increase frequency counts especially in highly inflecting languages like German. Stemming is not part of the general preprocessing component as it is only used sporadically depending on the particular class of collocations to be extracted from the corpus data. In the case of PP-verb collocations, for instance, reduction of verb forms is useful. In this work, mmorph[1] is used for reduction of verb forms to their bases.

3.3 Characteristics of the Extraction Corpus

3.3.1 Information Utilized for Selection of Collocation Candidates

The corpus used for collocation extraction is annotated similar to the training material as shown in figure 2.4, section 2.2.2. Each word in the extraction corpus is automatically annotated with a part-of-speech label, a mother category label, and a label representing the position of the word within syntactic structure. In particular NP, PP and ADJP chunks are annotated. The investigations presented in this work concentrate on PP-verb combinations, because they cover two classes of fundamental linguistic structure, namely nominal and verbal projections. Thus lexicalization phenomena within the NP[2], as well as the influence of lexicalization on argument structure and word order can be investigated. Figure 3.2 shows information used as basis for PP-verb collocation extraction, which is:

- A lexicon containing word forms, parts-of-speech and occurrence frequencies: For extraction of PP-verb collocations only nouns, verbs and prepo-

---

[1] Mmorph, MULTEXT morphology tool provided by ISSCO/SUISSETRA, Geneva, Switzerland.

[2] Recall, PPs are syntactically comparable to NPs with the preposition functioning as case marker.

sitions are of interest.

- Sentences annotated with parts-of-speech and basic syntactic structure: Sentences are the basic units within which the collocates of a collocation need to co-occur.
- Lexical tuples and their co-occurrence frequencies: In particular, preposition-noun bi-grams and preposition-noun-verb tri-grams are of interest. Preposition and noun need to be constructors of the same PP. PP and verb need to co-occur in a sentence. Arguments for this lax co-occurrence requirements are given below.
- PP instances constituted by a particular preposition-noun combination: Here the full variation of realizations of PPs constituted by individual preposition-noun combinations is stored.

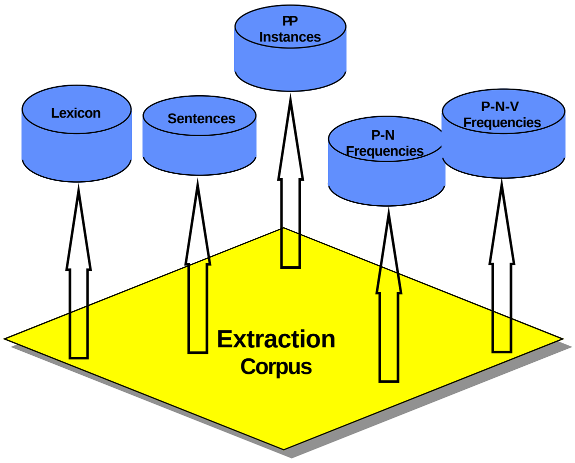


Figure 3.2: Information derived from the extraction corpus

Phrasal attachment is a major source of uncertainty in parsing as well as in chunking. PP-attachment cannot be decided on purely syntactic grounds. In the sentence *wir sahen heute den Sohn des Sängers mit den neuen Brillen* (we saw today the son of the singer with the new glasses), there are three potential attachment sites for the PP *mit den neuen Brillen*, it can be either attached to the verb *sahen* (high attachment), the NP *den Sohn* (low attachment to NP_1) or to the embedded genitive *des Sängers* (low attachment to NP_2). Current stochastic parsers tend to overestimate low attachment of PPs in the middle field when preceded by an NP. This reflects, on the one hand, the fact that in the training data instances of low attachment outnumber instances of high attachment and, on the other hand, it reflects the fact that access to purely structural information is not sufficient to decide on PP-attachment.[3] As a consequence, the chunker decisions are not reliable in the case of '(NP PP)'-chunks. Thus PP-verb co-occurrence frequencies are calculated for all PP-verb pairs co-occurring within a clause. This approach leads to over-generation of PP-verb tuples which, however, is a largely negligible factor, as a number of artificial PP-verb combinations will be excluded from consideration as collocation candidates because of low occurrence frequency.

Procedures for PN- and PNV-Extraction

In the following, the algorithm used for extracting preposition-noun-verb combinations is presented. The algorithm can be divided into three parts, namely

1. extraction of PN-combinations and PPs,
2. extraction of verbs, and
3. combinations of PN-pairs and verbs.

The extraction corpus is processed sentence by sentence. Words are normalized to small letters to avoid a distinction between PN- and PNV-combinations that only differ with respect to upper and lower case.

First, each phrase containing a preposition and a dependent noun is represented by the according preposition-noun pair. Both, PN-pairs and their occurrence frequencies, as well as the complete PPs[4] are stored. While the PN-combinations constitute the abstract representation of the PP-collocate, the PP instances are required as input for the PP-entropy model (cf. section 4.4.2).

Second, all verbs of a sentence are extracted. Infinitives with *zu* (to) are treated like single words, and separated verb prefixes are reattached to the verb. Only the main verbs are extracted from complex predicates.

Third, for all PPs and main verbs co-occurring within a sentence, PNV-triples are constructed. This strategy is a simple means to cover collocation instances where nominal and verbal collocates are part of different substructures for instance when the verbal collocate is part of a relative clause, but it also leads to generation of unwanted PP-verb combinations. This naive method, however, is justifiable as it allows to increase occurrence frequencies of true collocations,

---

[3] A better account for PP-attachment employing knowledge on lexical collocations is one of the motivations for investigating PP-verb collocations in the thesis.

[4] A PP here contains the preposition, the dependent noun, and the word string in between.

while a major part of arbitrary combinations will be left out from further processing because of infrequency. A different strategy is required for identification of collocations with verbal collocates that are homophonous to auxiliaries like *in Kraft sein* ('be in force'), *im Gespräch sein* ('be under discussion'), *in Führung sein* ('be in the lead'), *unter Druck sein* ('be under pressure'), or modals like *in Ruhe lassen* ('leave alone'), *im Stich lassen* ('forsake'). In this case, only sentences with simple predicates can be used for PNV-construction, to ensure that auxiliary and modal constructions are left out. In addition, information on systematic co-occurrence of PN-pairs and complex predicates is also useful, which however will not be elaborated in this work.

3.3.2 Distribution of Words and Word Combinations in Text

The distribution of words within a corpus approximates Zipf's law which says $n_c > n_{c+1}$, with n_c the number of words occurring c-times; i.e., with increasing count c the number of words occurring c-times decreases. In other words, there are more infrequently recurring words in texts than frequently recurring ones. Function words like articles, prepositions, auxiliaries are usually frequent, while content words such as nouns, main verbs and adjectives tend to be infrequent. A comparable distribution can also be found with respect to word combinations, i.e., there is only a small number of frequently occurring word combinations which represent preposition-noun (PN) and preposition-noun-verb (PNV) combinations compared to a large number of infrequent ones.

569 310 PNV-combinations (types) have been selected from the extraction corpus including main verbs, modals and auxiliaries. The set of triples covers 2 209 452 word tokens. As already explained, this is a theoretical maximum, because verbs are duplicated in sentences that contain more than one PP. Similarly PPs are duplicated in sentences where more than one main verb is found. For comparison, the number of prepositions and nouns identified sums to 965 902, and there are 971 012 verb forms identified in the 8 million word corpus amounting to 1 936 914 tokens. Considering only combinations with main verbs, the number of PNV-types reduces to 372 212 which represents a theoretical maximum of 1 362 264 tokens comprising a preposition, a noun and a main verb.[5] Table 3.1 shows the set of 372 212 PNV-types ranked by occurrence frequency. The first line at the left side of the table says that there are 323 768 PNV-combinations (full forms) that occur only once in the corpus. At the other end, there are few word combinations that occur more than 10 times, for instance

---

[5] The following example explains how the number of tokens is calculated. There are 372 212 PNV-types with different rank of occurrence frequency, multiplication of the types by their ranks results in 454 088 PNV-instances which multiplied by 3 leads to 1 362 264 word tokens.

118 PNV-instances occur 10 times, 17 occur 20 times, 7 occur 30 times, 3 occur 40 times, 1 occurs 50 times and 2 occur 60 and 70 times, respectively. At the far end where $c \geq 100$, there are only single occurrences of PNV-combinations. See the frequency n_c and rank order c pairs printed in bold face. The table also shows that, comparable to Zipf's law, n_c increases with decreasing c and vice versa with only a few exceptions when n_c is low.

n_c	c	n_c	c	n_c	c	n_c	c
323768	**1**	16	21	2	41	2	67
38014	2	19	22	5	42	3	68
4775	3	14	23	4	43	2	69
2792	4	12	24	1	44	**2**	**70**
826	5	9	25	1	45	1	71
603	6	7	26	1	46	3	74
320	7	10	27	3	47	1	75
235	8	6	28	1	48	1	78
133	9	5	29	1	49	1	92
118	**10**	**7**	**30**	**1**	**50**	3	95
74	11	4	31	4	52	1	98
78	12	5	32	2	53	**1**	**111**
56	13	5	33	2	54	**1**	**115**
38	14	5	34	2	57	**1**	**128**
47	15	4	35	2	59	**1**	**143**
34	16	2	36	**2**	**60**	**1**	**174**
28	17	4	37	2	61	**1**	**185**
32	18	1	38	1	62	**1**	**379**
18	19	3	39	1	63		
17	**20**	**3**	**40**	1	66		

Table 3.1: Preposition-noun-main verb occurrences in the extraction corpus

The diagram in figure 3.3 illustrates the partition of PNV-combinations according to the ranks $c = 1$, $c = 2$, $c \geq 3$. The majority of PNV-combinations (87 %) occurs only once (n_c = 323 768); 10 % occur 2 times (n_c = 38 014); and a small rest of 3 % (10 430 PNV-types) occur three times or more. Thus only a small subset of word combinations remains as a basis for statistics-based collocation identification.

Consider also the diagram in figure 3.4 where the composition of the subset where $c \geq 3$ is shown. While almost half (46 %) of the data occur three times, only 6 % occur more than 10 times. In total, there are at most 54 292 preposition, noun and main verb tokens covered by the word combinations that occur at least three times in the extraction corpus. Together with the tendency of statistical models to overestimate low frequency data, it becomes clear that large corpora are required as a starting point for collocation identification but only a very small percentage of the data is well suited for collocation identification.

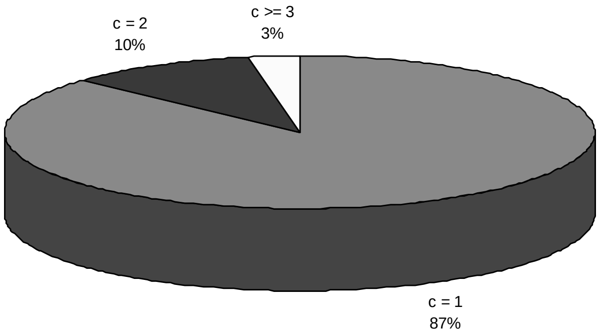


Figure 3.3: Distribution of PNV-combinations in the extraction corpus according to co-occurrence frequency c

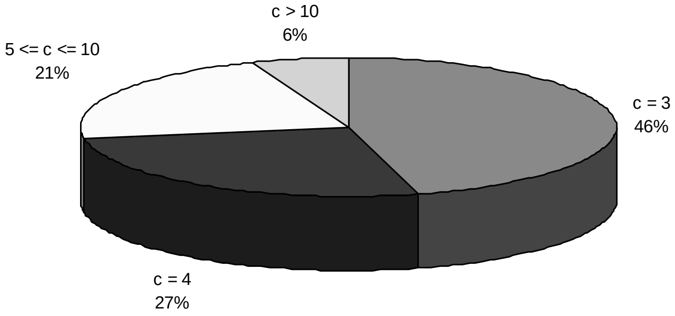


Figure 3.4: Distribution of PNV-combinations where co-occurrence frequency $c \geq 3$

In the case of highly inflecting languages, reducing words to their bases is an appropriate strategy for increasing occurrence frequencies. In this work, only verbs are reduced, as morphological variation of the verb does not influence the collocativity or noncollocativity of the PNV-combinations, as could be confirmed by studying the data retrieved from the newspaper corpus. In the following, the frequency distributions are summarized after the verbs have been reduced to base forms. The figures for full forms are enclosed in brackets. A decrease in the percentage of unique occurrences to 80.6 % (87 %) is opposed to an increase of recurrent data, i.e., 14.6 % (10 %) for $c = 2$ and 4.8 % (3 %) for $c \geq 3$. In general, the number of recurrent combinations increases with reduction of the verb forms. This tendency is also reflected in the set of PNV-combinations where $c \geq 3$. Here the percentage of combinations where $c = 3$ decreases while the proportions of the other subsets increase, i.e., 29.2 % (27 %) for $c = 4$, to 23.2 % (21 %) for $5 \leq c \leq 10$, and to 7 % (6 %) for $c > 10$.

In analogy to Zipf's law, the occurrence frequency of PNV-triples decreases with increasing rank. There are 4 774 types of full form triples with rank 3, but only 629 that rank higher than 10. Similarly there are 6 358 verb base triples with rank 3, and only 1 097 triples with rank above 10. The picture changes when looking at word tokens instead of PNV-types. Approximately the same amount of word forms is covered by the different groupings of full form triples, where the set of PNV-combinations defined by $c = 4$ covers a slightly smaller number of word tokens (33 504) than the other groups do. When the verbs are reduced to their bases, the distribution of word tokens divides at rank 5. In other words, the number of tokens is comparable in the groups $c = 3$ and $c = 4$, and the groups $5 \leq c \leq 10$ and $c > 10$. See table 3.2 for the occurrence frequencies of PNV-combinations in the extraction corpus.

	full forms		verb bases	
rank	pnv types	word tokens	colloc types	word tokens
$c = 3$	4 774	42 966	6 358	57 222
$c = 4$	2 792	33 504	4 585	55 020
$5 \leq c \leq 10$	2 235	42 735	3 643	70 488
$c > 10$	629	43 671	1 097	79 206
$\sum$	10 430	162 876	15 683	261 936

Table 3.2: Distribution of preposition-noun-main verb combinations ranked by occurrence frequency c

The word combinations are grouped according to rank of occurrence frequency and realization of the verb (full or base form).

Summing up,

1. Reduction to base forms is a simple means for increasing lexical co-occurrence frequencies.

2. A small number of frequently occurring word combinations covers a large portion of word tokens in text.

As a consequence of 1., a larger number of word combinations becomes appropriate for statistical evaluation. The discrepancy established by 2. is particularly clear with respect to the PNV-combinations containing verb bases, thus a proper treatment of highly recurrent word combinations is an important factor in natural language processing.

3.4 Classes of PNV-Combinations

3.4.1 An Overview

Recurrence is one of the key criteria for collocations in the present study. Single and very infrequent PNV-occurrences are useless for statistics-based collocation identification. Thus only PNV-combinations with co-occurrence frequency larger than two ($c > 2$) will be examined in the following.

First of all, an overview of classes of PNV-combinations found in the extraction corpus is given by examining the 20 most frequent PNV-combinations (table 3.3) and combinations where $c = 60, 50, 40, 30, 20$ (tables 3.4 and 3.5). Two cases are distinguished, namely full form triples (henceforth P.N.V(full form)-triples) and triples where the verb has been reduced to its base form (henceforth P.N.V(base form)-triples). Recall, the potential collocates have been normalized to small letters. The tables show that the number of support-verb constructions SVC (•) and figurative expressions (◇) increases when the verbs are reduced, i.e., there are 7 SVCs and 4 figurative expressions among the 20 most frequent full form combinations compared to 10 SVC and 6 figurative expressions among the triples containing verb bases. Accordingly the number of arbitrary word combinations decreases in the latter case.

The tables show that the number of highly recurrent data increases when morphological information is abstracted away. Density of SVCs and figurative expression among the data decreases, even though their total number increases. The examples also show that there is a large proportion of word combinations which are frequent but not collocational in a narrow sense, which requires closer examination. For further investigation, the following classes of PNV-combinations are distinguished:

Support-verb constructions • Apart from *vor Gericht gestellt*, there are 12 instances of support-verb constructions in the full form data which reduce to the following types: 1) *zur Verfügung* {{*stehen, steht, standen*}, {*gestellt, zu stellen*}} ('at the disposal be', 'make available'), 2) *ums Leben gekommen*

class	P.N.V(full form)	c
	um uhr beginnt	379
	bis uhr geöffnet	182
●	zur verfügung stehen	174
●	zur verfügung gestellt	143
●	zur verfügung stellen	128
●	zur verfügung steht	115
●	ums leben gekommen	111
◇	auf programm stehen	98
●	in anspruch genommen	95
○	am montag sagte	95
○	am dienstag sagte	95
◇	auf tagesordnung stehen	92
	am donnerstag sagte	78
	auf seite lesen	75
◇	im mittelpunkt steht	74
	auf kürzungen vor_behält	74
◇	auf programm steht	74
	am mittwoch sagte	71
●	zur verfügung zu_stellen	70
	auf seite zeigen	70

class	P.N.V(base form)	c
●	zur verfügung stellen	457
	um uhr beginnen	420
●	zur verfügung stehen	404
	bis uhr öffnen	196
●	ums leben kommen	195
◇	auf programm stehen	193
●	in anspruch nehmen	192
◇	im mittelpunkt stehen	176
◇	auf tagesordnung stehen	159
●	in frage stellen	146
●	in kraft treten	126
●	in frage kommen	120
◇	im vordergrund stehen	112
●	zur kenntnis nehmen	111
	am dienstag sagen	102
	am montag sagen	101
●	zu ende gehen	91
●	in griff bekommen	90
◇	ins leben rufen	89
◇	auf beine stellen	87

Table 3.3: The 20 most frequent PNV-combinations in the extraction corpus

('die'), 3) *in Anspruch genommen* ('claim'), 4) *in Frage stellt* ('question'), and 5) *zu Ende gehen* ('end'). For a closer discussion see section 3.4.3.

Figurative expressions ◇ Examples of figurative expressions are: *stehen* (to stand) + LOCATIVE such as *auf (dem) Programm* {*steht, stehen*} ('be on the programme'), *auf (der) Tagesordnung stehen* ('be on the agenda'), *im Mittelpunkt steht* ('be the center of attention'), *unter (Det) Motto steht* ('be the motto'); and *gehen* (to go) + LOCATIVE – *über (die) Bühne geht* like *gut über die Bühne gehen* ("go well"). For each example, except for *unter (Det) Motto steht*, literal and figurative interpretation is available, as the nouns can be interpreted as having spatial extension which is not the case for *Motto*. Literal reading, however, is in all cases less likely than figurative interpretation. Another example is *vor Gericht gestellt*. The expression also makes use of locative metaphor, but even though spatial interpretation is available for the noun *Gericht* literal interpretation is odd. A discussion of figurative expressions is presented in section 3.4.2.

class	P.N.V(full form)	c
	im anzeigenteil entnehmen	60
	im anzeigenteil bitte	60
	am dienstag mitteilte	50
	zur unsterblichkeit ägypten	40
	für sonntag lädt	40
	am samstag findet	40
●	zur verfügung standen	30
●	zu ende ging	30
●	zur kenntnis nehmen	30
	vor journalisten sagte	30
◇	unter motto steht	30
	um uhr gibt	30
	für donnerstag lädt	30

class	P.N.V(base form)	c
	zu hause bleiben	60
	im anzeigenteil entnehmen	60
	im anzeigenteil bitten	60
	am samstag treffen	60
	zum vorsitzenden wählen	50
	um uhr treffen	50
	auf seite bitten	50
	am freitag mitteilen	50
	zur unsterblichkeit ägypen	40
	um uhr hören	40
●	in führung gehen	40
◇	auf punkt bringen	40
	am sonntag spielen	40
	am samstag spielen	40
	am dienstag berichten	40
	zur schule gehen	30
	vor journalisten sagen	30
	vor jahren gründen	30
	um uhr kommen	30
	nach hause fahren	30
	in stadthalle sehen	30
●	in pflicht nehmen	30
	bis uhr stehen	30
◇	auf liste stehen	30
◇	auf bühne stehen	30
	am samstag geben	30
	am donnerstag treffen	30

Table 3.4: PNV-combinations that occur 60, 50, 40, 30 times in the extraction corpus

class	P.N.V(full form)	*c*
● ◇	vor gericht gestellt	20
	von sachspenden bitten	20
	um uhr hören	20
	um uhr hält	20
●	in frage stellt	20
●	in frage kommen	20
	im gespräch sagte	20
	im bürgerhaus beginnt	20
	auf anfrage bestätigte	20
	an anzeigenschaltern entgegengenommen	20
	am wochenende sagte	20
	am turm entgegengenommen	20
	am sonntag nachmittag	20
	am sonntag feiert	20
	am samstag beginnt	20
	am mittwoch findet	20
	am dienstag findet	20

class	P.N.V(base form)	*c*
●	zur diskussion stellen	20
●	zum zug kommen	20
	von sachspenden bitten	20
●	unter strafe stellen	20
	um kinder kümmern	20
◇	über wasser halten	20
	mit thema beschäftigen	20
●	ins gespräch bringen	20
●	in schranken weisen	20
●	in rechnung stellen	20
●	im zusammenhang stehen	20
◇	im stich lassen	20
	im rathaus geben	20
	für montag laden	20
◇	durch rechnung machen	20
◇	auf tisch kommen	20
◇	auf eis legen	20
●	auf distanz gehen	20
	auf anfrage bestätigen	20
	an anzeigenschaltern entgegennehmen	20
	am turm entgegennehmen	20
	am mittwoch teilen	20
	am freitag spielen	20
	am donnerstag finden	20
	am dienstag melden	20

Table 3.5: PNV-combinations with verb base form that occur 20 times in the extraction corpus

Other highly recurrent word combinations Here word combinations are subsumed which are frequent within a particular corpus, but not lexically determined. These word combinations may be extra-linguistically motivated as they describe conceptual aspects of the world in general like temporal and spatial situatedness of events, see the examples of temporal and spatial modification below; or they may express semantic relations that are determined by extra-linguistic facts like *in Regionalausgabe erscheint* (in local edition appears) which refers to the circumstance that the Frankfurter Rundschau has a local edition. Highly recurrent word combinations may also refer to lexical templates with particular functions, cf. sentential templates below. Other frequent word combinations are parts of collocations such as *auf Kürzungen vorbehalten* where *das Recht auf …* ('the right to do something') and *das Recht auf … vorbehalten* ('reserve the right to do something') are collocational.[6]

Temporal and spatial modification examples are *um (…) Uhr {gibt, hält, beginnt}, bis (…) Uhr geöffnet, an Anzeigenschaltern entgegengenommen am Turm entgegengenommen, im Bürgerhaus beginnt* (at o'clock {gives, holds, starts}, until o'clock open, at sales counters accepted, at the tower received, in the assembly rooms starts).

In addition, these word combinations are side effects of other, linguistically motivated word co-occurrences. The high occurrence frequency of *um (…) Uhr gibt*, for instance, results from the occurrence of *um (…) Uhr* as temporal modifier to the frequently occurring impersonal construction *es gibt* ('there is/are'), and *um (…) Uhr* in the combination *um (…) Uhr hält* is within the current corpus a preferred modifier to the noun-verb collocation *einen Vortrag halten* ('give a talk').

Am Turm entgegengenommen is a verb modifier which originates from 20 occurrences of the sentence *Geldspenden und Gutscheine werden auch an den Anzeigenschaltern im "Rundschau"-Haus am Eschenheimer Turm in Frankfurt entgegengenommen.* ("Donations and vouchers will also be accepted at the sales counters in the "Rundschau"-house at the Eschheimer tower in Frankfurt.")[7]

Sentential templates The previous sentence is like the following an example for an invariant sentence which has apart from its semantic interpretation a particular invariant function in the particular corpus, i.e., it refers to a Christmas charity of the Frankfurter Rundschau for the benefit of old people. The combination *auf Kürzungen vor_behält* originates from 74 occurrences of the sentence *Die Redaktion behält sich das Recht auf Kürzungen vor.* ('The editors reserve the right to edit contributions.'). The sentence functions is a disclaimer with respect to the letters to the editor.

Newspaper-specific combinations Examples for newspaper-specific combinations are:

---

[6] Note this word combination would be eliminated, if postnominal modifiers were taken into account. This, however, is not the case in the present study as automatic PP-attachment would lead to serious inaccuracies.

[7] The PP *im "Rundschau"-Haus* is not identified by the algorithm as the quotes are left unattached by the chunker. In order to cope with this kind of error, postprocessing of quotes is required.

PNV-combinations expressing information on the organization of the newspaper, e.g. *auf Seite* ... {*lesen, zeigen*} (at page {read, show}), *im Anzeigenteil entnehmen* ('see advertising section').

Combinations referring to statements such as *vor Journalisten sagte* ('at a press conference said'), *im Gespräch sagte* ('in a conversation said'), *am Dienstag mitteilte* (on Tuesday announced), *auf Anfrage bestätigte* (when questioned confirmed), *am* {*Montag, Dienstag, Mittwoch, Donnerstag, Wochenende*} *sagte* (on {Monday, Tuesday, Wednesday Thursday, weekend} said). In general, combinations like *am*+WEEKDAY+VERB are frequent in newspaper text; see also *am Sonntag feiert* (on Sunday celebrates), *am Samstag beginnt* (on Saturday starts).

The above PP-verb combinations are extra-linguistically motivated insofar as statements are temporally situated, directed towards someone, or a reaction to something or someone. The PP *auf Anfrage* is lexically fixed.

Errors The combinations *am* {*Mittwoch, Samstag*} *findet* and *für* {*Samstag, Donnerstag*} *lädt* are incomplete as verb prefixes are missing, *statt* in the case of *findet* and *ein* in the case of łädt. This kind of error results from a preprocessing error, namely the dot following a cardinal is misinterpreted as full stop instead of being recognized as part of a date like *9. September* where *9.* is a single token. *Im Anzeigenteil bitte* (in the classified section please), *zur Unsterblichkeit Ägypten* (to the immortality Egypt), *am Sonntag Nachmittag* (on Sunday afternoon), result from part-of-speech tagging errors, namely *Ägypten* and *Nachmittag* have been tagged as verbs instead of as nouns, and *bitte* in the given context is not a finite verb, as it has been tagged, but a particle.

Summing up, frequent word combinations that are neither identified as SVCs nor figurative expressions are grouped together. These data are largely topic-specific, and thus it is assumed that this kind of data can be helpful for topic identification and document retrieval. For more examples of pseudo-collocations extracted from the sample of the Frankfurter Rundschau see section 3.4.4.

Apart from the combinations covered by the small sample of frequent data, there are also occurrences where the verb-preposition combination is lexical, and the noun is selected according to semantic criteria. *Sorgen für*, for instance, combines with a variety of nouns such as *Aufsehen* (sensation), *Stimmung* (atmosphere), *Überraschung* (surprise), *Schlagzeilen* (head lines), *Furore* (sensation), *Aufwind* (up-drought), *Musik* (music), *Druck* (pressure), *Kinderbetreuung* (child care), *Diskussion* (discussion), and many more.

Other examples are combinations of lexicalized PPs and varying verbs like *nach Hause* {*gehen, fahren, bringen, laufen, tragen,* ...} (home {go, drive, bring, run, carry, ...}), or lexicalized PPs and semantically restricted verbs like *zu Boden* {*schlagen, stoßen, schleudern, werfen*} (to the ground {stike, push, hurl, throw}). All examples were found in the extraction corpus. The verbs in the latter case are transitive and express exertion of force. The word combinations are good examples of corpus-specific usage, as *zu Boden* can also combine with intransitive verbs like *segeln, gleiten, schweben* (sail, slide, float down) expressing gentle movement, a combination which, however, is not prominent in the newspaper corpus examined.

3.4.2 Idioms and Figurative Expressions

For the term idiom, different definitions exist. Some concentrate on semantic opacity and lexical invariance of idioms, cf. [Bußmann, 1990]. Thus the class of idioms is reduced to phraseological units with non-transparent meaning. Alternatively, idioms are classified into two types: expressions with still recoverable figurative meaning and expressions where a figurative interpretation is not possible any more, see for instance [Burger *et al.*, 1982]. The term idiom is also used as a generic term for a broad range of lexically determined constructions, cf. [Bußmann, 1990]. In the work presented, the term idiom is reserved for word combinations that are semantically opaque like *im Stich lassen* ('leave someone in the lurch') or *auf Teufel komm raus* – to want something [like the devil]–, where the meaning of *Stich* in this context is not clear, and *auf Teufel komm raus* is completely fixed. In many cases, no sharp borderline between idioms and figurative expressions can be drawn. For this reason, and because idioms are rare in the present corpus, idioms will be subsumed under figurative expressions in this work.

Figurative expressions emerge during language usage by reinterpretation of the literal meaning of a word combination, and may become conventionalized in the course of time. In addition, the process of lexicalization is also associated with restrictions in semantic compositionality and syntactic flexibility. Thus a broad range of expressions exist which vary with respect to semantic opacity and syntactic rigidity.

It is widely assumed that a relation between syntactic rigidity and semantic opacity exists, i.e., the more opaque the meaning is, the less flexible is the construction. A principled approach to the relation between the semantic and syntactic properties of collocations, however, is still missing. A rather differentiated position is advocated in [Nunberg *et al.*, 1994] who claim that the syntactic realization of idioms depends on the nature of the semantic relations among the parts of the idioms and on the meaning and discourse functions of the constructions. In order to systematically investigate these positions, automatic and flexible access to collocations in a broad range of corpora is required, as well as means for linguistically controlled representation and examination of the data retrieved, which are both topic of the current work.

In the following, examples of figurative expression occurring in the extraction corpus are given:

A major group of PNV-combinations that require figurative interpretation contains nouns that represent body parts. Note, the verb forms are normalized to bare infinitive.

(3.1) Arm (arm)

a. unter (die) Arme greifen
   (‘help somebody out with something’).

(3.2) Augen (eyes)

a. vor Augen {führen, halten}
   (‘to make something concrete to somebody’),
b. vor Augen liegen (‘be visible’/‘see’),
c. aus (den) Augen verlieren (‘lose sight of’)

(3.3) Beine, Füße (legs, feet)

a. auf (. . .) {Beine, Füße} stellen (‘to put something in motion’),
b. auf (. . .) {Beinen, Füßen } stehen (‘stand on one’s own two feet’)

(3.4) Fersen (heels)

a. auf (den) Fersen bleiben (‘be at someone’s heels’)

(3.5) Finger (finger)

a. auf (die) Finger schauen (‘keep a sharp eye on someone’)

(3.6) Gesicht (face)

a. ins Gesicht schreiben – like *etwas ist jemanden ins Gesicht geschrieben*
   (‘see something in someone’s face’),
b. zu Gesicht stehen (‘to suit someone’)

(3.7) Hand (hand)

a. in (die) Hand {bekommen, drücken, nehmen} (‘get hold of’, ‘(discretely) give’, ‘take something in hand’),
b. aus (der) Hand geben (‘to hand over’),

c. auf (der) Hand liegen (‘be obvious’)

d. in (die) Hände fallen (‘fall into someone’s hands’),

e. in (. . .) Hände kommen (‘come under the influence/control of someone’),

f. in (. . .) Händen liegen (‘be in someone’s hands’)

(3.8) Haut (skin)

a. unter (die) Haut gehen (‘get under someone’s skin’)

(3.9) Herz (heart)

a. ans Herz legen (‘enjoin someone to do something’),

b. am Herzen liegen (‘have at heart’),

c. ins Herz schließen (‘take to heart’),

d. übers Herz bringen (‘have the heart to do something’)

(3.10)Kopf (head)

a. auf (den) Kopf fallen – er ist nicht auf den Kopf gefallen
(‘he is quite smart’),

b. in (den) Kopf setzen
(‘put something into one’s head’/‘get something into someone’s head’),

c. auf (den) Kopf stellen
(‘turn things inside out’),

A characteristic feature of the examples is that the PPs are fixed, i.e., either there exists only one possible realization like *auf der Hand liegen, auf die Finger schauen, am Herzen liegen, zu Gesicht stehen* or variation is restricted see for instance example (3.3)b. for which variants such as the following were found in the corpus: *auf eigenen Beinen stehen* (on own feet stand), *auf finanziell dünnen Beinen stehen* (on financially thin legs stand), *auf wackligen Beinen stehen* (on shaky feet stand). The figurative expression *auf (. . .) Beinen stehen* is used as a metaphor for the sureness of a person, or a certain situation or particular circumstances. Accordingly, the potential for variation is determined by the semantics of *legs*, and the situation or circumstance that shall be expressed by means of the metaphor. Thus the expression *auf finanziell dünnen Beinen stehen* establishes a similarity between physical weakness expressed by ‘thin legs’ and financial weakness. A comprehensive discussion of relations which are established between different domains of experience by means of metaphor is given in [Lakoff and Johnson, 1981].

Another group of figurative expressions covers combinations of nouns for which a spatial interpretation is available with the verbs *stehen* and *stellen* like *Mittelpunkt* (center), *Vordergrund* (foreground), *Zentrum* (center), *Spitze* (top). Alternatively, there are words which combine with *stehen* but not with *stellen.* See for instance *auf dem Programm stehen* (on the programme stand, 'be in the programme'), but *auf das Programm setzen* (on the programme put, 'put in the programme') or *ins Programm nehmen* (in the programme take, 'include into the programme'); similarly we have *auf dem Spielplan stehen* (on the programme stand, 'be running') but *in den Spielplan aufnehmen* ('include into the programme') or *auf den Spielplan setzen* ('include into the programme'). These differences may result from the meaning of *Programm* which can be interpreted as 'list', and thus combine with the same verbs as *Liste* (list): *auf (...) Liste stehen* ('be on the list'), *auf (...) Liste setzen* ('put on the list') or alternatively *in (...) Liste aufnehmen* ('include into the list')

The verb pairs *stehen – stellen, stehen – setzen, stehen – aufnehmen* express noncausative-causative alternation. The causative variant introduces a new argument, namely the causer which becomes the subject of the construction. Causativity versus noncausativity is also expressed by verb pairs like *bringen – kommen* (bring – come), *legen – liegen* (lay – lie). Examples from the corpus are *über (die) Runden* {*kommen, bringen*} ('stay the course'), *ins Spiel* {*kommen, bringen*} ('come into play', 'bring into play') *in Aussicht* {*stellen, stehen*} ('promise', 'be promised'), *auf Eis* {*legen, liegen*} ('put on ice', 'be on ice'). For some combinations, only one variant exists like *unter (die) Räder kommen* ('fall into the gutter'), *zu Tode kommen* ('die'), *zum Zug kommen* ('get a chance') *im Regen stehen* ('be left out in the cold'), *auf (der) Stirne stehen* like *die Angst steht ihm auf der Stirne* ('fear is written all over his face').

Among figurative expressions there are also verbs with fixed prepositions that combine with a number of different nouns, like *ringen um* {*Identität, Macht, Kunden, Lösung, Chemiewaffenverbot, Reform, Zukunft, Sauerstoff*} (struggle for {identity, power, customers, solution, prohibition of chemical weapons, reform, future, oxygen})

Other examples of figurative expressions occurring in the extraction corpus are combinations like *ins Haus flattern* ('drop through the letter box'), *im Papierkorb landen* ('end up in the waste basket), *in den Müll wandern* ('go in the garbage'). While in the previous examples the collocations have a nominal and a verbal collocate, the combination *in Teufels bringen* is part of the larger collocation *in Teufels Küche bringen* ('bring someone into the devil of a mess'). The combination is also an example for an error in structural preprocessing, as the correct nominal dependent of *in* is *Küche.* Nevertheless, the full collocation can be identified by means of the PPs-instances which are exclusively *in Teufels Küche.* Other examples of collocations found which exceed PP-verb combinations are *Licht ins Dunkel bringen* ('cast light into something'), *Nägel mit Köpfen machen* ('to flesh out'), and the proverb[8] *die Spreu vom Weizen trennen* ('separate the wheat from the chaff'). These, however, cannot be identified by means of PNV-triples or PP-instances.

3.4.3 Support-Verb Constructions

Support-verb constructions SVCs are particular verb-object collocations constituted by a nominal and a verbal collocate, the predicative noun and the so called function verb, light verb, or support-verb. In the literature, various definitions and analyses for SVCs are presented. While the function of SVCs is commonly agreed on, i.e., they are predicates that allow Aktionsart and argument structure to vary, their syntactic realization is controversial. In particular, there is little agreement on the syntactic realization of the predicative noun. The phrase constituted by the predicative noun is either required to be a PP [Bußmann, 1990], or alternatively an accusative NP or PP [Helbig and Buscha, 1980]. [Mesli, 1991] also gives examples of SVCs with nominative, dative or genitive predicative nouns. [Yuan, 1986], in contrast, considers the question of the syntactic realization of the predicative phrase less important and instead focuses on the assumption that the predicative noun must be abstract. It can be deverbal, deadjectival or a primary noun. Similarly, [Heringer, 1968] speaks of "nomen actionis", which in his terms is a noun representing an action, event, or state, but is not necessarily deverbal. Different views are also taken with respect to the syntactic properties of SVCs. According to [Heringer, 1968] pluralization of the predicative noun is impossible in the SVC. [Polenz, 1963; Heringer, 1968; Herrlitz, 1973] do not allow articles in the PP or require the contraction of preposition and article. [Blochwitz, 1980] argues that pronominalization of the predicative noun and anaphoric reference is impossible.[9] Moreover morphosyntactic restrictions are used to distinguish SVCs from other verb-object collocations, e.g. [Helbig, 1979], which is questioned by researchers who argue that these restrictions are not a distinctive feature of SVCs but result from varying degrees of lexicalization, e.g. [Mesli, 1991; Blochwitz, 1980; Helbig, 1979; Heringer, 1968; Günther and Pape, 1976]. Such a position is also taken in the current work.

A greater consensus can be found with respect to semantic characteristics of the SVC. The support-verb is considered to be a main verb that has lost major parts of its lexical semantics and mainly contributes Aktionsart and information on causativity to the SVC, while the predicative noun contributes the core meaning. A generally acknowledged list of support-verbs, however, does

---

[8] 'A proverb is a short well-known saying which is supposed to sum up an important truth about life.' cf. the Collins English dictionary, [Col, 1996].

[9] For more examples of this kind see [Krenn and Volk, 1993].

not exist. Varying lists of SVCs are, for instance, presented in [Herrlitz, 1973; Persson, 1975; Yuan, 1986].

While idioms and figurative expressions are not restricted to noun-verb collocations, SVCs require at least a nominal and a verbal collocate. SVCs syntactically are comparable to head-argument structures where the verb is the head and the phrase containing the noun (henceforth predicative phrase) is an argument. In the current study, only SVCs containing a preposition, a noun and a verb are examined. The prepositions are also treated as collocates. Semantically, SVCs function as predicates comparable to main verbs in sentences. Thus it is not surprising that SVCs can usually be paraphrased by main verbs, e.g. *zu Besuch kommen* ≡ *besuchen* (visit) or adjective-copula constructions, e.g. *in Kraft treten* ≡ *wirksam werden* ('come into force'). Some SVCs can be used as active paraphrases of passive constructions, see for instance *zur Anwendung kommen* (SVC, active) ≡ *angewandt werden* (main verb, passive, En.: be applied). In a semantically transparent SVC the semantics of predicative noun and support-verb need to be compatible, see for instance *zu Besuch kommen* which primarily expresses a visiting-event expressed by *besuchen* the verb underlying *Besuch*, but it is also a coming-event which is expressed by the support-verb *kommen*. Syntactically, a vast majority of predicative nouns are deverbal or deadjectival, but primary nouns with argument structure can also function as predicative nouns. The noun usually combines with more than one verb. Accordingly, SVC instances with identical predicative noun can be grouped to more abstract types. An example for such a type is given in table 3.6 with the predicative noun *Betrieb* and the corresponding verbs.

preposition	predicative noun	support-verb
{in, außer}	Betrieb	{gehen, nehmen, setzen, sein, bleiben, lassen}

Table 3.6: SVC-type *Betrieb* + VERBS

Each individual combination of preposition, noun and verb constitutes an instance of the SVC type. Each instance represents a particular organization of the thematic structure (see examples 3.11, p. 78), and a particular phase of the process or state expressed by the predicate as well as causativity or noncausativity (see table 3.7). In the extraction corpus, realizations of the instances *in Betrieb* {*gehen, nehmen*} were found, when the threshold of occurrence frequency is set to three, and only full forms are used for PNV-construction. The corpus also contains realizations of *außer Betrieb* {*setzen, bleiben*} which, however, occur less than three times. The example show that a purely corpus-driven approach is infeasible for complete identification of SVC types. Nevertheless, corpora are an important resource as they provide substantial data on the actual usage of collocations, their syntactic variability and potential of modification. Examples for SVC types and their instances are given in table 3.8. The examples illustrate variation in Aktionsart (AA) and causativity (caus) as well as lexical variation. The table also shows that there are SVC types that comprise a single instance such as *in Frage kommen, in Erscheinung treten, in Anspruch nehmen.* In the case of *zur Auswahl*, a realization of the causative variant is missing in the corpus which is indicated by the brackets. *Außer Kraft treten* ('come to an end') occurs only once in the extraction corpus, thus it will not be accessible for collocation identification. SVC-instances where the verbal collocates are homophonous to auxiliaries or modals are also set in brackets.

predicative phrase	verbs	AA	caus	translation
in Betrieb	gehen,	incho	-	'go into operation'
	nehmen	incho	+	'put into operation'
	setzen,	incho	+	'start up'
	sein,	neut	-	'be running'
	bleiben,	contin	-	'keep on running'
	lassen	contin	+	'keep (something) running'
außer Betrieb	gehen	termin	-	'go out of service'
	nehmen,	termin	+	'take out of service'
	setzen,	termin	+	'stop'
	sein,	neut	-	'be out of order'
	bleiben,	contin	-	'stay out of order'
	lassen	contin	+	'keep out of order'

Table 3.7: SVC-instances of type *Betrieb* + VERBS

We distinguish four **Aktionsarten** AA: inchoative (incho, begin of process or state), terminative (termin, end of process or state), continuative (contin, continuation of process or state) and neutral (neut).[10] As already mentioned, Aktionsart in SVCs is mainly expressed by the support-verbs, but as can be seen from table 3.7, AA is not exclusively determined by the verb. In order to express inchoativity and terminativity, different prepositions are required, namely *in* for the inchoative variant, and *außer* for the terminative variant. Here the verbs *gehen, nehmen, setzen* express a change of process, but the prepositions add information on how the change has to be interpreted. **Causativity**, as already mentioned in the previous section, increases the argument structure by one. Causativity here is represented by the binary feature caus: $\{+, -\}$, where

---

[10] The distinction is taken from [Mesli, 1989] where the reader can also find a thorough discussion of Aktionsart and causativity in support-verb constructions.

causative variants are marked with ‘+’, noncausative variants with ‘−’. There are two verb pairs in table 3.7 that express causative-noncausative alternation: {*nehmen, setzen*} versus *gehen* which also express change of state or process, and *lassen* versus *bleiben*, both expressing duration. More examples for causative-noncausative alternation are given in table 3.8, i.e., *setzen* versus {*kommen, geraten, treten*}, *bringen* versus *kommen*, *stellen* versus *stehen*.

Prep	Noun	Verb	AA	caus
in	Kraft	treten	incho	-
in	Kraft	setzen	incho	+
(außer	Kraft	treten)	termin	-
außer	Kraft	setzen	termin	+
ins	Gespräch	kommen	incho	-
ins	Gespräch	bringen	incho	+
zur	Verfügung	stehen	neutral	-
zur	Verfügung	stellen	incho	+
in	Führung	gehen	incho	-
in	Führung	schießen	incho	-
in	Führung	bringen	incho	+
in	Führung	liegen	neutral	-
unter	Druck	geraten	incho	-
unter	Druck	kommen	incho	-
unter	Druck	setzen	incho	+
ins	Rollen	bringen	incho	+
ins	Rollen	kommen	incho	-
in	Frage	stellen	incho	+
in	Frage	stehen	neutral	-
in	Frage	kommen	neutral	-
zur	Auswahl	stehen	neutral	-
(zur	Auswahl	stellen)	incho	+
in	Erscheinung	treten	neutral	-
in	Vergessenheit	geraten	incho	-
in	Anspruch	nehmen	neutral	-

Table 3.8: Examples of support-verb constructions

How **argument structure** is varied by means of the verbs is shown in examples (3.11) . *Betrieb* being derived from the verb *betreiben* (run) has two thematic roles, henceforth the operator (causer) and the operand which are realized by the two NPs *die Firma* (the company) and *die Destille* (the distillery), respectively. While the main verb *betreiben* requires syntactic realization of both roles, see sentence (3.11)a., the realization of roles, i.e., their existence or their position in the the surface string, changes with the support-verb used. *Sein, gehen, bleiben* make the operand prominent, see sentences c. to e. Another way to make the operand prominent is passivization, see example b. In the case of passive transformation, syntactic realization of the operator becomes optional, in examples c. to e., however, no thematic role is available for the operator, while in the causative variants an argument position for the operator is available, see examples f. to h.

(3.11)a. die Firma betreibt die Destille
(the company runs the distillery)

b. die Destille wird (von der Firma) betrieben
(the distillery is (by the company) run)

c. die Destille ist in Betrieb
(the distillery is running)

d. die Destille geht in Betrieb
(the distillery goes into operation)

e. die Destille bleibt in Betrieb
(the distillery keeps on running)

f. die Firma nimmt die Destille in Betrieb
(the company puts the distillery into operation)

g. die Firma setzt die Destille in Betrieb
(the company puts the distillery into service)

h. die Firma läßt die Destille in Betrieb
(the company keeps the distillery running)

Summing up, SVCs function as predicates. Predicative nouns are abstract. They are typically deverbal or deadjectival, and thus have their own argument structures. The predicative noun is the semantic core of the SVC. It usually combines with more than one support-verb to allow for variation in thematic structure and Aktionsart.

These characteristics are valid for a broad range of PNV-combinations. However a number of PP-verb combinations exist that show characteristics of SVCs but are also comparable to figurative expressions, see for instance *am Anfang stehen* (at the beginning stand, 'be at the beginning'). *Anfang* is on the one hand deverbal *anfangen* (begin), on the other hand spatial interpretation is available. The figurative aspect is even more prevalent in the word combination *in den Anfängen stecken* ('be at the first stage'). Similarly *vor der Auflösung stehen* ('be in its final stages') is figurative, but can be paraphrased by the passive construction *aufgelöst werden*, where *auflösen* is the verb underlying the noun *Auflösung*. Another example is *in Kauf nehmen* ('put up with something') where the predicative noun relates to the verb *kaufen* (buy) which however does not very well fit as a paraphrase, even though a metaphoric connection can be established to some extent. From the few examples it becomes already clear that there are fuzzy borders between SVCs and figurative expressions.[11] This means, for a subset of PNV-combinations classification is fairly arbitrary. The following decisions have been made for classification of the reference data: Semantically opaque word combinations are classified as figurative expressions. In the case of semantically transparent word combinations, it is distinguished whether the nouns are abstract or concrete, and whether they contribute the main part of the semantics of the predicate. If the noun is concrete, the collocation is classified as figurative. If the noun is deverbal, deadjectival or another kind of abstract noun, and the noun contributes the major part of the meaning, the collocation is classified as SVC. Otherwise, the collocation is classified as figurative. The classification criteria are summarized in figure 3.5.

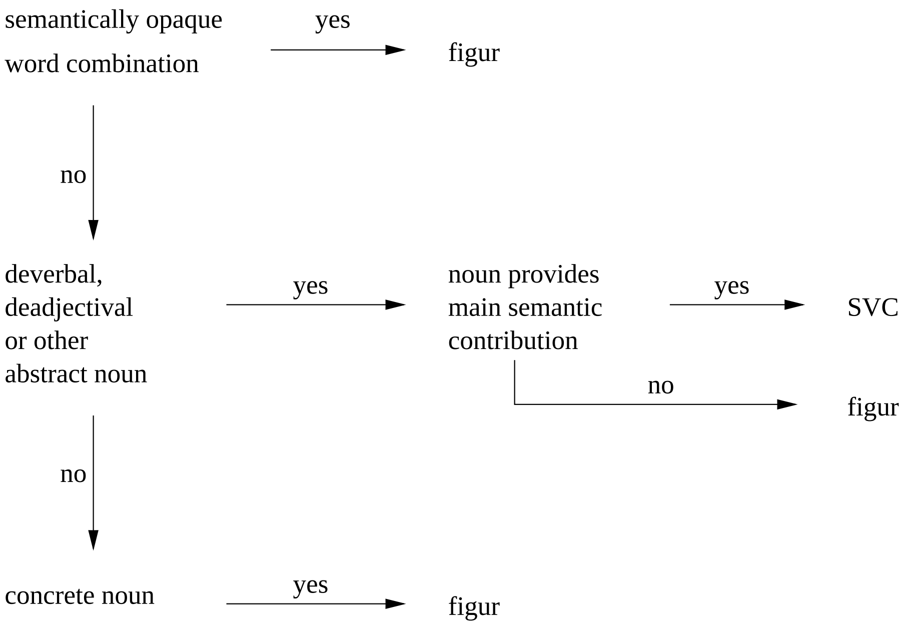

semantically opaque word combination
yes
figur
no
deverbal, deadjectival or other abstract noun
yes
noun provides main semantic contribution
yes
SVC
no
figur
no
concrete noun
yes
figur

Figure 3.5: Criteria for manual distinction of figurative expressions and SVCs

---

[11] Note, while figurative expression is a term which here is used to denote a particular group of word combinations, figurativity is a semantic property which characterizes a variety of word combinations within different groups of collocations.

3.4.4 Highly Recurrent PNV-Combinations

Considering the set highly frequent word combinations, it seems that prevalent topics in the corpus are reflected within this set, which shall be illustrated in the following. Systematic investigation, however, is topic of future research.

Two strategies for examination of frequent PNV-data have been pursued.

Strategy one: The most frequent PNV-combinations are examined. Two sets of data are compared; combinations with full forms, and combinations where the verb is reduced to its base form, i.e., information which is irrelevant for the distinction of collocations and noncollocations is abstracted away. The 500 most frequent PNV-combinations (full forms) derived from the extraction corpus are examined. The combinations range from 379 occurrences of the preposition-noun-verb triple *um Uhr beginnt* to 12 occurrences of *auf Tagesordnung standen*. The data contain 164 figurative expressions and support-verb constructions. Which means that there are potentially 336 pseudo-collocations. When using verbal base forms, the data reduce to 420 PNV-combinations, and the number of figurative expressions and SVCs reduces to 104. Thus 316 potential pseudo-collocations are among the data. For comparison, the most frequent 500 triples with verbal base forms range from 457 instances of *zur Verfügung stellen* ('make available') to 17 instances of *unter Kontrolle bringen* ('get under control'). The data contain 179 figurative expressions and SVCs, thus there could be 321 pseudo-collocations. In some cases, it may be more informative to preserve inflectional information, as morphological invariant word combinations an be useful indicators for particular text types; see for instance *unter Berufung berichtete* ('referring ... to informed'), *im Alter* {*gestorben, starb*} ('at the age ... died'), *nach Angaben getötet* ('according to ... killed') which are typical for press and news reporting. To what extent particular word combinations are distinctive for text types or domains need to be investigated on the basis of reference texts. The study aims at providing powerful methods and tools for such a task.

Strategy two: Lexical vectors are constructed. On the one hand, each PN-combination is associated with the co-occurring verbs (verb-vectors), on the other hand VP-combinations are associated with the co-occurring nouns (noun-vectors). In both cases, verb forms are reduced to their base forms. PNV-combinations and lexical vectors differ with respect to the information they represent. The former represent individual collocations while the latter also provide insights into the lexical company, and thus the semantic range of words. While individual PNV-combinations are well suited for identification of collocation instances, lexical vectors support identification of groups of collocations. Accordingly, sequential application of both strategies is useful. Thus verb and noun vectors are created from the 500 most frequent PNV-combinations with verbal base forms. Based on the vectors the PNV-triples can be grouped according to their thematic associations, examples of which are given in the following:

Times and Dates The most frequent PN-combinations express times and dates like {*um, bis*} *Uhr* ({at, until} o'clock), *am* (on) + WEEKDAY. The verbs refer to:

- Events like *beginnen* (start), *eröffnen* (open), *stattfinden* (take place), *aufführen* (perform), *spielen*, (play), *töten (kill)*. The "event"-related verbs typically refer to cultural events like *aufführen* and *spielen* with the exception *töten*.
- Utterances or announcements like *sagen* (say), *mitteilen* (inform), *berichten* (report), *erklären* (explain).

The data suggest, that announcements, cultural events and death are high ranking topics in present corpus.

Culture Frequent examples of word combinations related to cultural events are

- auf Bühne stehen ('be on stage'),
- aus Buch lesen ('read from book'),
- unter Leitung spielen ('to play conducted by').

Death *töten* (kill) and *sterben* (die) are two frequent verbs relating to death. Noun vectors reveal the preferred accompanying PN-combinations.

- *töten* – {*nach Angaben, am Montag*} ({'according to', 'on Monday'}), represent the two prevalent usages in the extraction corpus, namely reference to the source information indicated by *nach Angaben*, and reference to the date of the event, e.g. *on Monday*.
- *sterben* – {*im Alter, von Jahren, an Folgen*} ({'at the age', 'years ago', 'because of'}). With respect to these examples, it is worth noting that the identical rank of *im Alter sterben* and *von Jahren sterben*, see below, is a weak indicator that the two combinations are parts of a single-word combination which is *im Alter von X Jahren sterben* ('die at the age of X years') where *X* is a cardinal. Further evidence for the hypothesis is that in both cases the verbs show exactly the same realizations, see below. The verb full forms also reveal that the particular combination occurs only in past tense.

PNV-combination	rank	verb forms
von_APPR jahren_NN sterben_VV	64	gestorben_VVPP starb_VVFIN
im_APPRARTd alter_NN sterben_VV	64	gestorben_VVPP starb_VVFIN

Rescue and Life Saving

- in(s) Krankenhaus – {einliefern, bringen} (‘take to the hospital’)
- am Leben erhalten (‘keep alive’)

are the prevalent examples in the corpus.

Service Information Another large group of frequently occurring word combinations relates to services concerning the distribution of the newspaper, the location of information within the newspaper, and information on phone services

- Distribution of the newspaper: The following word combinations are another example for parts of a single collocation.

zur_APPRARTd zeitungszustellung_NN wenden_VV	28	wenden_VVFIN
in_APPR fragen_NN wenden_VV	28	wenden_VVFIN
an_APPR vertriebsabteilung_NN wenden_VV	28	wenden_VVFIN

  The data originate from the recurrent sentence *In allen Fragen zur Zeitungszustellung wenden Sie sich bitte an unsere Vertriebsabteilung.* (‘With respect to all questions concerning the delivery of the newspaper please contact our sales department.’) The sentence is comparable to phrasal templates, cf. page 24. Accessibility of the underlying sentences is required for proper identification of this kind of stereotypic combinations which may be useful for locating certain sections in text.

- Newspaper-internal information such as specification of the location of information within the newspaper
  auf Seite {*lesen, zeigen, entnehmen, stehen*}
  (at page {read, show, ‘learn from’, ‘say’),
  {*auf Freizeitseite, in Abendausgabe*} {*zeigen, lesen*}
  ({in supplement page, in evening edition} {show, read}),
  im Anzeigenteil entnehmen
  (‘learn from the advertisement pages’),
  {*in Regionalausgabe, Stadtteil-Rundschau*} *erscheinen*
  ({in regional edition, neighbourhood news} be published).

- Phone-numbers of services
  unter Telefonnummer {*gibt, zu erreichen, entgegennehmen, anmelden*}
  (unter phone number {'there is/are', to reach, take),
  unter {*Tel., Telefonnummer, Telefon*} *gibt*
  (under {tel., phone number, phone} there is/are),
  unter {*Telefonnummer, Telefon*} *zu erreichen*
  (under {phone number, phone} to reach)

  The examples are rather homogeneous, i.e., there is no variation of the preposition *unter*; all nouns occurring are variants of the word phone; interestingly, particular variants of phone combine with particular verbs.

Justice The following word combinations relate to jurisdiction.

- accusation
  vor Gericht stellen ('put on tiral'),
  im Verdacht stehen ('be suspected of')

- conviction
  verurteilen – zu {*Freiheitsstrafe, Haftstrafe, Haft*}
  (to sentence – to prison, arrest, arrest),
  zum Tode verurteilen ('to sentence to death'),
  auf Bewährung verurteilen ('to sentence on probation'),
  vom Landgericht verurteilt(werden)[12] ('to sentence by the Superior Court'),
  am Mittwoch verurteilen ('to sentence on Wednesday');
  unter Strafe stellen ('to punish')

- custody
  sitzen – {*im Gefängnis, in Untersuchungshaft*} ('be in jail', 'be on remand')

- release
  zur Bewährung aussetzen ('to release on probation')
  auf (freien) Fuß setzen ('to release (from jail)')

Dissemination of Information Prevalent expressions in the extraction corpus are:

- fixed PPs in combination with variable verbs such as

  auf Anfrage – {sagen, erklären, bestätigen, mitteilen}
  (on questioning – say, explain, confirm, impart),
  nach Angaben – { töten, geben, festnehmen, handeln, kommen}

---

[12] This is another example for invariant usage of word combinations in a particular corpus.

(according to – kill, there is, arrest, it refers to, come),
unter Berufung – {berichten, melden}
(refering to – report, announce),
aus Kreisen verlauten ('informed sources suggest ...')

- *heißen in* plus noun

  heißen in – {*Begründung, Aufruf, Erklärung, Bericht, Mitteilung, Schreiben, Pressemitteilung*}
  (it says in – statement of arguments, proclamation, declaration, report, communiqué, letter, press briefing)

- Verbs of utterance and co-occurring PPs like

  sagen – {*in Interview, im Rundfunk, im Deutschlandfunk, vor Journalisten, im/in Gespräch, im Fernsehen*}
  (say – in interview, in the news, in the Deutschlandfunk, at a press conference, in the/in conversation, on television)

While the fixed PPs and *heißen in* are typical for news speak, and thus are expected to have a strong potential as keys for the identification of news relevant sentences, the verbs of utterance get their relevance as identifier for news in combination with the PPs.

Politics and Business Organization The topics are grouped together as the vocabulary presented in the following for the most part can be used in both domains.

wählen – {zum Vorsitzenden, zum Präsidenten, für Jahre}
(elect – {for chairman, for president, for years})
im Amt – {bleiben, bestätigen} (stay in office, confirm a person in office)
sitzen – im {Stadtparlament, Parlament} (sit – in the {city parliament, parliament})
vertreten – im {Parlament, Ortsbeirat} ('be represented' – in the {parliament, local advisory board})
zum Rücktritt auffordern ('ask for resignation') über Mehrheit verfügen ('have the majority') in Sitzung beschließen ('decide in the meeting') in Resolution heißen ('the resolution says') auf Tagesordnung stehen ('be on the agenda') zur Jahreshauptversammlung {laden, treffen} (to the annual meeting {invite, meet}) in Kraft {treten, setzen} ('come into force', 'bring into force'), außer Kraft setzen ('to annul', 'to invalidate')

Sports The following expressions can be distinguished.

- placement
  auf Platz kommen ('be placed'),
  auf Platz folgen (on place follow),
- lead
  nach Minuten führen (after minutes lead)
  in Führung {gehen, bringen} ('take the lead', 'put someone ahead')
- start of race or competition
  an Start gehen ('to start')
  ins Rennen {schicken, gehen} ('to start')
  This kind of word combination is also used in politics to express participation of a politician in an election campain being another example for metaphoric use of expressions from one domain in another one.

Other prominent domains in newspaper text are catastrophies and desasters, traffic accidents and crime. The following word combinations are frequent:

- Catastrophies, desasters
  in Flammen aufgehen ('go up in flames')
- Traffic accidents
  ins Schleudern geraten ('to skid'),
  von Fahrbahn abkommen ('come off the carriageway')
- Crime
  flüchten {zu Fuß, in Richtung} (flee {on foot, in direction})
  auf Spur kommen ('get on the track of')
  im Wert stehlen ('steal something worth . . . ')

3.4.5 Frequency Distributions according to Collocation Classes

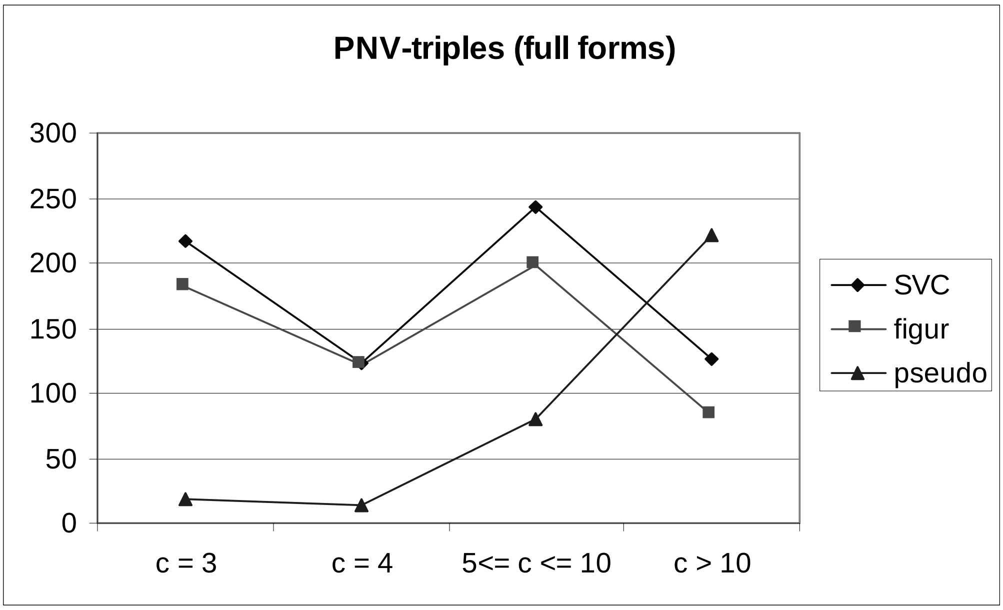
The number of collocations within the groups of PNV-combinations is fairly small, see table 3.9. In all cases, there is a strong decline in collocation density from the groups of frequent word combinations where $c \geq 10$ to infrequent ones. The data also support the claim made in [Breidt, 1993] that collocation density decreases when base forms are considered instead of full forms. In the case of full forms, SVCs and figurative expressions are rather evenly distributed over low frequency and high frequency word combinations, i.e., the set where $c \geq 5$ contains 369 SVCs, the sets where $c = 3$ and $c = 4$ together contain 340 SVCs.

The number of figurative expressions are 282 for the former set and 304 for the latter. The picture is different for pseudo-collocations, where approximately two thirds of the data occur among the highly recurrent word combinations where $c > 10$. This is also the case for PNV-triples with reduced verb forms. The data containing verb bases also show a clear tendency of SVCs and figurative expressions to occur among the more frequent data, i.e., 72 % of the SVCs and 62 % of the figurative expressions occur in the set of PNV-combinations where $c \geq 5$.

	full forms				verb bases			
rank	SVC	figur	pseudo	PNV total	SVC	figur	pseudo	PNV total
$c = 3$	217	182	21	4774	72	104	18	6358
	4.5	3.8	0.4	4774	1.1	1.6	0.3	6358
$c = 4$	123	122	14	2792	57	97	17	4585
	4.4	4.4	0.5	2792	1.2	2.1	0.4	4585
$5 \leq c \leq 10$	243	199	80	2235	166	220	72	3643
	10.9	8.9	3.6	2235	4.6	6.0	2.0	3643
$c > 10$	126	83	222	629	173	139	275	1097
	20.0	13.2	35.3	629	15.8	12.7	25.1	1097
$\sum$	709	586	337		468	560	382	

Table 3.9: Number and percentage of collocations according to collocation class and occurrence frequency

Summing up, pseudo-collocations do not change with respect to their frequency distributions when verb forms are reduced to their bases. SVCs and figurative expressions on the contrary do. The number of SVCs among full form triples is slightly higher than the number of figurative expressions. The frequencies change, when verb forms are reduced. For illustration see figure 3.6. Figurative expressions outnumber SVCs in the groups where $c = 3$, $c = 4$, $5 \leq c \leq 10$. Thus it can be concluded that there is more inflectional variation of the verbs in SVCs than in figurative expressions. In the current data, the average number of types of verb forms in SVCs is 2.7, and 2.2 in figurative expressions. The discrepancy is even more pronounced considering the total number of SVCs and figurative expressions: While there is a reduction of 34 % of the SVC-types from full forms to base forms, there is only a reduction of 4 % in the case of figurative expressions. The deviation in group $c > 10$, where SVCs outnumber figurative expressions, even though the verbs are reduced to base forms, suggests that in general individual SVCs are more frequently used than individual figurative expressions.



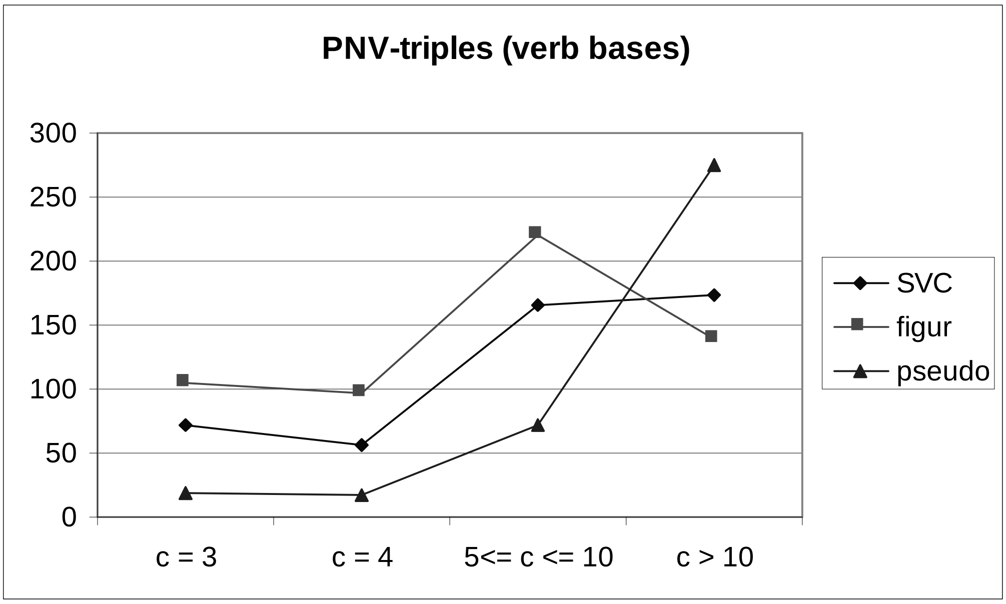


Figure 3.6: Distribution of SVCs, figurative expressions and pseudo-collocations within full form and base form data

3.5 Conclusion

Distribution of PNV-Combinations in Text: A well known characteristic of natural language is that words in texts are distributed comparable to Zipf's law. In other words, a vast majority of words occur only once in a corpus, whereas a very small number of words are highly frequent. As a consequence, only a small percentage of word-combinations in texts can be used for statistics-based collocation identification. This circumstance is exemplified by the frequency distribution of combinations containing a preposition (P), a noun (N) and a main verb (V). The PNV-triples are identified from the so called extraction corpus, an 8 million word sample selected from the Frankfurter Rundschau corpus. The sample has been automatically part-of-speech tagged and annotated with phrase chunks. While 87 % of the PNV-combinations (full forms) occur once in the extraction corpus, only 3 % occur three times or more, and only 6 % of this small sample occur more than 10 times. On the other hand, frequent word combinations cover comparably large portions of running text. For example, the 617 PNV-combinations (full forms) which occur more than 10 times in the extraction corpus cover 43 080 word tokens in running text. In other words, 6 % of the PNV-combinations that occur 3 times or more in the extraction corpus cover over 26 % of the tokens that can be covered by this group. By reducing the verbs to their base forms, the occurrence frequencies of the data relevant for PP-verb collocations can be increased without losing important collocation-specific information. This is due to the fact that PP-verb collocations are flexible with respect to the verbal collocate but rigid with respect to the PP-collocate. In particular, SVCs show strong inflectional variation in the verbal collocate. Thus larger numbers of word tokens in running text are covered by a smaller number of collocation types. A drawback, however, is that collocation density declines in the case of morphologically reduced data.[13]

Collocations Identified: Two major groupings of lexically determined combinations could be identified from the set of PNV-combinations, namely combinations where two elements are lexically selected, and combinations where preposition, noun and verb are lexically determined.

The first class comprises:

- Combinations where the verb-preposition combination is lexical, and the noun is selected according to semantic criteria; see for instance *sorgen für* which combines with a variety of nouns like *Aufsehen* (sensation), *Stimmung* (atmosphere), *Überraschung* (surprise), *Schlagzeilen* (head lines), *Fu-*

---

[13] The inverse behaviour of recall (number of collocations) and precision ($\frac{number\ of\ collocations}{sample\ size}$) with respect to collocations has also been stated in [Breidt, 1993].

rore (sensation), *Aufwind* (up-drought), *Musik* (music), *Druck* (pressure), *Kinderbetreuung* (child care), *Diskussion* (discussion).

- Combinations of lexicalized PPs and varying verbs like *zu Boden* {*schlagen, stoßen, schleudern, werfen*} (to the ground {stike, push, hurl, throw}) or *nach Hause* {*gehen, fahren, bringen, laufen, tragen, ...*} (home {go, drive, bring, run, carry, ...}).

The second class comprises:

- Support-verb constructions like *zur Verfügung* {*stehen, stellen, haben*} *('be at ones disposal', 'make available', 'have at ones disposal'), ins Gespräch* {*kommen, bringen*} ('engage in conversation with someone', 'open up a conversation').
- Figurative expressions like *am Herzen liegen* ('have at heart'), *unter die Lupe nehmen* ('have a close eye on someone or something')

In addition, PNV-combinations have been found among the higly recurrent ones which are neither SVCs nor figurative expressions. Most of them can be related to a certain topic or domain present in the particular extraction corpus, see for instance *unter Berufung berichtete* ('referring to ... reported') *aus Kreisen verlautete* ('informed sources suggested') which are typical for news speak. Other examples are combinations related to jurisdiction like *zum Tode verurteilen* ('sentence to death'), *zur Bewährung aussetzen* ('release on probation'), and many more. For this group of word combinations, the term pseudo-collocation has been introduced.

While general language collocations are rather evenly distributed over high and low frequency word combinations in a corpus, high occurrence frequency is an indicator for corpus-specificity of the particular word combination. Accordingly the three classes of PNV-collocations manually identified in the extraction corpus can be grouped as follows: SVCs and figurative expressions are either general language collocations or corpus-specific ones. The actual partition, however, needs to be determined in comparison with corpora from other domains. The remaining highly frequent word combinations, the pseudo-collocations, in contrast, are assumed to be corpus specific.

Chapter 4

Corpus-Based Collocation Identification

4.1 Introduction

Linguistically motivated strategies and statistical techniques for corpus-based collocation identification are discussed in this chapter. In section 4.2, the suitability of numeric and syntactic spans for accessing PP-verb collocations is examined. Based on a number of experiments, it is argued that corpus-based identification of PNV-collocation candidates leads to more appropriate results when collocation relevant linguistic information is taken into account during construction of the candidate data. In order to replace numeric by syntactic spans, the extraction corpus needs to be part-of-speech tagged and structurally annotated which has been topic of the previous chapter.

Linguistic constraints important for selecting candidate data for PNV-collocations, and their frequency distributions in the extraction corpus are discussed in section 4.3.

In section 4.4, computational methods are presented which account for the three defining characteristics of collocations as stated in section 1.5 which are: (i) over proportionally high recurrence of collocational word combinations compared to noncollocational word combinations in corpora; (ii) grammatical restrictions in the collocation phrases; and (iii) lexical determination of the collocates of a collocation.

Experiments for testing the feasibility of the different approaches for collocation identification are presented in the next chapter.

4.2 Lexical Tuples: Numeric versus Syntactic Span

A major characteristic of collocations is their lexical determination which is standardly dealt with by means of n-gram frequencies calculated on the basis of word combinations found within certain numeric spans. A numeric span covers words w_j to the left and/or right of a base word w_i within a particular distance r, i.e., the span delimits the lexical context within which collocation partners w_i w_j are to be found, with $|j - i| + 1 \leq r$. For the following reasons, numeric spans are a poor basis for collocation identification. If the span size is kept small, it is unlikely to properly cover nonadjacent collocates of structurally flexible collocations. Enlarging the span size, on the other hand, leads to an increase of candidate collocations including an increase of noisy data which need to be discarded in a further processing step. Another source of noise within the set of collocation candidates is due to the over-proportional frequency of function words within texts. This problem can be avoided by excluding function words from the construction of lexical tuples which is unproblematic as function words are nonproductive and thus easy to enumerate in so called stop word lists, a strategy which is widely used. Numeric spans are also insensitive to punctuation. Punctuation, however, is a suitable delimiter of word sequences containing syntactically motivated collocations. Using a sentence as the largest unit within which the collocates of a collocation may occur, as it is the case in the current study, is a first step in reducing the number of syntactically implausible collocation candidates. In addition, a large number of syntactically invalid n-grams is excluded beforehand, as parts-of-speech are known. Further improvement of the appropriateness of the collocation candidates selected is achieved by the availability of structural information, step by step replacing the numeric by a syntactic span. In the following, three experiments are described which illustrate the advantages of accessibility of syntactic information for collocation identification.

4.2.1 Extraction Experiments

Lexical tuples are selected from the extraction corpus varying the accessibility of linguistic information according to the following three strategies. For comparison, the 20 most frequent tuples resulting from employing numeric and syntactic spans are examined.

Strategy 1: Retrieval of n-grams from word forms only (w_i).

Bi- and tri-gram frequencies are calculated. For identification of preposition-noun co-occurrences, the numeric span is restricted to four, i.e., for each word form w_i, $i \in 1 \cdots n - 3$ in a corpus of size n, bi-grams with the three right neighbours are constructed resulting in the pairs $\langle w_i, w_{i+1}\rangle$, $\langle w_i, w_{i+2}\rangle$, $\langle w_i, w_{i+3}\rangle$. A span size of four is considered to be large enough to cover the majority of PPs assuming [P DET N], [P ADJ N], [P DET ADJ N], [P ADJ ADJ N], [P ADV ADJ N], where P stands for preposition, DET for determiner, ADJ for adjective, ADV for adverb and N for noun. In order to cover PP-verb collocations, frequencies of $\langle w_i, w_j, w_k\rangle$-triples are calculated where $j = i+1, i+2, i+3$ and $k = j+1$, i.e., the potential noun w_i and the potential verb w_j are considered to be adjacent, thus leading to a maximum span size of five words. This approach allows the span size and thus the amount of noisy data to be kept small. It covers, on the one hand, a broad range of PPs, and on the other hand takes advantage of the preference of PP- and verb collocates to be adjacent in verb final constructions.

Strategy 2: Retrieval of n-grams from part-of-speech annotated word forms (wt_i).

N-gram frequencies are calculated similarly to strategy 1, but using word-tag pairs wt_i instead of plain word forms. Thus co-occurrence frequencies are calculated for preposition-noun and preposition-noun-verb combinations resulting in

> $\langle wt_i, wt_{i+1}\rangle$, $\langle wt_i, wt_{i+2}\rangle$, and $\langle wt_i, wt_{i+3}\rangle$-pairs, where t_i is a preposition and $\{t_{i+1}, t_{i+2}, t_{i+3}\}$ are nouns, and
>
> $\langle wt_i, wt_j, wt_k\rangle$-triples, where t_i represents a preposition, t_j a noun and t_k a verb.

In strategy 1, linguistic knowledge has been implicitily used by applying syntactically motivated restrictions to span size and word positions. In strategy 2, the numeric span is enhanced with part-of-speech information.

Strategy 3: Retrieval of n-grams from word forms with particular parts-of-speech, at particular positions in syntactic structure (wt_ic_j).

N-gram frequencies are calculated exploiting the structural information provided by the chunk tags which are described in section 2.2.2. Numeric spans are entirely replaced by syntactic spans. Preposition-noun bi-grams are constructed only in these cases where preposition and noun are syntactic dependents. See the structure in figure 4.1 which contains two appropriate PN-tuples namely *von Gruppen*, and *am Wettkampf*.[1] PP-verb co-occurrences are identified as described in section 3.3.1.

---

[1] Recall, the part-of-speech tags are NN for noun, ART for article, APPR for preposiotion and APPRART for a fusion of preposition and determiner. The motivation for PP structures where preposition, article, adjectival modifier (*am Wettkampf beteiligen*) and noun are sisters is given in [Skut *et al.*, 1998].

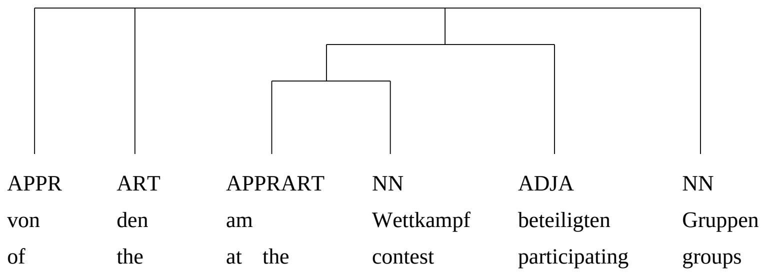


Figure 4.1: Example of a chunk containing two PN-tuples – *von Gruppen, am Wettkampf*

4.2.2 Results

Results of Strategy 1

Retrieval of PP-verb collocations from word forms only is clearly inappropriate as function words like articles, prepositions, conjunctions, pronouns, and cardinals outnumber content words like nouns, adjectives and verbs. As already mentioned, a commonly applied remedy are stop word lists which are used for excluding function words from processing. This strategy, however, leads to the loss of collocation-relevant information, as accessibility of prepositions and determiners may be crucial for the distinction of collocational and noncollocational word combinations. See for instance the PPs *in Betrieb* versus *im Betrieb* where the former is most likely the predicative phrase of the SVC *in Betrieb* {*gehen, setzen, nehmen, sein*} ('go into operation', 'start up', 'put into operation', 'be operating'), while the latter is a noncollocational word combination meaning 'in the enterprise'. Similarly the PP *zu Verfügung* is a predicative phrase referring to the SVC *zu Verfügung* {*stehen, stellen, haben*} ('be available', 'make available', 'have at one's disposal'), whereas the NP *eine Verfügung* (an injunction) is noncollocational.

Table 4.1 shows the 20 most frequent word bi-grams for each $\langle w_i, w_j \rangle$-pair with $j \in i+1, i+2, i+3$ derived from the extraction corpus. When no stop words are excluded, the bi-grams mainly consist of co-occurrences of function words. There are only two potential PPs – *bis* ... *Uhr* (until ... o'clock), *um* ... *Uhr* (at ... o'clock).

The majority of $\langle w_i, w_{i+1} \rangle$-combinations (14 out of 20) represent initial sequences of PPs comprising a preposition followed by an article. Note, the bi-gram *30 Uhr* is an indicator that the tokenizer used does not properly account for time expressions. Thus expressions like *20 : 30 Uhr* are split in four tokens – *20, :, 30, Uhr* – which is an additional source of noise in n-gram construction. The problem, however, can be easily solved by extra processing of time constructions.

Occurrences of *Uhr* as part of time expressions explain the high frequency of this particular noun in the extraction corpus.

w_i	w_{i+1}	freq
in	der	24412
für	die	11322
in	den	9733
30	Uhr	8352
20	Uhr	7090
in	die	5988
und	die	5949
mit	dem	5891
von	der	5785
auf	die	5474
an	der	5427
mit	der	5214
bei	der	5076
für	den	4748
sich	die	4666
über	die	4469
und	der	4419
daß	die	4381
auf	dem	4263
aus	dem	4225

w_i	w_{i+2}	freq
bis	Uhr	12023
die	der	11617
der	der	6957
der	und	5679
die	des	5529
der	in	5047
die	in	4986
die	von	4323
der	des	4249
die	und	3966
den	der	3813
von	und	3702
der	von	3300
und	der	3070
für	und	2932
zwischen	und	2750
und	in	2689
von	bis	2664
Die	der	2658
um	Uhr	2631

w_i	w_{i+3}	freq
der	der	7035
die	der	6153
(	)	6037
der	die	6001
die	die	5558
und	die	3891
und	der	3889
der	und	3715
der	in	3558
die	und	3500
10	Uhr	3330
die	in	3155
die	den	2924
in	der	2903
der	den	2606
den	der	2266
in	und	2263
und	in	2254
den	die	2095
für	der	1713

Table 4.1: ⟨Word,word⟩-bi-gram frequencies identified from the tokenized extraction corpus

$\langle W_i, w_{i+2}\rangle$-combinations are indicators for more complex syntactic structures such as NPs with a genitive modifier to the right, NP relative clause (S$_{rel}$) sequences or even two independent NPs in the middle field. Whereby the later is less probable for article-noun-article sequences when the second article can be interpreted as a genitive. There are six potential NP NP$_{gen}$ instances (*die der, der der, die des, der des, den der, Die der*)[2] four of which may also correspond to NP S$_{rel}$ structures, as articles and relative pronouns are lexically identical in German. An exception is *des* which only occurs as genitive article. Similarly *die von, der von* are most likely part of complex NPs with *von* indicating a pseudo-genitive[3]. Five bi-grams in the sample are related to co-ordinations involving NPs

---

[2] *der, die, den* (the), *des* (of the)

[3] Pseudo-genitives are PPs with the preposition *von* that function like genitive modifiers, e.g. *die Federn* {*[des Vogels]*$_{NP_{gen}}$, *[von dem Vogel]*$_{PP_{von}}$} (the feathers of the bird)

or PPs, see *der*-(noun)-*und* , *die*-(noun)-*und*, or *von*-(noun)-*und*, *für*-(noun)-*und*, *zwischen*-(noun)-*und*.

Similarly, seven of twenty $\langle w_i, w_{i+3}\rangle$-combinations represent NP NP or NP Srel sequences. Five combinations relate to co-ordinations involving an NP or PP. An interesting bi-gram is () suggesting that a large number of insertions in the extraction corpus consist of two words. Except for (), n-grams containing punctuation have been omitted from the example lists as punctuation is not part of collocations.

In the case of word triples, the following interpretations are possible: The set of the 20 most frequent $\langle w_i, w_{i+1}, w_{i+2}\rangle$-combinations contains temporal phrases such as *bis 17 Uhr* (until 17 o'clock), *Di. bis Fr* (Thuesday until Friday), *in diesem Jahr* (during this year); parts of temporal phrases like *10 bis 17* (10 until 17), *in den vergangenen* (in the past (years, months, ...)); and fixed word combinations such as the city name *Frankfurt a. M.*, or *Tips und Termine* (tips and dates) which is a headline in the information section of the Frankfurter Rundschau. See table 4.2.

w_i	w_{i+1}	w_{i+2}	freq	w_i	w_{i+1}	w_{i+2}	freq
bis	17	Uhr	2222	Di.	bis	Fr	807
bis	18	Uhr	2081	10	bis	17	779
bis	20	Uhr	1370	Uhr	in	der	768
bis	12	Uhr	1350	Di.	bis	So	733
bis	19	Uhr	1098	9	bis	12	717
FRANKFURT	A.	M.	949	bis	13	Uhr	713
in	diesem	Jahr	915	10	bis	12	605
bis	16	Uhr	889	Tips	und	Termine	597
bis	14	Uhr	864	in	den	vergangenen	583
um	20	Uhr	855	in	der	Nacht	582

Table 4.2: ⟨Word,word,word⟩-tri-gram frequencies identified from the tokenized extraction corpus

The $\langle w_i, w_{i+2}, w_{i+3}\rangle$-combinations relate to NP PP sequences such as *die ... in der* (the ... in the), co-ordinations (see the examples containing *und* (and)); time expressions like *9 ... 12 Uhr* where 483 instances relate to the sequence *9 bis 12 Uhr* (9 until 12 o'clock), *in ... Nacht zum* with 416 instances referring to *in der Nacht zum*, and 22 instances referring to *in jener Nacht zum*. All 438 instances are followed by a date expression. There are also instances of time-place combinations like *um ... Uhr in* (at ... o'clock in). The triple *Uhr Tel. 0* (o'clock phone 0) typically relates to sequences from the advertising section like *... Uhr, Tel. 0 61 72 / 71* or *... Uhr unter Tel. ...*. See table 4.3.

w_i	w_{i+2}	w_{i+3}	freq	w_i	w_{i+2}	w_{i+3}	freq
die	in	der	1102	9	12	Uhr	493
der	in	der	1018	10	20	Uhr	491
der	und	der	835	der	für	die	487
10	17	Uhr	780	10	12	Uhr	478
die	für	die	683	10	13	Uhr	458
die	in	den	580	um	Uhr	in	438
um	Uhr	im	578	in	Nacht	zum	438
bis	Uhr	und	547	15	18	Uhr	435
und	in	der	546	Die	und	das	427
die	und	die	539	Uhr	Tel.	0	416

Table 4.3: ⟨word,word,word⟩-tri-gram frequencies identified from the tokenized extraction corpus

$\langle W_i, w_{i+3}, w_{i+4}\rangle$-combinations are harder to interpret as they are less homogeneous because of the larger span they cover. Some of the data relate to NP PP sequences and co-ordinations. See table 4.4.

w_i	w_{i+3}	w_{i+4}	freq	w_i	w_{i+3}	w_{i+4}	freq
die	in	der	650	die	in	den	347
und	in	der	525	die	für	die	332
Sa.	bis	14	431	und	für	die	289
Die	das	Biest	404	von	18	Uhr	288
um	in	der	400	der	für	die	284
die	und	die	371	der	in	den	283
und	Tips	und	369	März	20	Uhr	274
der	und	der	369	und	in	den	254
Termine	und	Termine	369	1	Millionen	Mark	225
Mi.	bis	20	357	mit	in	der	223

Table 4.4: ⟨Word,word,word⟩-tri-gram frequencies identified from the tokenized extraction corpus

Results of Strategy 2

According to strategy 2, co-occurrence frequencies are only calculated for lexical tuples with appropriate parts-of-speech. The resulting 20 most frequent preposition-noun combinations are listed in table 4.5.

The $\langle wt_i, wt_{i+1}\rangle$-bi-grams can be grouped into the following classes:

- Arbitrary preposition-noun co-occurrences such as *am Samstag* (on Saturday), *am Wochenende* (at the weekend), *für Kinder* (for children), *im Rathaus* (in the town hall), *im Bürgerhaus* (in the assembly rooms) a subset of which, however, is typical for newspaper text, recall section 3.4.4.
- Fixed PPs such as *zum Beispiel* (for example).
- PPs with a strong tendency for particular continuation such as *nach Angaben* ('according to'), *im Jahr* (in the year).
- PP-collocates of verb-object collocations such as *zur Verfügung* (at the disposal).

wt_i	wt_{i+1}	freq
am	Sonntag	1865
am	Montag	1803
am	Dienstag	1698
am	Freitag	1675
am	Mittwoch	1669
am	Samstag	1662
am	Donnerstag	1564
zur	Verfügung	935
für	Kinder	866
nach	Angaben	775
zum	Beispiel	758
am	Wochenende	597
im	Bürgerhaus	539
im	Jahr	533
Nach	Angaben	523
zum	Thema	507
im	Rahmen	507
zur	Zeit	498
im	Rathaus	496
am	Ende	490

wt_i	wt_{i+2}	freq
bis	Uhr	12023
um	Uhr	2631
im	Jahr	1276
ab	Uhr	978
in	Jahr	972
vor	Jahren	825
seit	Jahren	812
in	Nacht	624
in	Stadt	569
in	Zeit	518
von	Mark	428
auf	Straße	424
auf	Weg	418
zum	Mal	411
um	Prozent	406
in	Bundesrepublik	372
aus	Gründen	359
in	Nähe	350
von	Millionen	342
an	Stelle	334

wt_i	wt_{i+3}	freq
von	Mark	785
in	Jahren	756
in	Straße	423
im	Jahres	299
mit	Mark	293
auf	Mark	267
in	Woche	253
in	Bundesländern	236
in	Ländern	220
in	Tagen	199
für	Mark	197
in	Monaten	196
um	Prozent	195
für	Kinder	195
in	Wochen	190
von	Millionen	180
auf	Seite	179
seit	Jahren	177
in	Kirche	168
in	Sitzung	162

Table 4.5: ⟨Preposition,noun⟩-bi-gram frequencies identified from the part-of-speech tagged extraction corpus

A native speaker of German would expect *nach Angaben* to be followed by either an NP_{gen} or a pseudo-genitive realized as PP_{von}. These expectations are clearly supported by the corpus: 722 of 775 *'nach Angaben'*-instances are immediately followed by an NP_{gen} (549 instances) or a PP_{von} (173 instances).

Comparable results are found with respect to *Nach Angaben*, which indicates that the behaviour of the PP does not change at sentence-initial position. There is also a strong expecation for the category following *im Rahmen* ('within the scope') which is confirmed by the corpus data, i.e., in 432 of 507 instances NP_{gen} is immediatley following. In the case of *im Jahr*, the corpus data are less biased. There are 213 instances of 533 total followed by a cardinal. Even though these are less than half of the cases, the tendency is clear as there is a big gap between the most frequent right adjacent category and the next one which are 34 instances of finite verbs. *Zur Verfügung* is the PP-collocate of a support-verb construction, and thus establishes a lexical expectation for the co-occurring verb.

A specific characteristic of $\langle wt_i, wt_{i+2}\rangle$-pairs is their tendency to cover PPs with pre-nominal modification (wt_{i+1}). Cardinal, for instance, is the most probable modifier category co-occurring with *bis ... Uhr* (12020 of 12023 cases total) and *um ... Uhr* (2574 of 2631 cases) like {*um, bis*} *10 Uhr* ({at, until} 10 o'clock). Adjective is the predominant modifier category related to *im ... Jahr* (1272 of 1276 cases total), *vergangenen* (last, 466 instances) , *letzten* (last, 74 instances), *kommenden* (coming, 161 instances), *nächsten* (next, 261 instances) are the four most frequent modifiers. *In ... Jahr*, on the other hand, perferably occurs with demonstratives (929 of 972), like *in* {*diesem, jenem*}*Jahr* (in {this, that} year). Co-occurence with an article is less frequent: there are 37 instances of *in dem Jahr*. Another information provided by the data is that datives are far more frequent than accusatives, 963 of 966 *in*-determiner-*Jahr* instances are datives. The remaining three instances are accusatives. In other words *in ... Jahr* is most likely to have locative reading in the current corpus.[4]

By means of $\langle wt_i, wt_{i+3}\rangle$-examples, it can be shown that a numeric span of four exceeds phrase boundaries. The bi-gram *im Jahres*, for instance, originates from PP NP_{gen} sequences in the extraction corpus like *im September dieses Jahres* (in the September of this year), *im Verlauf eines Jahres* (in the course of a year), *im Deutschland des Jahres ...* (in the Germany of the year ...). It is already clear from the morphological form that *im* and *Jahres* cannot constitute a PP as *im* assigns dative but *Jahres* is a genitive form. In the majority of cases, however, structural inappropriateness of the bi-gram cannot be detected from word form, see for instance the bi-gram *in Kirche* which may originate from a PP like *in der schönen Kirche* (in the beautiful church), but is amongst others derived from *in Räumen der Kirche* (in rooms of the church), where *in* and *Räumen*, and *Räumen* and *Kirche* are syntactic dependents, but not *in* and *Kirche*. While the previous examples represent complex PPs with a postnominal genitive modifier, a numeric span of four also covers unrelated phrases, or cuts

---

[4] The German preposition *in* either assigns dative or accusative, where the former expresses locativity, and the latter directivity.

through phrases. The bi-gram *in Jahren* (in years) is for example supported by the sequence of PPs *in Wiesbaden seit Jahren* (in Wiesbaden since/for years), the bi-gram *auf Seite* is supported by *auf holpriger Spur Seite* (on bumpy road page). Here the span covers a PP (*auf holpriger Spur*) and cuts through an NP beginning with *Seite*. Similarly *für Autos da Kinder* (for cars because children) – which is a PP followed by the initial sequence of a clause (*da Kinder*) – is an example of a word sequence from which the bi-gram *für Kinder* is extracted.

Tables 4.6 to 4.8 show the 20 most frequent preposition-noun-verb co-occurrences based on a span size of maximally five words, whereby the following classes of tri-grams have been extracted: $\langle wt_i, wt_{i+1}, wt_{i+2}\rangle$, $\langle wt_i, wt_{i+2}, wt_{i+3}\rangle$ and $\langle wt_i, wt_{i+3}, wt_{i+4}\rangle$.

$< wt_i, wt_{i+1}, wt_{i+2} >$	freq	$< wt_i, wt_{i+1}, wt_{i+2} >$	freq
zur Verfügung gestellt	143	**zur Verfügung steht**	43
ums Leben gekommen	112	auf Asyl bleibt	42
in Anspruch genommen	95	**zur Kasse gebeten**	38
zur Verfügung stehen	85	ins Krankenhaus gebracht	35
zur Verfügung stellen	58	**in Auftrag gegeben**	34
ins Leben gerufen	57	**zum Opfer gefallen**	33
in Frage gestellt	53	**in Kraft treten**	33
in Betrieb genommen	47	**in Verbindung setzen**	29
zur Kenntnis genommen	44	**in Aussicht gestellt**	29
in Anspruch nehmen	44	zur Zeitungszustellung wenden	28

Table 4.6: ⟨Preposition,noun,verb⟩-tri-gram frequencies identified from the part-of-speech tagged extraction corpus

It becomes evident from the morphological properties of the verbs, the majority of which are either participles or infinitves, that the frequent trigrams mainly originate from sentences with complex predicates. It can be seen from the examples in the tables that frequent preposition-noun-participle or -infinitive sequences are good indicators for PP-verb collocations, especially for collocations that function as predicates such as support-verb constructions and a number of figurative expressions.[5] In [Hoberg, 1981] it is already argued that there is a strong tendency for PP and verb to be adjacent in the surface string in the case of SVCs. Evidence of this kind has been utilized in [Docherty *et al.*, 1997] where very large amounts of data are examined for tri-grams of adjacent preposition, noun and participle. The assumption is also supported by the data in table 4.6 where 17 of 20 PNV-combinations are collocations, and the majority of which

---

[5] Figurative expressions and support-verb constructions are printed in bold face.

are support-verb constructions. The proportion of collocations decreases with increasing span size in the PP, and the type of collocations occurring changes as well, see tables 4.7 and 4.8 where the number of collocations is 12 and 1, respectively, and all collocation instances are figurative. Thus the data suggest that PP-collocates in SVCs are typically composed of a preposition and a noun.

$\langle wt_i, wt_{i+2}, wt_{i+3}\rangle$	freq	$\langle wt_i, wt_{i+2}, wt_{i+3}\rangle$	freq
bis Uhr geöffnet	141	**auf Weg gebracht**	20
nach Smogverordnung überschritten	62	**über Bühne gehen**	19
Auf Programm stehen	51	**auf Fuß gesetzt**	19
Auf Tagesordnung stehen	46	in Krankenhaus gebracht	18
mit Augen gesehen	37	**auf Straße gegangen**	18
in Tasche greifen	28	**auf Beine stellen**	17
mit Hause nehmen	26	Um Uhr beginnt	17
von Jahren gestorben	24	**auf Beine gestellt**	16
unter Lupe genommen	21	**in Lage versetzt**	14
um Uhr beginnt	21	um Prozent gestiegen	13

Table 4.7: ⟨Preposition,noun,verb⟩-tri-gram frequencies identified from the part-of-speech tagged extraction corpus

$\langle wt_i, wt_{i+3}, wt_{i+4}\rangle$	freq	$\langle wt_i, wt_{i+3}, wt_{i+4}\rangle$	freq
auf Mark geschätzt	24	über Mißerfolg gestritten	8
zu Haft verurteilt	22	in Saison aufhorchen	8
zu Gefängnis verurteilt	13	Mit Hessentiteln blieben	8
von Mark entstanden	11	Für Jahr wünscht	8
mit Mark veranschlagt	11	zu Haftstrafe verurteilt	7
durch Frauen helfen	11	von Mark verursacht	7
mit Mark angegeben	10	für Anspruch nehmen	7
in Müll wandern	10	auf Mark veranschlagt	7
am Menschen getötet	10	auf Mark belaufen	7
auf Mark beziffert	9	von Mark zahlen	6

Table 4.8: ⟨Preposition,noun,verb⟩-tri-gram frequencies identified from the part-of-speech tagged extraction corpus

Tables 4.7 and 4.8 also contain other kinds of lexically determined co-occurrences such as

pseudo-collocations, e.g.

> *bis (...) Uhr geöffnet, um (...) Uhr beginnt* (until (...) o'clock open, at (...) o'clock starts),
> *zu (...) {Haft, Haftstrafe, Gefängnis} verurteilt* ('sentenced to prison'),
> *mit (...) Mark {veranschlagt, angegeben}, auf (...) Mark {beziffert, geschätzt, veranschlagt, belaufen}*

and verb-preposition combinations, e.g.

> *verurteilen zu* (sentence to),
> *schätzen auf* (estimate at),
> *veranschlagen mit* (assess at),
> *angeben mit* (specify),
> *beziffern auf* (amount to),
> *belaufen auf* (amount to),
> *streiten über* (quarrel about),
> *wünschen für* (wish for).

Results of Strategy 3

In contrast to the previous examples which are at least partially based on numeric spans, the bi-grams in table 4.9 are taken from prepositional phrases identified by the chunk tagger.

	wt_ic_k	wt_jc_k	freq		wt_ic_k	wt_jc_k	freq
$\star$	um	uhr	2768	# $\star$	im	jahr	1496
$\star$	bis	uhr	2748	$\star$ +	seit	jahren	1307
#	am	sonntag	2179	$\star$	in	jahr	1073
#	am	montag	2015	+	in	jahren	1060
#	am	dienstag	2004	$\star$	ab	uhr	1041
#	am	samstag	1983	# +	für	kinder	993
#	am	freitag	1979	$\star$	vor	jahren	979
#	am	mittwoch	1903	#	zur	verfügung	921
#	am	donnerstag	1810	#	zum	beispiel	833
#	nach	angaben	1577	+	auf	seite	799

Table 4.9: ⟨Preposition,noun⟩-bi-gram frequencies identified from the part-of-speech tagged and chunked extraction corpus

Among the 20 most frequent preposition-noun combinations, there are 12 examples that occur as well among the 20 most frequent $\langle wt_i, wt_{i+1}\rangle$-bigrams (see #), 7 examples that occur also in the $\langle wt_i, wt_{i+2}\rangle$-list (see $\star$), and only three of the $\langle wt_i, wt_{i+3}\rangle$-examples (see +) which gives further evidence that $\langle wt_i, wt_{i+3}\rangle$-pairs tend to exceed phrase boundaries in German.

Co-occurrence frequencies based on PP-chunks rank higher than individual $\langle wt_i, wt_{i+1}\rangle$-, $\langle wt_i, wt_{i+2}\rangle$- or $\langle wt_i, wt_{i+3}\rangle$-frequencies, because of two reasons: (1) the material allowed between preposition and noun is not restricted by a particular span size, and (2) the words have been normalized to lower case, thus no orthographic distinction is made between a PP at the beginning of or within a sentence.

Table 4.10 shows the 20 most frequent preposition-noun-verb co-occurrences identified using parts-of-speech and structural information.

$\langle wt_i c_k, wt_j c_k, wt_l c_m\rangle$	freq	$\langle wt_i c_k, wt_j c_k, wt_l c_m\rangle$	freq
um uhr beginnt	379	am dienstag sagte	95
bis uhr geöffnet	182	**auf tagesordnung stehen**	92
zur verfügung stehen	174	am donnerstag sagte	78
zur verfügung gestellt	143	auf seite lesen	75
zur verfügung stellen	128	**im mittelpunkt steht**	74
zur verfügung steht	115	auf kürzungen behält_vor	74
ums leben gekommen	111	**auf programm steht**	74
auf programm stehen	98	am mittwoch sagte	71
in anspruch genommen	95	**zur verfügung zu stellen**	70
am montag sagte	95	auf seite zeigen	70

Table 4.10: ⟨Preposition,noun,verb⟩-tri-gram frequencies according to syntactic spans

Among the examples we find 11 collocations which is less than in the case of $\langle wt_i, wt_{i+1}, wt_{i+2}\rangle$-sequences. Comparing the data reveals the following differences: While employing the notion of syntactic span allows identifying more instances per collocation type, triples of preposition, noun and main verb, on the other hand, coincide with characteristic linguistic properties of PP-verb collocations namely the tendency of the PP-collocates to comprise only a preposition and a noun, and the tendency of PP-verb collocations to be adjacent in the surface string, as this is the case in complex time constructions (examples 4.4), modal constructions (example 4.5), relative clauses 4.3, infinitive clauses (example 4.2), and deverbal adjective phrases (example 4.1). All examples have been found in the extraction corpus.

(4.1) ... **zur Verfügung gestellten** Gelder seien für den Ankauf wichtiger Materialien eingesetzt worden.
('the finances made available were used for the purchase of important ma-

terials')
(adjective phrase)

(4.2) ..., die Räume auch über den regulären Kündigungstermin hinaus **zur Verfügung zu stellen ,** ...
('to make the rooms also longer available than the regular day of notice to quit')
(infinitive clause)

(4.3) So sucht die RSG einen Sponsor, der das Material kostenlos **zur Verfügung stellt.**
('Thus the RSG is looking for a sponsor who makes available the material for free.')
(relative clause)

(4.4) Für jede Mannschaft werden mindestens ein Trainer und ein Betreuer **zur Verfügung stehen.**
('For each team there will be available at least on trainer and one coach.')
(complex time)

(4.5) Mindestens ein Stellplatz muß je 80 Quadratmeter Nutzfläche **zur Verfügung stehen.**
('There must be available at least one parking lot per 80 squaremeters floor-space.')
(modal construction)

PP-Collocate	V-Collocate	Right Neighbour	Co-occurring Main Verb
zur Verfügung	stehen	189	404
	stellen	240	457
in Kraft	treten	99	126
	setzen	12	23
	bleiben	0	5
in Anspruch	nehmen	139	192

Table 4.11: Occurrence frequencies of verbal partner collocates for *zur Verfügung, in Kraft* and *in Anspruch*

But there is also a large number of co-occurrences where PP and verb are not adjacent in the surface string. The differences in occurrence frequency using syntactic and numerical spans are illustrated in table 4.11 which presents the co-occurrence frequencies between the PPs *in Kraft, zur Verfügung* and *in Anspruch* and related main verbs. The verbs are either selected by means of numerical span (right neighbour of the PP-collocate) or syntactic span (co-occurring main verb). Verb forms have been reduced to their bases. In the examples, co-occurrence frequencies are without exception higher for the data extracted applying syntactic spans. The concept of numeric span performs fairly well as long as preposition, noun and verb are adjacent. The group of collocations covered, however, is restricted. In the case of *in Kraft bleiben*, no such example occurred in the data.

4.3 Characteristics of the Collocation Candidates

After having learned that syntactic constraints are important for the selection of collocation candidates from corpora, linguistic constraints relevant for PNV-combinations are discussed in more detail, and the resulting distributions of collocations and noncollocations in the candidate data are illustrated (section 4.3.1). In addition, it is discussed how a frequency-based selection of collocation candidates influences the distribution of SVCs, figurative expressions and pseudo-collocations within the data (section 4.3.2).

4.3.1 Linguistics-Driven Candidate Selection

A number of PNV-samples is drawn from the extraction corpus. The samples differ with respect to the morphosyntactic and syntactic properties of the data covered, thus allowing for insights into the relation between syntactic properties of the test data and the distribution of the different collocation classes within the samples. As basic requirement, the word combinations being part of a test sample need to occur at least three times in the extraction corpus.

P.N.VVPP-Trigrams

The set comprises all sequences where a preposition P, a noun N and a past participle of a main verb VVPP are adjacent in the extraction corpus. This set is closely related to the one described in [Breidt, 1993], as it covers the PP-verb combinations where the verb complex is in sentence final position, and the dependent noun occurs to the left of the main verb. The differences to the data described in Breidt are: (i) part-of-speech information is available; (ii) the verb is not lexically specified; (iii) only the noun immediately to the left of the verb is examined, in contrast to Breidt where the two words w_{j-2}, w_{j-1} to the left of each key verb w_j are considered; (iv) only combinations with minimal PPs are taken into account. Applying the method, 319 word combinations are selected from the extraction corpus, 102 (32 %) of which are manually identified as SVCs, 39 (12.2 %) as figurative expressions, and 25 (7.8 %) as pseudo-collocations. In the following, the test data are successively enlarged by broadening the syntactic and morphological coverage.

P.N.VVPP-Triples

Other than in the case of the trigram sample, no adjacency or order requirements are stated. The PN-combinations need to be constituents of the same PP, PP and past participle need to co-occur in the same sentence. While the dependency requirements for the preposition and the noun are strict, no dependency relation is required for the PP-verb combination. Even though the latter decision leads to the stipulation of syntactically inappropriate PP-verb combinations, it is justified for practical reasons as automatic PP-attachment is highly inaccurate. A large number of the syntactically independent PP-verb combinations are directly eliminated because of infrequency. This is also the case for the other samples. In the current sample, the number of PNV-combinations increases to 2 959 containing a larger number of collocations than the previous set, i.e., 139 (4.7 %) SVCs, 107 (3.6 %) figurative expressions, and 75 (2.5 %) pseudo-collocations. Collocation density, on the other hand, drastically declines, see the percentages values in brackets.

P.N.VVPP|IZU|VVINF-Triples

PN-combinations are constituted as above. Instead of looking for combinations with past participles only, bare infinitives (VVINF) and to_infinitives (IZU) are considered as well. The number of collocation candidates and true collocations further increases, and there is also a slight increase in collocation density compared to the previous set. The set contains 5 042 word combinations of which 335 (6.6 %) are SVCs, 277 (5.5 %) figurative expressions, and 106 (2.1 %) pseudo-collocations.

P.N.VVPP|IZU|VVINF-Trigrams

This set is constructed similar to the set of P.N.VVPP-trigrams except that the verbs can be either bare infinitives, to_infinitives or past participles. Compared to the set of P.N.VVPP-trigrams, the number of collocation instances increases. There are 161 (33.3 %) SVCs and 61 (12.6 %) figurative expressions, and 27 (5.6 %) pseudo-collocations. SVC-density is slightly higher than in the set of P.N.VVPP-trigrams.

P.N.V(Full Form)-Triples

Here the only co-occurrence restriction between PN and V is that V must be a main verb occurring in the same sentence as PN. This relaxation leads to a set of 10 430 word combinations, 710 (6.8 %) of which are SVCs, 586 (5.6) are figurative expressions and 337 (3.2 %) are pseudo-collocations.

P.N.V(Base Form)-Triples

All main verbs are reduced to their base forms. This way, PNV-combinations (types) containing the same verb stem are reduced to a single type. Thus the occurrence frequency of morphologically flexible types increases, resulting in a larger sample set with a grown proportion of high frequency co-occurrences. Unfortunately the density of true collocations further decreases. The sample consists of 14 660 PNV-triples, of which 412 (2.8 %) are SVCs, 527 (3.6 %) are figurative expressions, and 345 (2.4 %) are pseudo-collocations.

Kwic-Based Reduction of the Test Sets

In addition, the test samples are reduced by applying the kwic-method to the the above samples. In the case of P.N.V(full forms), the verbs are first reduced to base forms, and the kwic-method is then applied to the morphologically reduced triples. In the following, the effects on the different samples are described.

P.N.VVPP-trigrams

The number of PNV-combinations selected reduces to 129, of which 93 (72 %) are SVCs, 14 (11 %) are figurative expressions, and 9 (7 %) are pseudo-collocations. This is a recall of 91.2 % of the SVCs occurring in the original set of 319 word combinations.

P.N.VVPP|IZU|VVINF-trigrams

SVC-density is comparable to the previous sample, whereas the density of figurative expressions and pseudo-collocations has declined.

P.N.VVPP-triples

412 word combinations are identified 118 of which are SVCs. This is a recall of 84.9 % with a precision of 28.6 %. For comparison, only 43 figurative expressions and 10 pseudo-collocations are identified.

P.N.VVPP|IZU|VVINF-triples

816 PNV-combinations are extracted which contain 236 SVCs, 106 figurative expressions and 15 pseudo-collocations. This is 70.4 % recall for SVCs and an increase in precision to 29 %.

P.N.VVPP|IZU|VVINF-trigrams

The data reduce to 213 PNV-combinations with 148 (69.5 %) SVCs, 25 (11.7 %) figurative expressions and 10 (4.7 %) pseudo-collocations. Recall of SVCs is 92 %.

P.N.V(full form)-triples

Reduction of the verbs to their bases, and application of the kwic-model lead a stepwise decrease of the number of applicable PNV-combinations, such that 2 348 combinations are left from the original 10 430. The number of SVCs reduces to 272 (11.6 %), which amounts to 75 % of the initial number of SVCs. The number of figurative expressions reduces to 176 (7.5 %), and the number of pseudo-collocations to 71 (3 %).

P.N.V(base form)-triples

The number of PNV-triples here reduces to 2 299 of which are 239 (10.4 %) SVCs, 195 (8.5 %) figurative expressions and 51 (2.2 %) pseudo-collocations. In this case, the kwic-strategy leads to a recall of 58 % for SVCs.

The raw data are presented in tables 4.12 to 4.14, p. 109. For illustration, the distributions of collocations and noncollocations within the full test sets (all data) and the reduced test sets (kwic-based) are graphically represented, see tables 4.2 to 4.4, p. 110 – 112. In the case of P.N.V(full forms), “full all” represents the distribution in the set of full form data, “base all” represents the set after reduction of the verbs to their bases, and “kwic bases”represents the morphologically reduced data after application of the kwic-strategy.

Observations

- The number of collocational and noncollocational data covered grows with relaxation of the syntactic constraints applied during the construction of the test set.[6] Collocation density, however, decreases. The largest amount of data is covered by the set of P.N.V(base forms), the set with the most permissive construction criteria. The percentage of collocations is fairly small. In contrast, the sets of P.N.VVPP- and P.N.VVPP|IZU|VVINF-trigrams, on the one hand, are the smallest sets, as they are most rigid with respect to the construction criteria. On the other hand, the sets show the highest density of collocations. This discrepancy illustrates a central dilemma in corpus-based collocation identification, i.e., the number of collocations contained in a candidate set and collocation density within the set is inversely proportional.

---

[6] Numeric spans can be viewed as the extreme case of syntactic relaxation.

- A peculiarity of the data with reduced verb forms is that figurative expressions outnumber SVCs. Reduction to verb bases has strongest effect on SVCs. Reduction of verbs in the set of P.N.V(full form)-triples leads approximately to a 50%-reduction of SVCs, a 30%-reduction of figurative expressions and a 10%-reduction of pseudo-collocations. This kind of discrepancy is also found comparing the number of collocations among the P.N.V(full form)-triples and the P.N.V(base form)-triples. While the number of SVCs declines from 710 to 412, the number of figurative expressions just slightly decreases from 586 to 527, and the number of pseudo-collocations, on the contrary, increases from 337 to 345 as can be seen from the raw data in the tables above.

- The graphical representations depict that in all samples a kwic-based data reduction leads to a substantial increase of the percentage of SVCs among the test data. There is also a moderate increase of the percentage of figurative expressions, which shows that a certain subset of verbs take part in SVCs and figurative expressions. The kwic-strategy leads to a substantial decrease in proportion, even though pseudo-collocations are not fully eliminated.

- P.N.VVPP- and P.N.VVPP|IZU|VVINF-trigrams
  are particularly well suited for the identification of SVCs. SVC-density is highest, when the kwic-strategy is applied. This is due to the following reasons: (1) The collocates of SVCs tend to occur in close neighbourhood to each other in verb final constructions, which has been already argued for in [Hoberg, 1981]. (2) PP-collocates of SVCs tend to be minimal, i.e., a large number of PP-collocates consist only of a preposition and a noun. (3) A basic set of support-verbs is easy to determine.

P.N.VVPP-trigrams		
	$c \geq 3$	kwic
total	319	129
SVC	102	93
figur	39	14
pseudo	25	9

P.N.VVPP-triples		
	$c \geq 3$	kwic
total	2 959	412
SVC	139	118
figur	107	43
pseudo	75	10

Table 4.12: Raw data: P.N.VVPP-trigrams versus -triples

P.N.VVPP|IZU|VVINF-trigrams		
	$c \geq 3$	kwic
total	484	213
SVC	161	148
figur	61	25
pseudo	27	10

P.N.VVPP|IZU|VVINF-triples		
	$c \geq 3$	kwic
total	5 042	816
SVC	335	236
figur	277	106
pseudo	106	15

Table 4.13: Raw data: P.N.VVPP|IZU|VVINF-trigrams versus -triples

P.N.V(full form)			
	$c \geq 3$ full	$c \geq 3$ base	kwic base
total	10 430	8 828	2 348
SVC	710	362	272
figur	586	400	176
pseudo	337	306	71

P.N.V(base form)		
	$c \geq 3$	kwic
total	14 660	2 299
SVC	412	239
figur	527	195
pseudo	345	51

Table 4.14: Raw data: P.N.V(full forms) versus -(base forms)

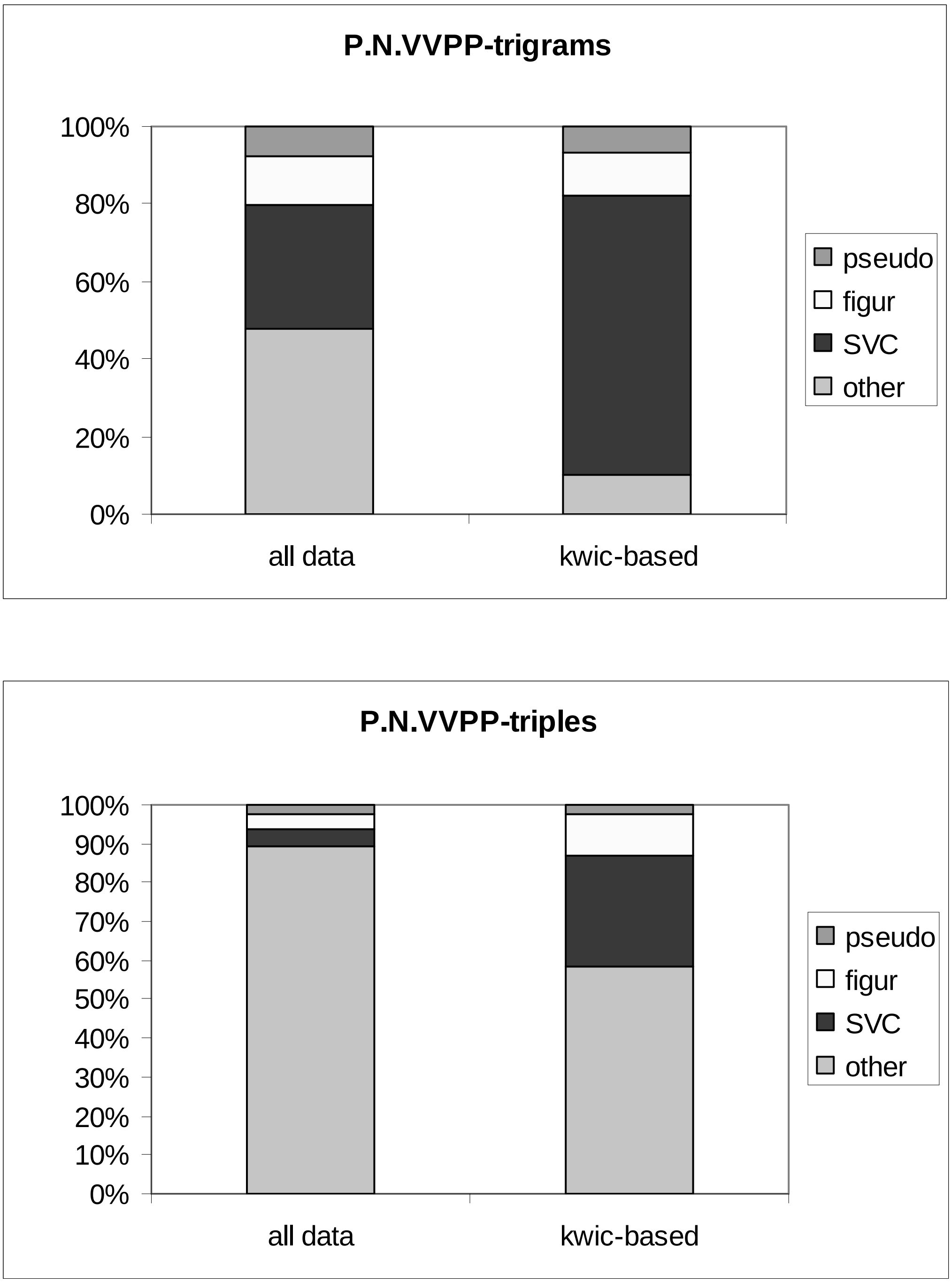
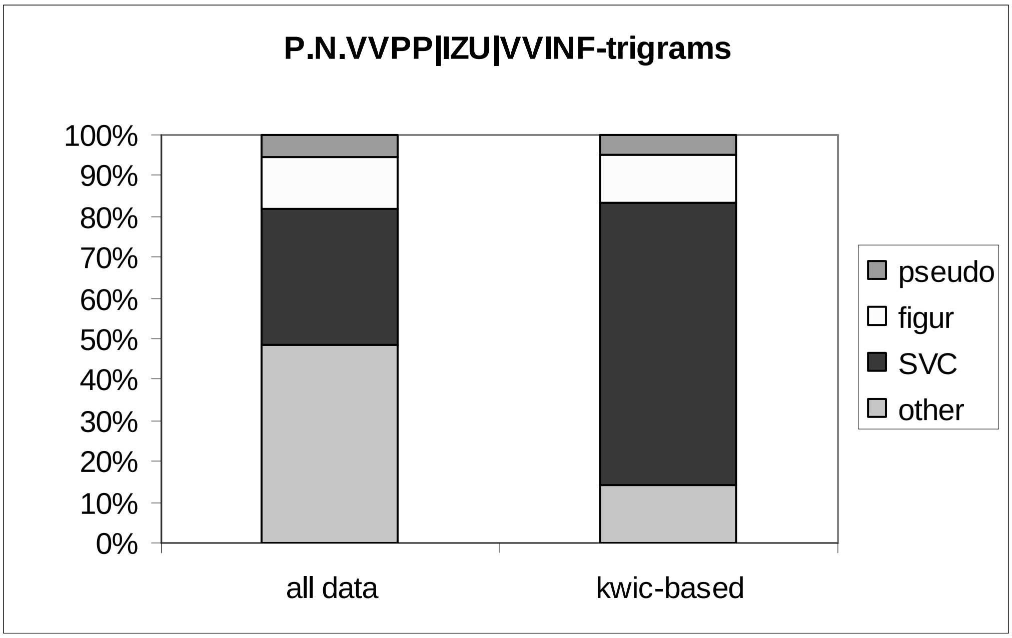


Figure 4.2: Graphical representation: P.N.VVPP-trigrams versus -triples



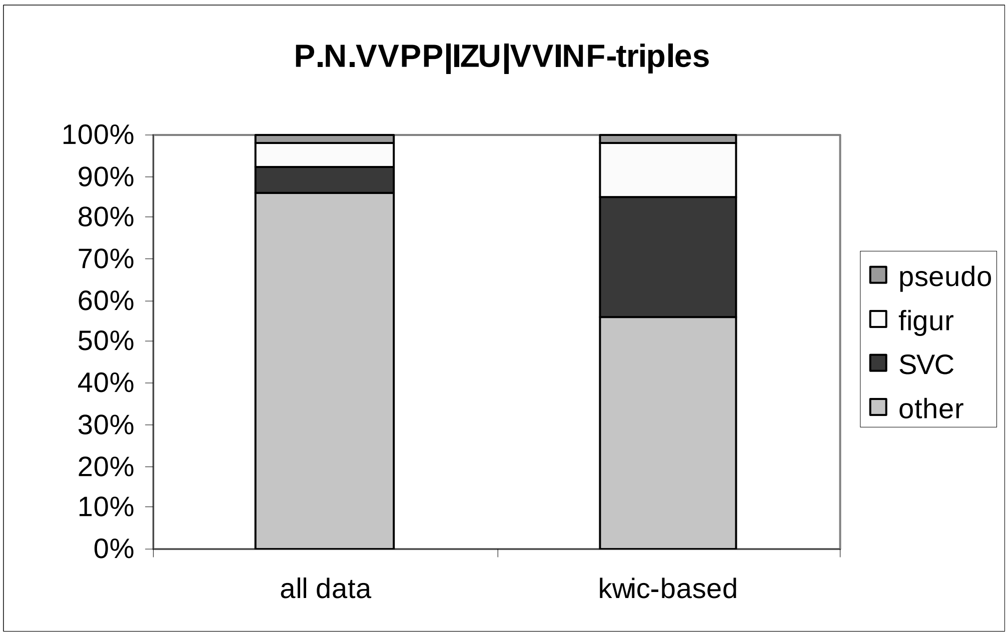
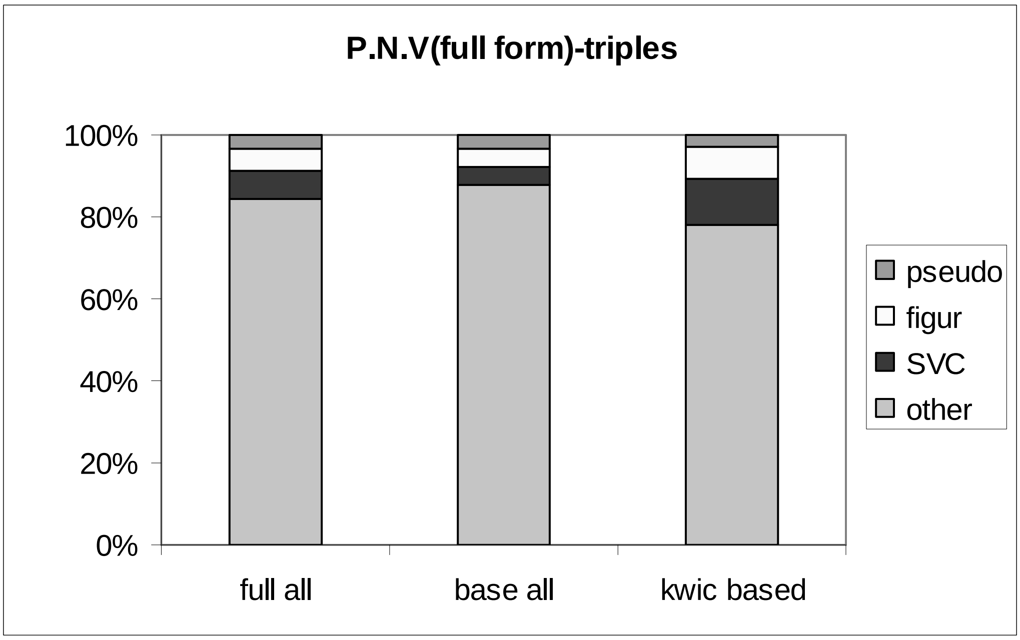


Figure 4.3: Graphical representation: P.N.VVPP|IZU|VVINF-trigrams versus -triples



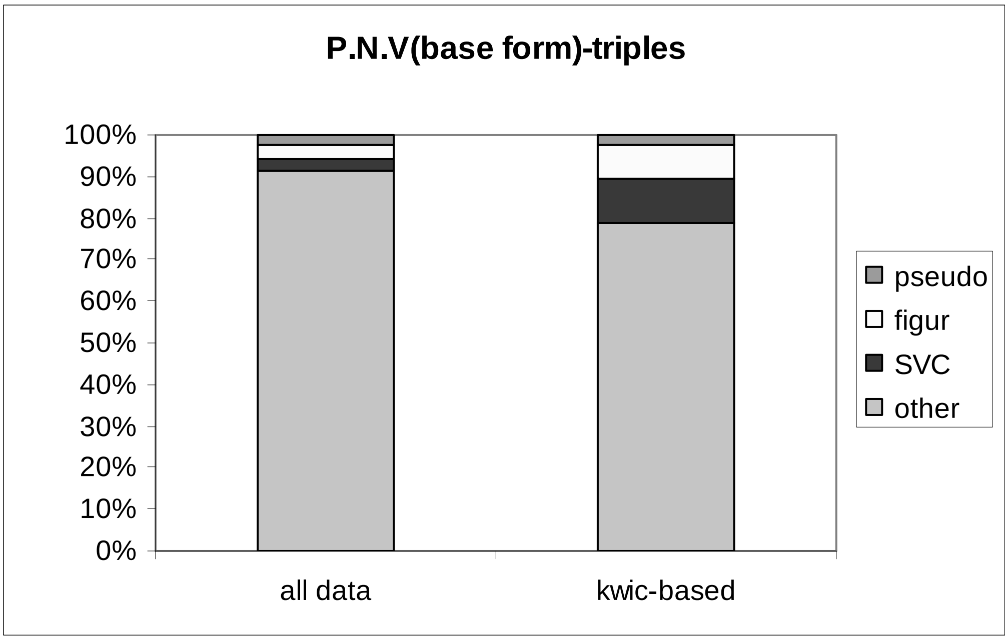


Figure 4.4: Graphical representation: P.N.V(full forms) versus -(base forms)

4.3.2 Frequency-Based Candidate Selection

In the following, the motivations for selecting the particular thresholds are given.

$t = 10$ defines set A of PNV-combinations with co-occurrence frequency $c \geq 10$. Statistics-based collocation identification from this set is expected to be easier than from the following two sets, because overestimation of co-occurrences of infrequent words is avoided. Due to the high proportion of collocations among the data, however, even models that perform slightly better than chance will lead to relatively good results.

$t = 5$ defines set B of word combinations with co-occurrence frequency $c \geq 5$. The set is well suited for testing the models on a broader range of co-occurrence frequencies without an extreme loss of accuracy because of overestimation of low frequency data.

$t = 3$ defines set C with co-occurrence frequency $c \geq 3$. 3 is the minimal occurrence frequency for PNV-combinations to take part in collocation identification in this study. Set C includes the previous sets. Because of the large number of low frequency co-occurrences, the set is a challenging test suite for statistical models.

Differences between Sets A, B and C

After comparison of the test sets against each other, the sets are described with respect to internal differences between the full set C ($c \geq 3$), and the subsets A ($c \geq 10$) and B ($c \geq 5$). The sets of P.N.VVPP-trigrams, P.N.V(full form)-triples and P.N.V(base form)-triples are used for illustration. Raw data and graphical representations are presented on page 115ff.

Examining the data, the following observations can be made:

- The proportion of collocations among the data increases with increasing co-occurrence frequency.
- The proportion of SVCs is largest in the subset of P.N.VVPP-trigrams where $c \geq 10$. The set, in general, has a large proportion of collocations, as it contains only word combinations with high co-occurrence frequency, and its construction criteria meet characteristic syntactic properties of verb-object collocations.
- The proportion of pseudo-collocations and figurative expressions also increases from C to A in the sets constituted by P.N.V(base)- and P.N.V(full form)-triples.

- On the other hand, there is little difference in the proportion of figurative expressions between sets A, B and C when P.N.VVPP-trigrams are considered. The proportion is highest in set B. Suggesting that figurative expressions are more broadly distributed than SVCs.

- The percentage of pseudo-collocations overproportionally increases with growing co-occurrence frequency.

- The highest proportions of figurative expressions were found in the sets of P.N.V(base)- and -(full form)-triples where $c \geq 10$, and in the set of P.N.VVPP-trigrams where $c \geq 5$. The highest recall is achieved from the first set providing further evidence that there is less variation in the verb inflection and as a consequence in the syntactic constructions in the case of figurative expressions than in the case of SVCs.

	P.N.VVPP-trigrams			P.N.V(full form)			P.N.V(base form)		
sets	C	B	A	C	B	A	C	B	A
SVC	102	60	33	710	369	144	412	304	174
figur	39	17	6	586	282	96	527	338	150
pseudo	25	12	8	337	302	237	345	315	262
other	153	36	3	8 798	1 911	270	13 376	3 532	663

Table 4.15: Raw data according to frequency-based candidate selection

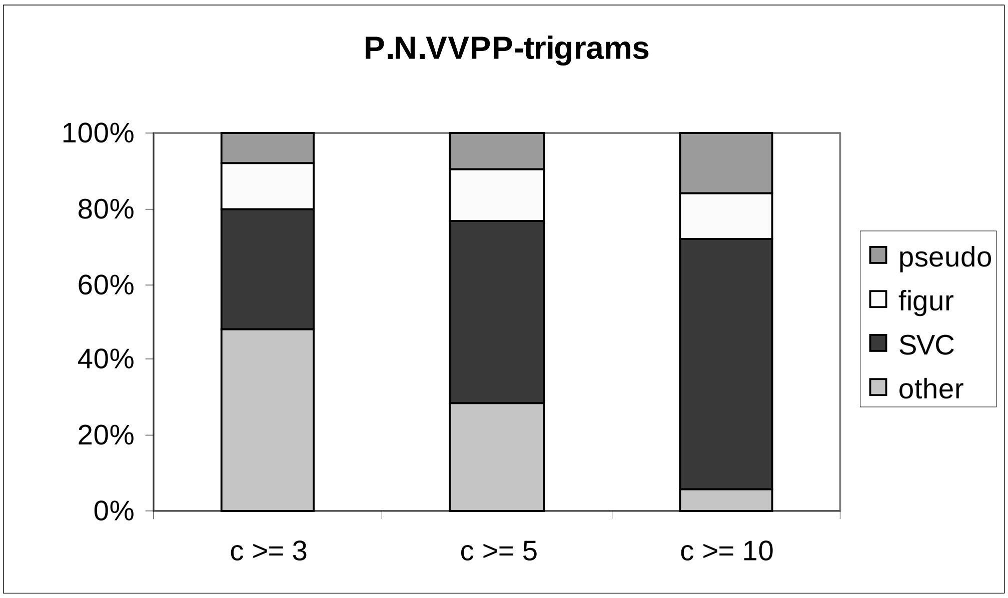


Figure 4.5: Graphical representation: P.N.VVPP-trigrams

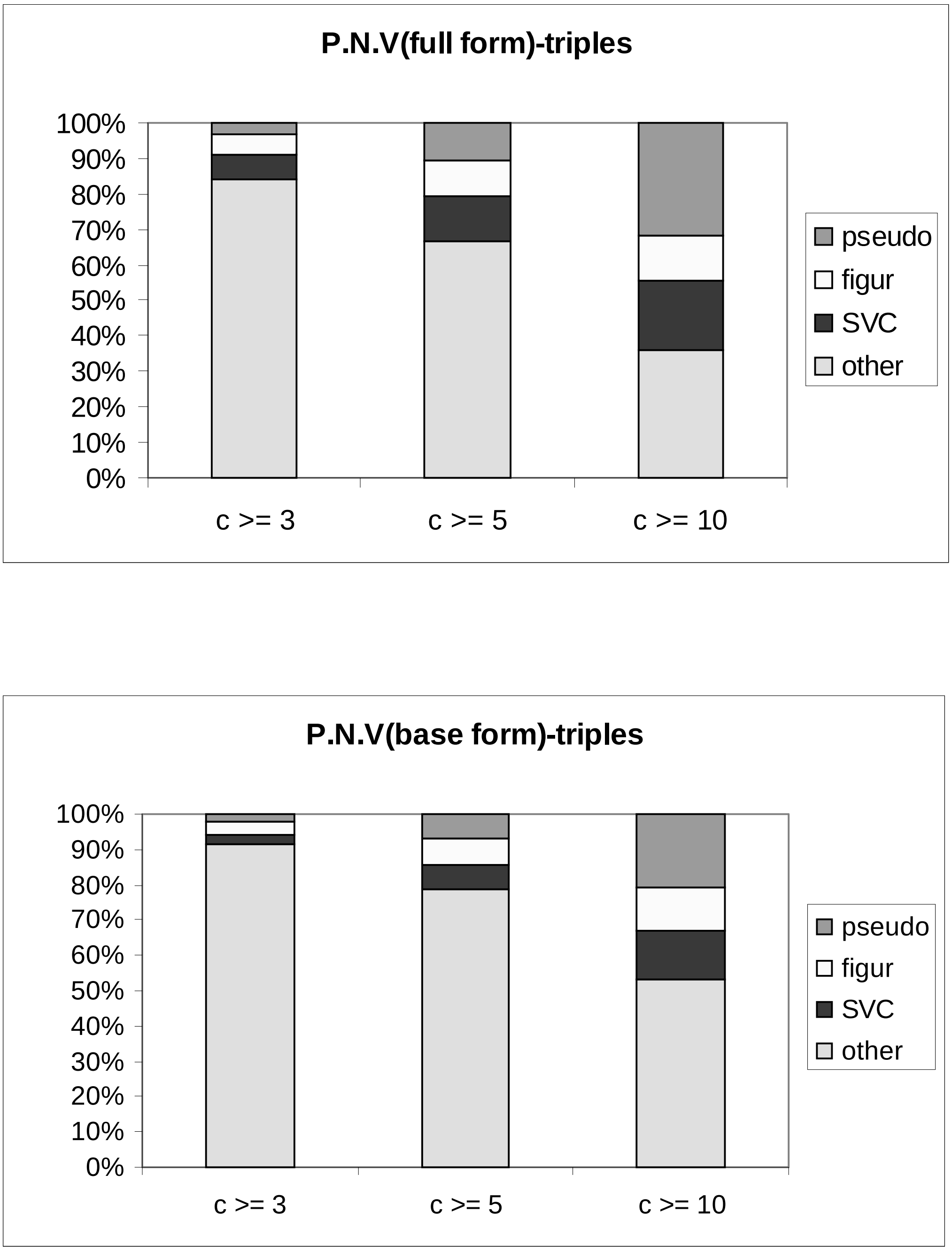


Figure 4.6: Graphical representation: P.N.VVPP(full and base form)-triples

4.4 Models for Collocation Identification

In the following, three models for collocation identification are presented, each one accounting for a specific characteristic of collocations, namely over proportional recurrence of particular word combinations in text corpora, grammatical restrictions that typically coincide with particular word combinations, and lexical selection between the collocates of a collocation.

4.4.1 Lexical Cohesion Measures

State-of-the-art approaches to statistics-based collocation identification all make use of the recurrence of the collocates of a collocation in text corpora. As already described in section 2.1.1, different measures have been proposed mainly as improvements to specific mutual information MI as it was defined in [Church and Hanks, 1989]. Accordingly, MI and its most successful alternative, the log-likelihood statistics Lgl presented in [Dunning, 1993] must not be missing in the present study. In addition, two other measures are employed, the $Dice$ coefficient and relative entropy I. $Dice$, because it is a simple association measure accounting for positive[7] word combinations only. This is also the case for MI. I has been selected, as it measures the informativity of one random variable with respect to another one. All four measures employ frequency information over PNV-triples $f(PNV)$ which are the estimates for the joint probabilities $p(PN, V)$, and frequency information over PN-tuples $f(PN)$ and V-unigrams $f(V)$ which are the estimates for the marginal probabilities $p(PN)$ and $p(V)$. When applied to a set of collocation candidates, each of the measures imposes its own order on the test set, ranking the PNV-combinations in terms of likelihood. Information on the particularities of the individual measures, and guidelines interpreting the rankings are presented in section 2.3.1.

In addition to the statistical association measures, mere co-occurrence frequency $freq$ is also taken into account. In this case, the PNV-combinations are ranked according to their occurrence frequency. Word combinations that occur more often in a text corpus are expected to be more likely collocations than word combinations that occur only rarely in the corpus under investigation. Occurrence frequency is the most simple means to model recurrence. Thus it is used as a baseline against which the other association measures are compared.

---

[7] In such an approach, the lexical association between, for example, *zur Verfügung* and *stellen*, is determined by the occurrence frequencies of *zur Verfügung stellen*, *zur Verfügung* and *stellen* only. Combinations where *zur Verfügung* or *stellen* occur with other partners are not considered.

Estimates

For calculation of the statistics-based association measures, relative frequencies (Maximum Likelihood Estimates MLE) are used. The following values are specified: absolute frequencies of collocations $f(c1c2)$ and collocates $f(c1)$, $f(c2)$, and the values in the contingency table 2.1 where $a = f(c1c2), b = f(c1\neg c2), c = f(\neg c1c2), d = f(\neg c1\neg c2)$. The frequencies are normalized by n, the number of words in the corpus, in order to keep the figures used for calculation small. The ranking of the PNV-combinations according to a particular measure, however, is not influenced by the normalization factor. In the following an overview of the formulas used is given.

Mutual information as described in [Church and Hanks, 1989]

$$MI_{Church;Hanks} = \log \frac{\frac{f(c1c2)}{n}}{\frac{f(c1)}{n} * \frac{f(c2)}{n}}$$

which can be reformulated as

$$MI = \log \frac{\frac{a}{n}}{\frac{a+b}{n} * \frac{a+c}{n}}$$

Dice coefficient as described in [Smadja *et al.*, 1996]

$$Dice = \frac{2 * \frac{f(c1c2)}{n}}{\frac{f(c1)+f(c2)}{n}}$$

which can be reformulated as

$$Dice = \log \frac{2 * \frac{a}{n}}{\frac{2a+b+c}{n}}$$

Log-Likelihood according to Dunning. The formula below is valid for cases where $\{a, b, c, d\} > 0$; $N = a + b + c + d$. The following conventions hold: $0 \log \frac{0}{y} = 0$, otherwise: $\log x$ is undefined for values $x \leq 0$.

$$\begin{aligned} Lgl \quad = \quad 2 * (\quad & \frac{a}{n} * \log \frac{\frac{a}{n} * N}{\frac{a+b}{n} * \frac{a+c}{n}} \quad + \\ & \frac{b}{n} * \log \frac{\frac{b}{n} * N}{\frac{a+b}{n} * \frac{b+d}{n}} \quad + \\ & \frac{c}{n} * \log \frac{\frac{c}{n} * N}{\frac{c+d}{n} * \frac{a+c}{n}} \quad + \\ & \frac{d}{n} * \log \frac{\frac{d}{n} * N}{\frac{c+d}{n} * \frac{b+d}{n}} \quad) \end{aligned}$$

Relative entropy, i.e., mutual information as defined in information theory. The formula below represents the case where $\{a, b, c, d\} > 0$, additional formulas are required for cases where $0 \log \frac{0}{y} = 0$, $\log x$ is undefined for values $x \leq 0$.

$$\begin{aligned} I \; = \; & \tfrac{a}{n} * \log \frac{\frac{a}{n}}{\frac{a+b}{n} * \frac{a+c}{n}} & + \\ & \tfrac{b}{n} * \log \frac{\frac{b}{n}}{\frac{a+b}{n} * \frac{b+d}{n}} & + \\ & \tfrac{c}{n} * \log \frac{\frac{c}{n}}{\frac{c+d}{n} * \frac{a+c}{n}} & + \\ & \tfrac{d}{n} * \log \frac{\frac{d}{n}}{\frac{c+d}{n} * \frac{b+d}{n}} & \end{aligned}$$

4.4.2 An Entropy Model for PP-Instances

Here, restrictions in linguistic variability of the PP-collocates are applied for collocation identification. Invariance of the PP may be due to collocation-specific restrictions in determination, and blocked or restricted modification. The according linguistic properties are reflected in the surface realizations of the PPs, in particular in frequency counts over the surface strings comprising the preposition, the noun, and the intervening lexical material.

Entropy H is a suitable means for modeling the (in)variation of PP-instances related to a particular PN-combination. For a mathematical discussion of entropy see [Cover and Thomas, 1991] or any other standard book on information theory.

$$H = -\sum_{i=1}^{n} p(X = x_i) \log p(X = x_i)$$

Entropy measures the informativity of a probability distribution $p(X)$: the larger the entropy the more information is contained in the distribution which also indicates that there is little certainty with respect to the outcome. Distributions with distinct peaks are less informative than flat distributions. Applied to collocation phrases, the following holds: Given a PN-tuple and its related PP-instances[8], the instances with identical surface strings are grouped together. Each PN-combination is associated with k classes of m instances. In accordance with linguistic observations, the occurrence of classes with overproportionally large m is an indicator for collocativity, as collocational PN-combinations have low entropy values. The approach, however, requires the definition of a threshold which needs to be empirically specified.

Estimates

Maximum likelihood estimates are also used for calculation of PP-entropy. The probability distribution constituted by the minimal instances of a PN-tuple j, $p(PPinstance_{i_{PN_j}})$, with $i = 1 \ldots k$ is estimated by

---

[8] i.e. all PPs in the extraction corpus which are constituted by the particular P and N irrespective of their occurrence within collocations or noncollocations;

$$\frac{f(PPinstance_{i_{PN_j}})}{f(PN_j)}$$

with $f(PPinstance_i) = m$, $m = 1, 2, 3, \ldots$, the number of occurrences of $PPinstance_i$ in the extraction corpus, and $f(PN_j)$ the number of PN-tuples j in the extraction corpus.

Thus we calculate

$$PPentropy_{PN_j} = -\sum_{i=1}^{k} \frac{f(PPinstance_{i_{PN_j}})}{f(PN_j)} \log \frac{f(PPinstance_{i_{PN_j}})}{f(PN_j)}$$

4.4.3 Kwic-Based Identification of SVCs

In this case, lexical selection is modeled, in particular, lexical keys are used for selecting potential collocations. Thus the model is called kwic-based; kwic means "key word in context". Such an approach is expected to be particularly suitable for SVCs, as a set of typical support verbs is easy to specify, see for instance [Breidt, 1993] where the following verbs are listed: *bleiben* (stay), *bringen* (bring), *erfahren* (experience), *finden* (find), *geben* (give), *gehen* (go), *gelangen* (get), *geraten* (get), *halten* (keep), *kommen* (come), *nehmen* (take), *setzen* (set), *stehen* (stand), *stellen* (put), *treten* (step), *ziehen* (draw). A method of automatic identification of potential support-verbs is specified in [Grefenstette and Teufel, 1995].

In the ongoing study, the same list of verbs as suggested in [Breidt, 1993] is employed, because the particular verbs have proven to be representative for SVCs, i.e., 91 to 92 % of the SVCs in the samples with the highest SVC-density, such as the sets of P.N.VVPP- and P.N.VVPP|IZU|VVINF-trigrams,[9] are covered by these verbs.

4.5 Conclusion

Numeric versus Syntactic Spans: The results from the previous experiments confirm that accessibility of syntactic information is important for retrieval of appropriate collocation candidates from corpora. Retrieval of $n-$grams over word forms only results in a huge number of word combinations comprising function words only. Thus usually stop word lists (lists of function words) are employed to discard according word combinations. In general, part-of-speech tagged text is a better basis for collocation identification, as the selection of

---

[9] See 4.3.1 for details on the test samples.

collocation candidates is determined by the co-occurrence of words with certain part-of-speech labels. Thus collocation-relevant function words like determiners and prepositions can be accounted for. Span size is another important factor that influences the appropriateness of the collocation candidates. In the experiments, it has been shown that preposition-noun pairs over a span size of two or three words are more likely to cover PPs than preposition-noun pairs over larger spans, as in the latter case the $n-$grams tend to exceed phrase boundaries. A span size of two, on the other hand, coincides with the linguistic observation that a large number of PP-collocates consists exactly of a preposition and a noun. Similarly, spans of size three or four (with the verb as rightmost element) are well suited for identifying preposition-noun-verb collocations from verb final constructions. The appropriateness of such constructions for identifying SVCs has already been stated in [Breidt, 1993]. Breidt achieves good results assuming the nominal collocate to occur one or two words to the left of a key verb. The results from the previous experiment, as well as the result from Breidt show that numeric spans are appropriate for identification of collocation candidates as long as the spans are defined such that collocation-specific linguistic units are covered. The notion of numeric span, however, needs to be replaced by syntactic span, in order to access the full variety of PP-verb combinations without unnecessarily increasing the number of inappropriate PNV-combinations. This is particular important for languages with flexible word order such as German.

Characteristics of the Collocation Candidates:

P.N.VVPP- and P.N.VVPP|IZU|VVINF-trigrams contain the highest proportions of SVCs.

Reduction of the sets of collocation candidates by a kwic-based approach using a set of typical support-verbs as keys allows increasing the percentage of SVCs and figurative expressions among the candidate data. As expected, the effect is much stronger for SVCs than for figurative expressions, and marginal for pseudo-collocations.

Relaxation of morphosyntactic and syntactic constraints on sample construction allows increasing the number of collocations covered, but also leads to an over proportional growth of the number of noncollocational word combinations among the candidate data. The discrepancy between the number of collocations covered and collocations density is a central problem in corpus-based collocation identification.

Another peculiarity is that SVCs and figurative expressions are reversely distributed within P.N.V(base form)- and P.N.V(full form)-triples. Whereas the number of SVCs is higher in full form data, the number of figurative expressions is higher in base form data. This means, that with respect to language usage there is more variation of verb inflection in SVCs than in figurative expressions.

The proportion of SVCs and pseudo-collocations is largest in samples of full form data where co-occurrence frequency c is high, i.e., in sets where $c \geq 10$. In terms of proportion, figurative expressions are more evenly distributed over sets where $c \geq 5$ and $c \geq 10$. However, a general tendency holds for all collocations, i.e., collocation density among the candidate data increases with increasing occurrence frequency.

Summing up, there are two major strategies for increasing the proportion of true collocations among the candidate data: (i) constrain the construction of the candidate data by collocation-specific syntactic properties; (ii) consider only word combinations with high occurrence frequency.

Models for Collocation Identification: Three kinds of models for collocation identification have been presented, each of which accounting for one of the three defining characteristics of collocations employed in the thesis.

1. Recurrence of collocations in text corpora is modeled by mere co-occurrence frequency and four well known statistical association measures, i.e., mutual information MI, $Dice$ coefficient, relative entropy I and a log-likelihood statistics Lgl. Statistical word association measures are employed for calculating the ratio between joint and marginal probabilities of word combinations. Simple association ratios (MI and $Dice$) and measures that account for the significance of individual word combinations with respect to a particular sample (I and Lgl) are distinguished.

2. Based on the linguistic observation that restrictions are an indicator for collocativity, a stochastic model for grammatical restrictions in PP-collocates is introduced. The restrictions are modeled by calculating the entropy of the minimal phrases constituted by a PN-combination being part of a PNV-tuple found in the extraction corpus.

3. Lexical selection between the collocates is modeled by employing typical support-verbs for a distinction of collocational and noncollocational PNV-combinations. The approach is comparable to the one described in [Breidt, 1993].

Maximum likelihood estimates for the statistical models are presented. While models 1. and 2. are expected to be equally well suited for different types of PNV-collocations, model 3. is particularly designed for identifying SVCs.

Chapter 5

Experiments

5.1 Introduction

The aim of the current chapter is testing the different models described in the previous chapter with respect to their feasibility for collocation identification.

First, the association measures presented in section 4.4.1 are tested, see section 5.3. The questions of interest are

- Do the mathematical differences between the statistical association models have significant effects when applied to German PNV-combinations?
- Is there one best measure for identifying collocations from German PNV-data?
- Is there a difference between the more sophisticated statistical association measures and a simple frequency-based approach?

Secondly, the results achieved by the association models are compared with the kwic-based strategy (section 4.4.3), where collocations are selected by purely lexical constraints, see section 5.4. Here, the questions of interest are

- Is a purely lexicon-based approach as the kwic-model in its results comparable to approaches based on lexical co-occurrence frequency, such as the association models?
- Can the kwic-strategy further improve the results achieved by the association measures?

Third, the PP-entropy model (section 4.4.2) is compared to the association measures and the kwic-strategy. It is then tested whether combination with the kwic-strategy leads to an improvement of the results, see section 5.5. The questions of interest are:

- Is PP-entropy an alternative to the association measures?
- Which results can be achieved by combining the entropy- and the kwic-model?

In section 5.7, a number of key experiments is repeated, employing a German newsgroup corpus which strongly differs from the newspaper corpus, the basis for the initial experiments. While the newspaper corpus is controlled with respect to style and spelling, the newsgroup corpus is an example of much more spontaneous language production. Thus the newsgroup corpus is assumed to be appropriate for testing the generality of the results gained by employing the newspaper corpus.

5.2 Hypotheses to be Tested

The arguments presented hitherto shall now be condensed into a number of hypotheses which will be examined in the experimental section below.

Hyp: Mere occurrence frequency is well suited for corpus-based collocation identification in general, and in particular it is comparable to the statistical association measures such as MI, $Dice$, I and Lgl.

Hyp: The accuracy (precision) of collocation identification can be improved by employing collocation-class-specific linguistic information – such as lexical keys (kwic) or the rigidity of collocation phrases (PP-entropy) – for collocation identification in addition to mere co-occurrence frequency.

Hyp: The statistical association measures can be divided in two groups: MI and $Dice$ versus I and Lgl, because of the differences and similarities between the models.

Hyp: Because of the different distributions of collocation classes within a test sample, the statistical association models will differ in their feasibility to identify a particular collocation class.

Hyp: The kwic-based approach is particularly well suited for identifying SVCs.

Hyp: PP-entropy is equally well suited for identifying SVCs and figurative expressions.

5.3 Evaluation of the Association Measures

This section aims at testing the feasibility of the statistical association measures I, Lgl, MI and $Dice$ for collocation identification with respect to sets A, B and C from P.N.V(full form)- and -(base form)-triples. The models are compared against each other and against mere co-occurrence frequency $freq$. The aim is testing whether the association models in practice fall into the two groups – I, Lgl on the one hand, and MI, $Dice$ on the other hand – as it is expected considering the mathematical similarities and differences of the models. Comparison with mere occurrence frequency is of interest, in order to determine how far the naive approach of recurrence leads in collocation identification. In other words, occurrence frequency is used as baseline in judging the statistical association models.

In the experiments, three groupings of collocations are distinguished:
(1) Collocations$_{all}$: No distinction is made between SVCs, figurative expressions and pseudo-collocations.
(2) Collocations$_{SVC,figur}$: Only SVCs and figurative expressions are considered. This way, a strong bias towards high frequency co-occurrences in the collocation data is avoided, because SVCs and figurative expressions are more evenly distributed over high and low ranks of co-occurrence frequencies than it is the case for pseudo-collocations.
(3) SVCs and figurative expressions are examined separately, in order to test the feasibility of the different models for identifying a specific class of collocations.
Groups (1) and (2) are examined in experiments I. The distinction in (3) is examined in experiments II.

In order to ensure equal conditions for comparing the association models, the n highest ranked word combinations per measure are compared. This method is employed, because each association model imposes a particular order on the PNV-combinations when applied to a test sample. The particular orders are interpreted as collocability rankings, with the n highest ranked word combinations per measure being considered collocational, in terms of probability. Specification of a threshold is another possibility to distinguish collocations from noncollocations. An appropriate threshold, however, is hard to define, because it needs to be determined on a case-by-case basis, and moreover it is not clear how it could be ensured that the thresholds employed for the individual measures are comparable, and because of the differences between the models it is also not possible to employ a single threshold to all models. The following sample classes are examined in the experiments: A, $n = 500$; B, $n = 500, 1\ 000, 1\ 500, 2\ 000$; C, $n = 500, 1\ 000, 1\ 500, 2\ 000$. For a description and motivation of the test statistics applied see section 2.3.2.

5.3.1 Experiment I

Tables 5.1 and 5.2 show the distribution of $\text{collocations}_{all}$ and $\text{collocations}_{SVC,figur}$ as identified from the set of P.N.V(full form)-triples by the four association measures, and by occurrence frequency. The data are used for testing whether there are differences between the association models at all including mere co-occurrence frequency, and if yes, whether a single best model can be identified.

$\text{collocations}_{all}$						
set	sample size n	MI	Dice	I	Lgl	freq
A (≥ 10)	500	325	325	342	341	353
B (≥ 5)	500	134	283	217	217	353
	1 000	328	372	458	458	513
	1 500	570	585	618	618	655
	2 000	169	310	749	749	780
C (≥ 3)	500	30	60	121	113	353
	1 000	71	135	254	254	513
	1 500	111	217	392	392	655
	2 000	169	310	548	548	780

Table 5.1: Number of collocations identified by the association measures including frequency; $\text{collocations}_{all}$

$\text{collocations}_{SVC,figur}$						
set	sample size n	MI	Dice	I	Lgl	freq
A (≥ 10)	500	214	189	180	180	166
B (≥ 5)	500	98	128	162	162	166
	1 000	246	253	330	330	273
	1 500	441	393	437	437	385
	2 000	141	236	518	518	495
C (≥ 3)	500	29	44	104	96	166
	1 000	63	105	206	206	273
	1 500	95	167	327	327	385
	2 000	141	436	468	468	495

Table 5.2: Number of collocations identified by the association measures including frequency; $\text{collocations}_{SVC,figur}$

Experiment Ia

The initial research hypothesis H_1 and its related H_0 are:

H_1: MI, $Dice$, I, Lgl and $freq$ differ in their ability to identify collocations.

H_0: The models are equally well suited for collocation identification.

The χ^2-values for collocations$_{all}$ and collocations$_{SVC,figur}$ are given in tables 5.3 and 5.4, respectively.

collocations$_{all}$			
set	sample size n	χ^2	significance level when $df = 4$
A (≥ 10)	500	5.33	.30 n.s.
B (≥ 5)	500	215.56	.001
	1 000	90.54	.001
	1 500	12.09	.02
	2 000	838.43	.001
C (≥ 3)	500	656.85	.001
	1 000	617.57	.001
	1 500	634.14	.001
	2 000	623.37	.001

Table 5.3: Results: χ^2-values for the differences between MI, $Dice$, I, Lgl and $freq$ with respect to collocations$_{all}$; n.s. = not significant

The data show that there are significant differences between the association measures for collocations$_{all}$ samples B and C, and for all samples derived from collocations$_{SVC,figur}$. Thus H_0, the assumption that there is no significant difference between the measures, must be rejected in all of these cases, i.e., there are significant differences between the models except for one case:

H_0 cannot be rejected with respect to collocations$_{all}$, set A. H_0 is not in the region of rejection, as the observed value $.30 > p < .20$ is above the critical value for rejection of H_0 which is $\alpha = .05$. Therefore the models do not differ with respect to this particular case.

Interpretation

There is a significant difference between the association models including frequency for sets A collocations$_{SVC,figur}$, and B and C collocations$_{all}$ and collocations$_{SVC,figur}$. Thus in all cases but one, at least one model is better or worse than the rest. The exception is set A collocations$_{all}$, where all models (including frequency) are equally well suited for identifying collocations from PNV-fullform data.

collocations$_{SVC,figur}$			
set	sample size n	χ^2	significance level when $df = 4$
A (≥ 10)	500	10.83	.05
B (≥ 5)	500	34.26	.001
	1 000	32.93	.001
	1 500	9.82	.05
	2 000	418.27	.001
C (≥ 3)	500	163.31	.001
	1 000	204.06	.001
	1 500	281.21	.001
	2 000	269.93	.001

Table 5.4: Results: χ^2-values for the differences between MI, $Dice$, I, Lgl and $freq$ with respect to collocations$_{SVC,figur}$

Experiment Ib

As we have learned from the results of applying the χ^2 test for k-samples, the models differ significantly in almost all cases. Thus the question arises whether a best model can be identified. To answer this question, the two models with the highest number of collocations identified will be compared. In cases where there is only a minimal difference between the best results – such as I, Lgl and $freq$ in B, $n = 500$, collocations$_{SVC,figur}$ (table 5.2) – the models are assumed to be equally good, and the significance of the differences is calculated between the similar models and the next best one which in our example is $Dice$. The number of collocations identified by I and Lgl, in general, is identical in the majority of cases, which is an empirical proof for the similarity of the two measures.

The research hypothesis employed in experiment Ib is:

H_1: There are differences between the models identifying the first and second highest number(s) of true collocations.

H_0: There is no difference between the two best models.

Tables 5.5 and 5.6 show the values gained by applying the χ^2 test for two independent samples to the two measures or groups of measures with highest recall of true collocations. From the results, we see that H_0 must be partially rejected:

On the one hand, there is a significant difference between the two best performing models or groups of models in the case of sets B, $n = 500$, 1 000 collocations$_{all}$ and collocations$_{SVC,figur}$, all sets C collocations$_{all}$, and sets C, $n = 500$, 1 000, 1 500 collocations$_{SVC,figur}$.

On the other hand, H_0 cannot be rejected for set A collocations$_{SVC,figur}$, for sets B, $n = 1\ 500$, 2 000 collocations$_{all}$ and collocations$_{SVC,figur}$, and for set C, $n = 2\ 000$ collocations$_{SVC,figur}$, i.e., no single best model can be identified here.

collocations$_{all}$					
set	sample size n	measure(s)		χ^2	significance level when $df = 1$
		best	second best		
A (≥ 10)	500	no significant difference between the measures			
B (≥ 5)	500	freq	Dice	20.57	.001
	1 000	freq	I, Lgl	5.84	.02
	1 500	freq	I, Lgl	1.77	.20 n.s.
	2 000	freq	I, Lgl	0.95	.35 n.s.
C (≥ 3)	500	freq	I	214.02	.001
	1 000	freq	I, Lgl	140.77	.001
	1 500	freq	I, Lgl	100.71	.001
	2 000	freq	I, Lgl	60.15	.001

Table 5.5: Results: χ^2-values comparing the best association measures; collocations$_{all}$; n.s. = not significant

Interpretation

The result gained from set A, collocations$_{SVC,figur}$, i.e., that there is no significant difference between the highest scoring models MI and $Dice$, provides empirical support for the mathematically motivated assumption that MI and $Dice$ are comparable. This is at least the case for test samples consisting of high frequency data. In this particular case, the models outperform mere frequency. This conclusion can be drawn, because experiment Ia has shown that the results from the models including frequency differ significantly (table 5.4). Moreover the number of collocations identified by frequency is smallest, see table 5.2.

The results from sets B parallel each other, i.e., using the smaller test samples – with the $n = 500$, 1 000 highest ranked PNV-combinations – allows a single best measure (group of measures) to be identified, whereas this is not the case for the larger samples of B with $n = 1\ 500, 2\ 000$. Considering the samples taken

collocations$_{SVC,figur}$					
set	sample size n	measure(s) best	measure(s) second best	χ^2	significance level when $df = 1$
A ($\geq$ 10)	500	MI	Dice	2.39	.20 n.s.
B ($\geq$ 5)	500	freq, I, Lgl	Dice	6.59	.02
	1 000	I, Lgl	freq	7.45	.01
	1 500	MI, I, Lgl	Dice	3.67	.10 n.s.
	2 000	I, Lgl	freq	0.64	.45 n.s.
C ($\geq$ 3)	500	freq	I	18.89	.001
	1 000	freq	I, Lgl	11.96	.001
	1 500	freq	I, Lgl	5.98	.02
	2 000	freq	I, Lgl	0.93	.35 n.s.

Table 5.6: Results: χ^2-values comparing the best association measures; collocations$_{SVC,figur}$; n.s. = not significant

from sets B which show a significant difference between the models with highest collocation recall, frequency turns out to be among the best models in three of four cases, namely for B collocations$_{all}$ $n = 500$, 1 000 and B collocations$_{SVC,figur}$ $n = 500$. I and Lgl outperform $freq$ with respect to B collocations$_{SVC,figur}$ $n = 1\ 000$.

The results gained from set C clearly show that occurrence frequency outperforms the statistical measures. Frequency is significantly better in all cases but one: No significant difference between the best models $freq$, I and Lgl could be found with respect to collocations$_{SVC,figur}$, $n = 2\ 000$.

Summing up, the following reasons are evident for the partial superiority of $freq$ over the statistical association measures: (1) High frequency is a major characteristic of pseudo-collocations, and also an indicator for collocativity in general. Thus frequency is a particularly good identifier for collocations$_{all}$. (2) Selection of collocations by mere frequency leads to a cut-off of low frequency occurrences, which has particularly strong effects on collocation identification from sets C, because a large portion of low frequency data is cut off. (3) Statistical measures tend to overestimate low frequency occurrences which leads to prediction of false collocation candidates in sets with large proportions of infrequent data. Accordingly, the performance of the association measures is very poor with respect to set C. The potential for overestimation of infrequent data is clearly smaller in set B which comprises only word combinations where $c \geq 5$, and is no factor in set A where $c \geq 10$. All in all, set B is a more fair test suite for statistical association measures, because it is less biased towards low frequency co-occurrences than C and, on the other hand, it is much more demanding than set A, as collocation density in B is less high than in set A.

5.3.2 Experiment II

While in experiment I no difference between collocation types has been made, experiment II aims at investigating the feasibility of the association measures for identifying SVCs on the one hand, and figurative expressions on the other hand. Pseudo-collocations are left out from consideration, as it could be concluded from experiment I that they are best identified by high co-occurrence frequency using full form data. The same procedures as in experiment I are now applied to the individual collocation classes. In the ideal case, a best association model is identified for SVCs and another one for figurative expressions. The results gained by experimenting with the set of P.N.V(full form)-triples are compared with the results based on the set of P.N.V(base form)-triples. The two sets have been chosen because of the reverse distribution of SVCs and figurative expressions within the sets, see figure 3.6 at page 87. Thus it is expected that the differences between full and base form data also affect the preformance of the models.

Two sampling strategies are employed: On the one hand, the statistical association measures, MI, $Dice$, I and Lgl are compared against each other. On the other hand, co-occurrence frequency $freq$ is also included in the comparison. The distinction has been made in order to find out (i) whether MI and $Dice$ versus I and Lgl also form two classes when identifying SVCs or figurative expressions, and (ii) whether the results gained by applying mere frequency are comparable to the results achieved by the statistical models.

Experiment IIa

The research hypotheses to be pursued are:

for SVCs:

$H1_{SVC}$: The lexical association models differ in their feasibility to identify SVCs.

$H0_{SVC}$: There are no differences between the association models with respect to SVCs.

for figurative expressions:

$H1_{figur}$: The lexical association models differ in their feasibility to identify figurative expressions.

$H0_{figur}$: There are no differences between the association models with respect to figurative expressions.

Tables 5.7 and 5.8 show the results from applying the χ^2 test for k-samples. As can be seen from the tables, $H0_{SVC}$ must be rejected for all samples A, B and C except one, which is set C, $n = 500$, base forms excluding frequency. In other words, in all cases but one there are significant differences between the models when employed for identifying SVCs from full and base form data.

$H0_{figur}$ cannot be rejected for set A, but it must be partially rejected for sets B and C, in particular: $H0_{figur}$ must be rejected for all samples taken from set C when frequency is one of the models tested, i.e., in this particular case, the models differ significantly. $H0_{figur}$ must be rejected for six out of eight samples taken from set C when frequency is not among the models tested. Thus there is also a significant difference between the models in the majority of cases C excluding frequency. With respect to set B, $H0_{figur}$ must be rejected in three out of eight cases when only statistical association measures are considered. The same number of cases where $H0_{figur}$ must be rejected was also found for set B when frequency is taken into account.

Lexical association measures without *freq*									
		P.N.V(full forms)				P.N.V(base forms)			
set	n	SVC		figur		SVC		figur	
		χ^2	signif. level	χ^2	signif. level	χ^2	signif. level	χ^2	signif. level
A	500	151.57	.001	1.34	n.s.	24.08	.001	4.78	n.s.
B	500	44.65	.001	0.11	n.s.	42.82	.001	29.52	.001
	1 000	43.25	.001	0.16	n.s.	95.34	.001	1.03	n.s.
	1 500	8.95	.05	1.75	n.s.	84.73	.001	2.48	n.s.
	2 000	386.82	.001	29.95	.001	79.98	.001	11.95	.01
C	500	97.89	.001	9.51	.05	2.09	n.s.	52.49	.001
	1 000	175.27	.001	7.34	n.s.	37.43	.001	45.58	.001
	1 500	273.15	.001	11.2	.02	109.31	.001	33.66	.001
	2 000	374.11	.001	13.57	.01	157.73	.001	7.14	n.s.

Table 5.7: Differences between MI, $Dice$, I and Lgl; df = 3; n.s. = not significant

Interpretation

SVCs: Totally clear results have been achieved for identifying SVCs from sets A and B of full and base form data: significant differences are found between the statistical models, and also when the frequency-based strategy is taken into account. In other words, for each constellation tested there must be at least a model which is distinct from the others.

Lexical association measures including $freq$									
		P.N.V(full forms)				P.N.V(base forms)			
set	n	SVC		figur		SVC		figur	
		χ^2	signif. level	χ^2	signif. level	χ^2	signif. level	χ^2	signif. level
A	500	150.08	.001	2.05	n.s.	24.69	.001	5.37	n.s.
B	500	56.1	.001	0.55	n.s.	101.63	.001	39.46	.001
	1 000	43.24	.001	2.3	n.s.	125.48	.001	3.79	n.s.
	1 500	12.33	.02	2.69	n.s.	93.88	.001	3.96	n.s.
	2 000	400.85	.001	39.58	.001	79.76	.001	12.04	.01
C	500	231.74	.001	21.91	.001	360.11	.001	105.09	.001
	1 000	225.63	.001	20.42	.001	323.63	.001	103.1	.001
	1 500	303.79	.001	30.27	.001	262.57	.001	139.27	.001
	2 000	391.31	.001	29.54	.001	253.75	.001	89.97	.001

Table 5.8: Differences between MI, $Dice$, I, Lgl and $freq$; df = 4; n.s. = not significant

With respect to set C, there is one exception to the previous result: no significant difference between the statistical association models can be found when the 500 highest ranked base form data (C, $n = 500$) are considered. The differences between the models, however, become significant when $freq$ is taken into account. All in all – except for one case (C, $n = 500$, base forms) – there are significant differences between the models irrespective of threshold, of including or excluding $freq$, and of full forms or base forms.

Figurative expressions: Other than for SVCs, a smaller number of significant differences can be detected for figurative expressions.

For set A, no differences between the models are found, i.e., all models perform equally well supporting the assumption of equal performance of the models when applied to high frequency data.

With respect to set B, no difference between the models are detected for base form samples where $n = 1\ 000, 1\ 500$, and for full form samples where $n = 500$, 1 000, 1 500. The results hold for association measures including and excluding frequency, which means that $freq$ performs neither better nor worse than the other models. The composition of the data (full versus base forms), however, has a slight effect on the significance of model differences, which will be described below.

Considering set C, there are two cases where $freq$ differs significantly from the statistical association measures, these are C, $n = 1\ 000$ full forms, and C, $n = 2\ 000$ base forms.

Full versus base forms: Changes in the significance of the model differences between full and base form data can be found: (i) In set B, $n = 500$ where the models differ significantly for identifying figurative expressions from base form data, but not from full form data. This holds for including and excluding $freq$. (ii) Similarly the significances for the differences between the statistical association measures vary between base form and full form figurative expressions for sets C, $n = 1\ 000, 2\ 000$, and between base form and full form SVCs for set C, $n = 500$. Thus it can be concluded that inflectional constraints in the candidate data influence the applicability of the association models.

Experiment IIb

While experiment IIa has revealed general tendencies concerning the differences between the lexical association models, the current experiment aims at identifying single best models. The following hypotheses are tested:

for SVCs:

$H1_{SVC}$: There are single best models for identifying SVCs from set A, B and C of full and base form data.

$H0_{SVC}$: The first and second best models do not differ for SVCs.

for figurative expressions:

$H1_{figur}$: There are single best models for identifying figurative expressions from set A, B and C of full and base form data.

$H0_{figur}$: The first and second best models do not differ for figurative expressions.

Tables 5.9 to 5.12 show the results achieved by the two best (groups of) models identifying SVCs or figurative expressions from P.N.V(full form)- and P.N.V(base form)-triples. Again the examination is twofold. On the one hand, statistical models (MI, $Dice$, I, Lgl) are compared against each other (assoc. meas. excl. freq). On the other hand, mere co-occurrence frequency ($freq$) is also included (assoc. meas. incl. freq).

SVCs: $H0_{SVC}$ cannot be rejected for set A, i.e., no single best model can be identified. Whereas $H0_{SVC}$ must be partially rejected for sets B and C.

$H0_{SVC}$ must be rejected for B, base form data when only statistical association measures are compared. In all of these cases two best models could be identified (I, Lgl)[1]. If frequency is also taken into account, $H0_{SVC}$ cannot be rejected for $n = 1\ 000$ and $n = 1\ 500$. In these cases, I, Lgl and $freq$ perform

---

[1] I and Lgl here select the same number of SVCs.

equally well. In the two remaining cases, $freq$ outperforms I and Lgl ($n = 500$), I and Lgl outperform $freq$ ($n = 2\ 000$).

Support-Verb Constructions assoc. meas. excl. freq					
set	n	full forms		base forms	
		best mods	χ^2	best mods	χ^2
A	500	MI Dice 134 118	1.19 n.s.	I/Lgl MI 112 89	3.01 n.s.
B	500	I/Lgl Dice 90 58	7.62 .01	I/Lgl Dice 59 27	12.23 .001
	1 000	I/Lgl Dice 201 129	18.29 .001	I/Lgl Dice 133 56	33.75 .001
	1 500	I/Lgl MI 269 253	0.52 n.s.	I/Lgl MI 192 94	36.37 .001
	2 000	I(Lgl) Dice 310(298) 89	134.74 .001	I/Lgl MI 251 180	12.74 .001
C	500	I(Lgl) Dice 54(51) 9	32.8 .001	no signif. diff. between the models	
	1 000	I/Lgl Dice 124 27	66.02 .001	I(Lgl) Dice 38(36) 18	6.63 .01
	1 500	I/Lgl Dice 205 54	95.08 .001	I(Lgl) Dice 92(91) 29	33.1 .001
	2 000	I/Lgl Dice 298 89	123.77 .001	I(Lgl) Dice 134(133) 41	50.58 .001

Table 5.9: Results: the best association models for identifying SVCs from PNV-full and -base forms comparing MI, $Dice$, I and Lgl; df = 1; n.s. = not significant

With respect to set B full form data, $H_{0_{SVC}}$ must be largely rejected for the statistical measures, i.e., in three out of four cases I and Lgl differ significantly from the second best model which is $Dice$. In the case where $n = 1\ 500$, there is no difference between the highest ranking models which are I, Lgl and MI. When frequency is taken into account, $H_{0_{SVC}}$ cannot be rejected in three out of four cases, i.e., there are no significant differences between the highest ranking models for the samples where $n = 500$, 1 500, 2 000. For $n = 500$ and $n = 2\ 000$ I, Lgl and $freq$ do not differ in their feasibility for identifying SVCs. For the sample where $n = 1\ 500$, I, Lgl and MI are the highest ranking measures. $H_{0_{SVC}}$, however, must be rejected for the sample where $n = 1\ 000$; here I and

Lgl outperform *freq*.

Support-Verb Constructions assoc. meas. incl. freq					
set	n	full forms		base forms	
		best mods	χ^2	best mods	χ^2
A	500	MI Dice 134 118	1.19 n.s.	I/Lgl freq 112 103	0.38 n.s.
B	500	freq I/Lgl 101 90	0.65 n.s.	freq I/Lgl 103 59	13.62 .001
	1 000	I/Lgl freq 201 163	4.6 .05	freq I/Lgl 159 133	2.51 n.s.
	1 500	I/Lgl MI 269 253	0.52 n.s.	I/Lgl freq 192 189	0.01 n.s.
	2 000	I(Lgl) freq 310(298) 288	0.23 n.s.	I/Lgl freq 251 210	3.92 .05
C	500	freq I(Lgl) 101 54(51)	16.16 .001	freq I/Lgl/Dice 103 4	100.51 .001
	1 000	freq I/Lgl 163 124	5.87 .02	freq Lgl 159 38	81.08 .001
	1 500	freq I/Lgl 223 205	0.79 n.s.	freq I 189 92	36.19 .001
	2 000	I/Lgl freq 298 288	0.16 n.s.	freq I/Lgl 210 134(133)	17.46 .001

Table 5.10: Results: the best association models for identifying SVCs from PNV-full and -base forms comparing *MI*, *Dice*, *I*, *Lgl*, mere occurrence frequency *freq*; df = 1; n.s. = not significant

$H0_{SVC}$ must be rejected for set C base forms if frequency is taken into account. In all of these cases, *freq* is the best model. $H0_{SVC}$ must also be rejected for set C full forms when only the statistical models are tested. Here in all cases *I* and *Lgl* outperform *Dice*. $H0_{SVC}$ must be partially rejected for set C full forms, if frequency is taken into account: *freq* outperforms *I* and *Lgl* for $n = 500$ and $n = 1\ 000$. In the two remaining cases no difference between the three models could be found. $H0_{SVC}$ must as well be partially rejected for set C base forms, if only the statistical models are compared. Here *I* and *Lgl* outperform *Dice* for $n = 1\ 000$, 1 500, 2 000.

Figurative expressions: $H_{0_{figur}}$ must not be rejected for set B when the statistical association measures and frequency are compared. With respect to this sample there is either no difference between the models at all (cf. experiment IIa), or there is no difference between the highest ranking models.

Figurative Expressions assoc. meas. excl. freq					
set	n	full forms		base forms	
		best mods	χ^2	best mods	χ^2
A	500	no significant differences between the measures			
B	500	no signif. diff. between the measures		MI(Dice) I/Lgl 62(61) 27	14.26 .001
	1 000	no significant differences between the measures			
	1 500	no significant differences between the measures			
	2 000	I Lgl 208 170	8.84 .01	MI I/Lgl 232 192	4.01 .05
C	500	I(Lgl) Dice 50(45) 35	2.52 n.s.	Dice MI 41 33	0.72 n.s.
	1 000	no signif. diff. between the measures		Dice MI 70 50	3.2 n.s.
	1 500	I/Lgl Dice 122 113	0.3 n.s.	Dice MI 95 80	1.19 n.s.
	2 000	I/Lgl Dice 170 147	1.66 n.s.	no signif. diff. between the measures	

Table 5.11: Results: the best association models for identifying SVCs from PNV-full and -base forms comparing *MI*, *Dice*, *I* and *Lgl*; df = 1; n.s. = not significant;

$H_{0_{figur}}$, however, must be rejected for set B, if only the statistical models are compared: There are only 3 applicable cases left after experiment IIa, i.e., (1) B full forms $n = 2\ 000$, with I being significantly better than Lgl; (2) B base forms $n = 500$, with MI and $Dice$ outperforming I and Lgl; and (3) B base forms $n = 2\ 000$, with MI outperforming I and Lgl.

$H_{0_{figur}}$ must also be rejected for set C of base form triples when $freq$ is taken into account. In other words, $freq$ is the best model for identifying figurative expressions from P.N.V(base from) triples.

On the other hand, $H_{0_{figur}}$ cannot be rejected for the sets C of full and base form data, when only statistical models are compared, i.e., all statistical models perform equally well. The picture, however, changes for set C $n = 1\ 000,\ 1\ 500$ of full forms when *freq* is taken into account. In these cases, *freq* outperforms the statistical models. For the remaining two cases ($n = 500,\ 2\ 000$), the models are equally well suited.

Figurative Expressions assoc. meas. incl. freq					
set	n	full forms		base forms	
		best mods	χ^2	best mods	χ^2
A	500	no significant differences between the measures			
B	500	no signif. diff. between the measures		freq MI 71 62	0.56 n.s.
	1 000	no significant differences between the measures			
	1 500	no significant differences between the measures			
	2 000	I freq 208 207	0.87 n.s.	MI freq 232 202	2.17 n.s.
C	500	freq I 65 50	1.93 n.s.	freq Dice 71 41	8.46 .01
	1 000	freq I/Lgl 110 82	4.2 .05	freq Dice 121 70	14.47 .001
	1 500	freq I/Lgl 162 122	5.92 .02	freq Dice 173 95	24.29 .001
	2 000	freq I/Lgl 207 170	3.795 n.s.	freq Dice 202 112	27.38 .001

Table 5.12: Results: the best association models for identifying SVCs from PNV-full and -base forms comparing *MI*, *Dice*, *I*, *Lgl* and mere occurrence frequency *freq*; df = 1; n.s. = not significant;

Interpretation

The following general tendencies could be observed:

With respect to SVCs:

- *MI* and *Dice* are the highest ranking models for identifying SVCs from set A of full forms.
- *I*, *Lgl* and *freq* are the highest ranking models for identifying SVCs from set A of base forms.

- I and Lgl are the best statistical models for identifying SVCs from sets B and C of full and base forms.
- $Freq$ is always among the best models when SVCs are identified from sets B and C full and base form data. It performs best for C base forms and is equal to I and Lgl in the other samples taken from B and C.

With respect to figurative expressions:

- Other than for SVCs, there is no such clearcut difference between the models in identifying figurative expressions.
- I, Lgl, $Dice$, MI and $freq$ are equally well suited for identifying figurative expressions from set A of full and base form data.
- I, Lgl, $Dice$, MI and $freq$ are in the majority of cases also equally well suited for identifying figurative expressions from set B of full and base form data. There is, on the one hand, a slight preference for MI in the case of base form data and, on the other hand, a slight preference for I in the case of full form data.
- I, Lgl and $Dice$ in the majority of cases outperform MI when figurative expressions are identified from set C of full form data, provided $freq$ is not taken into account.
- $Dice$ and MI are the best statistical models when figurative expressions are identified from set C of base form data, provided $freq$ is not taken into account.
- $Freq$ outperforms the statistical models when figurative expressions are identified from set C of base form data, and $freq$ is among the best models for the full form data.

With respect to full and base form data:

- In the case of full form data compared to base form data, the numbers of true collocations identifed approximate for the best models. This is due to the fact that
- the performance of statistical association models, especially I and Lgl, strongly increases from base to full forms; but
- the performance of $freq$ stays approximately the same.

5.4 Evaluation of the Kwic-Based Model

As shown in section 4.3.1, a kwic-based reduction of the test samples leads to an increase of the proportion of SVCs and figurative expressions with a stronger increase of SVCs. Thus it is expected that a kwic-based strategy where support-verbs are employed as lexical keys significantly improves identification accuracy for SVCs, while the effect on figurative expressions is expected to be less strong. In order to investigate the assumption, the kwic-model is compared with the best performing association models (see experiments IIIa and IIIb). In addition, the kwic-strategy and two of the best models for identifying SVCs, namely I and $freq$, are combined, and compared against each other (see experiment IV).

5.4.1 Experiment III

Two sets of candidate data are employed for testing: P.N.VVPP-trigrams and P.N.V(base form)-triples. The former is of interest, because the proportion of SVCs among the data is very high. The latter set is employed, because, in contrast to the former, it contains a larger number of figurative expressions than SVCs, and the proportion of the collocations in general is very low.

The experiments are performed according to the following procedure: The kwic-strategy is applied to each set A, B and C of the candidate data. This way, for each set all PNV-combinations but the ones containing a verb which is among the lexical keys specified in section 4.4.3 are discarded. The number of true and false collocations identified by this procedure is used for comparison with the best result achieved by the association models for sets A, B and C. The χ^2 test for two independent samples is applied for testing the significance of the difference between the kwic-based strategy and the one single best model for sets A, B and C, respectively.

The one single best model is determined according to the following procedure: First of all, a single best combination of association measure and sample must be identified for sets B , $n = 500$ to $n = 2\ 000$ and C, $n = 500$ to $n = 2\ 000$. To achieve this goal, χ^2 tests are employed. If no significant differences between the samples can be found, the F-score is used to identify the “best” sample.[2] F-score is a measure that combines recall and precision into one value. Recall is defined as $\frac{\#true\ collocations\ found}{\#true\ collocations\ total}$. Precision is defined as $\frac{\#true\ collocations\ found}{\#candidates\ retrieved}$. # stands for “number of”.

The formula for computing the F-score is taken from [Carrol *et al.*, 1999].

$$Fscore = \frac{2 * recall * precision}{recall + precision}$$

---

[2] Best here is under quotes, because due to statistical insignificance no strictly best sample exists.

Experiment IIIa

In the following, the kwic-strategy and the single best measures are compared employing the set of P.N.VVPP-trigrams.

The research hypotheses are:

for SVCs:

$H1_{SVC}$**:** The kwic-model and the single best model differ with respect to the identification of SVCs from sets A, B and C.

$H0_{SVC}$**:** The kwic-model and the single best model do not differ with respect to identification of SVCs.

for figurative expressions:

$H1_{figur}$**:** The kwic-model and the single best model differ with respect to the identification of figurative expressions from sets A, B and C.

$H0_{figur}$**:** The kwic-model and the single best model do not differ with respect to identification of figurative expressions from sets A, B and C.

See table 5.13 for the data. Note that the numbers for figurative expressions in sets A and B are particularly small. Thus no significant results are possible. As the sets A, B and C are much smaller here than the sets employed in the tests before, only one sample per set A, B and C has been selected. The information associated with sets A, B and C is illustrated with respect to set A SVC: set A contains a total number of 33 SVCs. The single best models allow 26 - 29 SVCs to be identified; "all meas" indicates that there is no significant difference between the measures, and $n = 40$ means that the 40 highest ranked PNV-combinations per measure have been considered; "kwic 29" indicates that 29 SVCs are among the set of PNV-combinations selected by applying the kwic-strategy; the total number of word combinations identified by means of the kwic-model "kwic total" is 33; the χ^2 value is 1.76 resulting from comparing the best association measure ("best meas") with the kwic-model "kwic". As the observed $\chi^2 = 1.76$ is below the required theoretical value $p = 3.84$ for $\alpha = .05$ with $df = 1$, H_0 cannot be rejected, and thus the difference between the kwic and the association models is not significant ("n.s.").

As can be seen from the χ^2 values in table 5.13, $H0_{SVC}$ must be rejected for sets B and C, whereas $H0_{figur}$ must be rejected only for set C, i.e., there is a significant difference between the best association model and the kwic-model with respect to identifying SVCs from sets B and C, and with respect to identifying figurative expressions from set C.

P.N.VVPP-trigrams			
		SVC	figur
A	total best meas	33 26 - 29 (all meas, n = 40)	6 4 (all meas, n = 40)
	kwic	29	1
	kwic total = 33		
	χ^2	1.76 n.s.	0.5 n.s.
B	total best meas	60 55, 54, 52 (I,Lgl, freq, n = 100)	17 12 - 15 (all meas, n = 100)
	kwic	**52**	7
	kwic total = 72		
	χ^2	4.57	0.63 n.s.
C	total best meas	96 75 (I, n = 150)	35 **22** (Dice, n = 150)
	kwic	**85**	8
	kwic total = 129		
	χ^2	6.23	4.33

Table 5.13: Comparison of the kwic-model and the single best measure; numbers in bold face indicate the superior model; df = 1

Interpretation

The results based on P.N.VVPP-trigrams can be interpreted as follows: As expected, the kwic-approach is significantly more accurate for identifying SVCs from sets B and C, than this is the case for the single best models including mere co-occurrence frequency. For set A, no significant differences between the models could be found. This may be attributed to the fact that SVC-density is highest in set A, and that the set contains highly recurrent data. Because of the latter, statistics-based methods do not deteriorate as strongly as in sets where the proportion of low frequency data is high.

The results also support the expectation that the kwic-model does not improve identification of figurative expressions, which can be attributed to the following three factors: (i) the lexical keys employed better suit SVCs than figurative expressions; (ii) the set of P.N.VVPP-trigrams is strongly biased towards SVCs; (iii) the number of figurative expressions is small, especially in A, and therefore significant differences are difficult to obtain.

Experiment IIIb

Here the same hypotheses are tested as in experiment IIIa, but with respect to P.N.V(base forms), a candidate sample with a very low density of SVCs as well as figurative expressions.

According to table 5.14, $H0_{SVC}$ must be rejected for sets A and C, whereas $H0_{figur}$ must be rejected for set C. In other words, the differences between the kwic-model and the best association models are significant in these particular cases.

Interpretation

When P.N.V(base form)-triples are used as a basis for identification, there is only one case where the kwic-strategy leads to a significantly better result than the other strategies, which is the identification of SVCs from set A. This may be mainly attributed to the fact that the verbal keys are particularly representative for frequent SVCs. With respect to set B there is no significant difference in accuracy between a kwic-based data reduction and a frequency-based reduction. The result from set C shows that a frequency based cut-off is significantly better than a kwic-based data reduction when the proportion of SVCs among the data is low. On the other hand, recall is much higher using the kwic-approach.

A similar observation can be made with respect to identifying figurative expressions from set C. Here too, $freq$ outperforms the kwic-strategy, whereas recall is much higher employing the kwic-model. For identifying figurative expressions from sets A and B, however, there is no significant difference in accuracy between the kwic-based approach and the best association measure. The results provide evidence that there is a set of verbs which occur in SVCs and figurative expressions.

P.N.V(base form)			
		SVC	figur
A	total	174	150
	best meas	112	87
		(I,Lgl, n = 500)	(MI, n = 500)
	kwic	**147**	86
	kwic total = 458		
	χ^2	10.91	0.22 n.s.
B	total	304	338
	best meas	103	232
		(freq, n = 500)	MI, n = 2 000
	kwic	249	169
	kwic total = 1 252		
	χ^2	0.07 n.s.	2.4 n.s.
C	total	412	527
	best meas	**103**	**71**
		(freq, n = 500)	(freq, n = 500)
	kwic	328	238
	kwic total = 2 985		
	χ^2	35.63	19.79

Table 5.14: Comparison of the kwic-model and the single best model; numbers in bold face indicate the superior model; df = 1

5.4.2 Experiment IV

Experiment IVa

In the following, it is investigated whether combining the kwic-strategy and a high performing association measure leads to an improvement in identifying SVCs. A lesson learned from experiments II is that I, Lgl and $freq$ are the best performing association measures for identifying SVCs from base form data. Accordingly, I/Lgl[3] on the one hand, and $freq$ on the other hand are combined with the kwic-strategy. As on the one hand the kwic-strategy is particularly designed for identifying SVCs, and on the other hand identification of SVCs from P.N.V(base forms) is assumed to be hard because of the low percentage of SVCs among the data, it is expected that employing the kwic-strategy will

---

[3] As I and Lgl, if at all, differ only marginally, the I-values are used in experiment IV.

increase identification accuracy. The following hypotheses are tested:

for I

H_{1_I}**:** The I+kwic-model differs significantly from I with respect to identifying SVCs from sets A, B and C of P.N.V(base form)-triples.

H_{0_I}**:** The I+kwic-model does not differ from I with respect to identifying SVCs from sets A, B and C of P.N.V(base form)-triples.

for $freq$

$H_{1_{freq}}$**:** The freq+kwic-model differs significantly from $freq$ with respect to identifying SVCs from sets A, B and C of P.N.V(base form)-triples.

$H_{0_{freq}}$**:** The freq+kwic-model does not differ from $freq$ with respect to identifying SVCs from sets A, B and C of P.N.V(base form)-triples.

As can be seen from the data in table 5.15, H_{0_I} must be partially rejected, i.e., there is a significant difference between the I+kwic- and the I-model for sets B $n = 1\ 500, 2\ 000$. In all other cases, the combined and the simple model do not differ significantly.

P.N.V(base form)-triples						
set	n	I	I+kwic	kwic total	χ^2	signif. level
A	500	112	105	380	2.91	n.s.
B	500	59	59	423	0.77	n.s.
	1 000	133	124	756	3.07	n.s.
	1 500	192	**177**	1 057	7.5	.01
	2 000	251	**230**	1 191	26.13	.001
C	500	4	4	419	0.01	n.s.
	1 000	38	36	864	0.08	n.s.
	1 500	92	84	1 020	3.812	n.s.
	2 000	134	127	1 520	3.21	n.s.

Table 5.15: Comparison of I+kwic and I; n.s. = not significant; the recall values of the better model are printed in bold face; df = 1

The data in table 5.16 reveal that $H_{0_{freq}}$ must be rejected in all cases. In other words, there is a difference between freq+kwic and $freq$ in all cases examined.

P.N.V(base form)-triples						
set	n	freq	freq + kwic	kwic total	χ^2	signif. level
A	500	103	**88**	220	28.51	.001
B,C	500	103	**88**	220	28.51	.001
	1 000	159	**134**	391	55.96	.001
	1 500	189	**159**	523	85.03	.001
	2 000	210	**176**	653	105.86	.001

Table 5.16: Comparison of freq+kwic and freq; the recall values of the better model are printed in bold face; df = 1

Interpretation

An I- or Lgl-based approach to identification of SVCs from base form data does not gain in accuracy from employing the kwic-strategy. In both cases, approximately the same number of SVCs is identified. As the samples selected by the association measures cannot be significantly reduced by the kwic-strategy, the difference between the approaches is marginal. The only exceptions are sets B $n = 1\ 500,\ 2\ 000$ where the combined model has a better accuracy than I.

The freq+kwic-model on the other hand is significantly better than simply applying $freq$. This holds for sets A, B and C. Because of a frequency-based cut-off, the results are not determined by the complete samples A, B and C, but only by the n most frequent word combinations which are the same in A, B, C $n = 500$, as well as in B and C $n = 1\ 000 \ldots 2\ 000$. A further data reduction by employing the kwic-strategy discards a small number of true collocations, but a considerably large number of noncollocations, which drastically increases the percentage of true collocations among the remaining data.

Experiment IVb

Finally, it is investigated whether the I+kwic-model and the freq+kwic model differ with respect to identifying SVCs. The according research hypothesis is:

H_1: The I+kwic- and the freq+kwic-model differ significantly with respect to identifying SVCs from sets A, B and C of P.N.V(base form)-triples.

H_0: The I+kwic- and the freq+kwic-model do not differ with respect to identifying SVCs from sets A, B and C of P.N.V(base form)-triples.

Based on the results in table 5.17, H_0 must be rejected in all cases, which means that the two combined models differ significantly in accuracy.

P.N.V(base form)-triples						
set	n	I + kwic	total	freq + kwic	total	χ^2
A	500	105	380	**88**	220	9.21
B	500	59	423	**88**	220	54.23
	1 000	124	756	**134**	391	46.18
	1 500	177	1 057	**159**	523	38.16
	2 000	230	1 191	**176**	653	13.9
C	500	4	419	**88**	220	175.29
	1 000	36	864	**134**	391	205.75
	1 500	84	1 020	**159**	523	126.36
	2 000	127	1 520	**176**	653	130.11

Table 5.17: Identification of SVCs from the set of P.N.V(base form)-triples; the recall values of the better model are printed in bold face; df = 1

Interpretation

Applying the kwic-strategy subsequently to *I* and the frequency-based approach leads to the following results: Accuracy in identifying SVCs is significantly higher for the model combining frequency and the kwic-strategy than the model combining *I* and the kwic-strategy.

Summing up, the kwic-model is well suited for identification of SVCs, but only when SVC-density is high in the set used for identification. This can, for instance, either be achieved by using an adequate base set, such as the set of P.N.VVPP-trigrams, or by applying a (statistical) measure which is suitable for identification of SVCs. Thus, the kwic-model is not a general alternative to statistical association measures, but leads to higher identification accuracy when combined with other strategies that increase the number of SVCs in the set of collocation candidates. *Freq* + *kwic* has turned out to be the best combination for the samples under investigation.

5.5 Evaluation of the Entropy Model

Based on the PP-entropy values of selected PNV-combinations derived from the extraction corpus, a threshold $t = 0.7$ has been empirically determined which divides the PN-tuples being part of a PNV-combination into potential collocates and noncollocates. PN-tuples with entropy values lower than 0.7 are considered to be collocates. Similar to the kwic-strategy, application of the entropy model drastically reduces the number of collocation candidates. See table 5.18 where the number of collocations identified by means of PP-entropy is compared to the total number of collocations in sets A, B and C. The data reveal that the entropy model is a fairly poor estimate for pseudo-collocations, as at best only 13.9 % of the pseudo-collocations contained in the P.N.V(full form)- and 12.2 % in the P.N.V(base form)-data are covered. On the other hand, low PP-entropy is particularly well suited for identification of SVCs. The method allows identifying at best 74.3 %, and in the worst case 56.7 % of the SVCs contained in the P.N.V(full form)-data, and between 51.7 % and 39.8 % in the P.N.V(base form)-data. The differences in recall between full and base form data are due to the fact that the full form data contain inflectional variants of individual SVCs. Thus the figures from the base form data provide a clearer picture of the feasibility of the entropy model for collocation identification. PP-entropy is also useful for identifying figurative expressions. In this case, recall ranges from 52.1 % with respect to set A to 45.8 % with respect to set C of the P.N.V(full form)-triples, and 35.3 % to 34.5 % of the P.N.V(base form)-triples. While recall of SVCs decrease with increasing proportion of low frequency data, recall of figurative expressions is fairly constant over sets A, B and C. The data provide evidence for a correlation between the rigidity in the PP-collocate and occurrence frequency of SVCs. Such a correlation has not been found for figurative expressions. In general, the precision in identifying SVCs and figurative expressions increases when the entropy model is applied, whereas the precision for pseudo-collocations decreases.

P.N.V(base form)-triples						
	A	H	B	H	C	H
	$c \geq 10$	< 0.7	$c \geq 5$	< 0.7	$c \geq 3$	< 0.7
$\sum_{pnv}$	1 249	249	4 489	792	14 660	2872
$\sum_{coll_{SVC}}$	174	90	304	129	412	164
$\sum_{coll_{figur}}$	150	53	338	116	527	182
$\sum_{coll_{pseudo}}$	262	26	315	37	345	42

P.N.V(full form)-triples						
	A	H	B	H	C	H
	$c \geq 10$	< 0.7	$c \geq 5$	< 0.7	$c \geq 3$	< 0.7
$\sum_{pnv}$	747	212	2864	605	10 430	2 093
$\sum_{coll_{SVC}}$	144	100	369	202	709	357
$\sum_{coll_{figur}}$	96	49	282	114	586	219
$\sum_{coll_{pseudo}}$	239	18	302	31	337	36

Table 5.18: Number of collocations identified by means of PP-entropy (H) using a threshold of 0.7

5.5.1 Experiment V

Experiments are presented which (i) investigate the difference between the entropy model and the single best models for identifying SVCs and figurative expressions (experiment Va); (ii) examine whether a combination of the entropy- and the kwic-model leads to significant differences to the simple entropy model in identifying SVCs and figurative expressions (experiment Vb); (iii) test whether the combined models, freq+kwic and entropy+kwic, differ with respect to identifying SVCs.

Experiment Va

The identification results for SVCs and figurative expressions achieved by the entropy model are compared with the best results achieved by the association models. Full and base form data are used for the investigations. The test procedure is comparable to the one described on page 140f. (one single best model). The following hypotheses are tested:

for SVCs:

$H1_{SVC}$**:** The entropy model and the single best models differ significantly with respect to identifying SVCs from sets A, B and C of P.N.V(base form)- and -(full form)-triples.

$H0_{SVC}$**:** The entropy model and the single best models do not differ for the respective sets.

for figurative expressions:

$H1_{figur}$**:** The entropy model and the single best models differ significantly with respect to identifying figurative expressions from sets A, B and C of P.N.V-(base form)- and -(full form)-triples.

$H0_{figur}$**:** The entropy model and the single best models do not differ for the respective sets.

The data in tables 5.19, p. 151 and 5.20, p. 152 reveal the following: $H0_{SVC}$ must be rejected for sets A and B of P.N.V(full form)-triples, and for sets A and C of base form triples. In these particular cases, the models significantly differ.

$H0_{figur}$ also must be rejected for sets A and B of full form triples, and B and C of base form triples. The models differ significantly.

Interpretation

For sets A and B of full form data, the entropy model leads to significantly better accuracy results in identifying SVCs and figurative expressions than the best association measure. The difference between the models is insignificant for set C. Recall of SVCs from set C of full form data, however, is 3 times higher employing the entropy model compared to the best association model. Thus the association measures should be replaced by the entropy model for identifying SVCs and figurative expressions from full form data.

The results for the base form data are much more heterogeneous. The entropy model performs significantly better than the best association model for identification of SVCs from set A, and figurative expressions from set B. In the case of set C, a simple approach based on occurrence frequency significantly outperforms the entropy model with respect to accuracy in identifying SVCs and figurative expressions. Recall, however, is higher for both SVCs and figurative expressions when the entropy model is applied to set C. On the other hand, in three out of four cases (A SVC, A figur, B figur), higher recall is achieved by the best association models. Summing up, the entropy model outperforms the association measures for identifying SVCs and figurative expressions from sets A and B base form data, because it is at least as good as or better than the best association measure. The entropy model, however, is inferior to a frequency-based approach when applied to set C base forms.

P.N.V(full form)			
		SVC	figur
A	total	144	96
	best meas	134 (MI, n = 500)	80 (MI, n = 500)
	entropy	**100**	**49**
	entropy total = 212		
	χ^2	27.08	4.61
B	total	369	282
	best meas	101 (freq, n = 500)	162 - 188 (all meas), n = 1 500
	entropy	**202**	**114**
	entropy total = 605		
	χ^2	27.27	13.46
C	total	709	586
	best meas	101 (freq, n = 500)	207 (freq, n = 2 000)
	entropy	357	219
	entropy total = 2 093		
	χ^2	2.53 n.s.	0.0046 n.s.

Table 5.19: Comparison of the entropy model and the single best model; n.s. = not significant; figures in bold face indicate the superior model; df = 1

P.N.V(base form)			
		SVC	figur
A	total best meas	174 112 (I,Lgl, n = 500)	150 87 (MI, n = 500)
	entropy	**90**	53
	entropy total = 249		
	χ^2	15.25	1.41 n.s
B	total best meas	304 103 (freq, n = 500)	338 232 MI, n = 2 000
	entropy	129	**116**
	entropy total = 792		
	χ^2	3.58 n.s.	4.55
C	total best meas	412 **103** (freq, n = 500)	527 **71** (freq, n = 500)
	entropy	164	182
	entropy total = 2 872		
	χ^2	127.46	36.81

Table 5.20: Comparison of the entropy model and the single best model; figures in bold face indicate the superior model; df = 1

Experiment Vb

In the following, a model combining entropy and the kwic-based strategy is compared with the simple entropy model. Here again, base form data are used for testing, because this is the set with the smallest percentage of SVCs and figurative expressions. Thus identification of both collocation classes is hard. The hypotheses are:

for SVCs:

$H_{1_{SVC}}$: The entropy+kwic-model and the simple entropy model differ significantly with respect to identifying SVCs from P.N.V(base form)-triples.

$H_{0_{SVC}}$: The entropy+kwic-model and the simple entropy model do not differ with respect to identifying SVCs.

for figurative expressions:

$H1_{figur}$: The entropy+kwic-model and the simple entropy model differ significantly with respect to identifying figurative expressions from P.N.V(base form)-triples.

$H0_{figur}$: The entropy+kwic-model and the simple entropy model do not differ with respect to identifying figurative expressions.

P.N.V(base form)-triples								
	entropy + kwic			entropy			χ^2	χ^2
set	SVC	figur	total	SVC	figur	total	SVC	figur
A	**86**	34	145	90	53	249	18.97	0.14 n.s.
B	**129**	**72**	347	129	116	792	58.9	6.09
C	**164**	**108**	850	164	182	2 872	148.92	36.15

Table 5.21: Identification of SVCs and figurative expression by means of the entropy model from the set of P.N.V(base form)-triples; figures in bold face indicate the superior model; df = 1

As shown in table 5.21, $H0_{SVC}$ must be rejected for sets A, B and C. In other words, the combined and the simple model do significantly differ when employed for identifying SVCs. $H0_{figur}$ must as well be rejected for sets B and C, i.e., the models differ with respect to identifying figurative expressions when applied to sets B and C.

Interpretation

When the entropy model is combined with the kwic-model, identification of SVCs from base form data becomes significantly more accurate compared to simply applying the entropy model.

A similar result is achieved for identifying figurative expressions from sets B and C, providing further evidence for the occurrence of certain verbs in SVCs and figurative expressions.

The major advantage of the combined model over the simple entropy model is that the kwic-strategy leads to a strong reduction of the candidate data, resulting in higher identification accuracy. The important difference between applying the kwic-strategy to SVCs and figurative expressions is that recall of figurative expressions considerably declines, whereas recall of SVCs remains fairly constant. Thus the combined model is without doubt for SVCs the better alternative to the simple entropy model, but is restricted with respect to figurative expressions. In the latter case, the trade-off between recall and precisions must be considered carefully.

Experiment Vc

In this experiment, the entropy+kwic-model is compared with the frequency+kwic-model for identifying SVCs from P.N.V(base form)-triples.

The research hypothesis is:

H_1: The entropy+kwic-model and the frequency+kwic-model differ significantly with respect to identifying SVCs from P.N.V(base form)-triples.

H_0: The entropy+kwic-model and the frequency+kwic-model do not differ with respect to identifying SVCs.

According to the results presented in table 5.22, H_0 must be rejected for sets A and C. In other words, the entropy+kwic- and the freq+kwic-model differ significantly for identifying SVCs from sets A and C of base form data, but both models are equally well suited for identifying SVCs from set B of base form data.

P.N.V(base form)-triples					
	entropy + kwic		frequency + kwic		χ^2
set	SVC	total	SVC	total	
A	**86**	145	88	220	12.3
B	132	347	134	391	0.98 n.s.
C	164	850	**134**	391	32.11

Table 5.22: Comparison of the entropy+kwic- and the frequency+kwic-model; figures in bold face indicate the superior model; df = 1

Interpretation

The entropy+kwic-model performs significantly better for set A, i.e., for data with high occurrence frequency, whereas the frequency+kwic-model is significantly better for set C which contains a large portion of low frequency data. Here again, we find the widely experienced superiority of the frequency-based approach over statistics-based approaches with respect to data containing a large proportion of low frequency occurrences. It is also noteworthy that the entropy+kwic-model is in all cases significantly better than the I+kwic-model.[4]

5.6 Summary

In the following, a summary of the experiments conducted hitherto is presented, and answers to the questions asked in the introduction are given.

---

[4] No table is given.

Differences between the models (*MI*, *Dice*, *Lgl*, *I* and *freq*) have been found concerning recall and accuracy (precision) of collocation identification. The models differ in their suitability for collocation identification depending on the sample employed and on the type of collocation to be identified, i.e., SVCs, figurative expressions or pseudo-collocations.

Sample characteristics having an impact on collocation identification are the threshold of cooccurrence frequency $c \geq 3, 5, 10$ which corresponds to sets C, B and A; and the (morpho)syntactic constraints applied during candidate selection, i.e., full form versus base form data, P.N.VVPP-triples etc.

There are more significant differences between the models concerning identification of SVCs, especially identification from medium frequency data. For figurative expressions, a "best model" is more difficult to define. *Freq* is a good identifier for collocations$_{all}$, which is particularly due to the frequency-based definition of pseudo-collocations. *Freq* is also well suited when samples containing large portions of low frequency data are used, and with some restrictions in the case of medium frequency data.

Given base form data, especially sets B and C, the dominance of *freq* is more obvious than given full form data. This is due to the fact that under full form data the two or more best models approximate, because the performance of *freq* does not increase singnificantly from base to full form data, whereas the performance of *I* and *Lgl* in identifying SVCs and figurative expressions drastically increases from base to full form data.

MI and *Dice* are the best association models for identifying SVCs from highly recurrent full form data (sets A), and for identifying figurative expressions from sets A, B and C base forms.

I and *Lgl*, on the other hand, are equally well suited for identifying SVCs from data containing large portions of medium (sets B) and low (sets C) frequency PNV-tuples. While *MI* and *Dice* are better suited for identifying figurative expressions from base form data, *I* and *Lgl* are more appropriate for identifying figurative expressions from full form data.

The particular strength of the kwic-based approach lies in its ability to improve the identification accuracy for SVCs when combined with a frequency-based or an entropy-based candidate selection.

PP-entropy is a clear alternative to the association measures for identifying SVCs and figurative expressions from high and medium frequency full form data, but also for identifying SVCs from high frequency base form data, and for identifying figurative expressions from medium frequency base form data.

All in all, there is no single best measure for identifying different types of collocations from different samples. In general, statistical measures tend to overestimate low frequency data. The effect is less strong with measures that take the significance of the data into account, which applies to two of the measures used in this study, namely the log-likelihood statistics (*Lgl*) and relative entropy (*I*). Thus the measures become superior to *MI* and *Dice* with increasing number of low frequency word combinations among the data. However, frequency-based cutting-off of the data in test samples containing large portions of low frequency data (sets C) leads to better results in collocation identification than applying the statistical association measures to the full samples.

In the following, the results are presented in more detail.

The results for identifying SVCs and figurative expressions by means of lexical association measures are:

Identification from high frequency data (sets A): All lexical association measures tested, i.e., mutual information according to Church and Hanks 1989 *MI*, the Dice coefficient *Dice*, relative entropy *I*, the log-likelihood statistics introduced by Dunning 1993 *Lgl* and simple co-occurrence frequency *freq* are equally well suited for identifying figurative expressions from full as well as from base form data, and for identifying SVCs from base form data. The picture is different for identifying SVCs from full form data. In this case *MI* and *Dice* are significantly better than the other models. All in all, *MI* and *Dice* have shown to be the best measures for identifying SVCs and figurative expressions from high frequency data.

Identification from samples containing large portions of medium frequency data (sets B): There are clearly two superior models for identifying SVCs from full form as well as from base form data, these are *I* and *Lgl*. The dominance of *I* and *Lgl*, however, is not valid for figurative expressions. The situation is also different for figurative expressions, here *freq* is among the best measures for identifying figurative expressions from full and base form data, but there are also two statistical measures among the best models: *MI* for base form data, and *I* for full form data.

Identification from samples containing large portions of low frequency data (sets C): In most cases, *freq* outperforms the statistical association models in samples C, because of two reasons: On the one hand, statistical measures tend to overestimate low frequency data, and thus identify word combinations as collocations because of their low frequency occurrence in the sample under investigation. On the other hand, the frequency-based approach cuts off low frequency data, thus only highly recurrent collocation candidates are left which is the sample with the highest proportion of collocations, as we know from section 4.3.2. In particular, *freq* is significantly better than *I* or *Lgl* for identifying SVCs from base form data, and from full form data with sample sizes $n = 500$, 1 000. *Freq*, *I* and *Lgl* are equally well suited for full form data with sample sizes $n = 1\ 500$, 2 000. *Freq* is also the best model for identifying figurative expressions from base form data consistently outperforming *Dice* which is the second best model. In the case of full form data, *freq* is always among the best models, but I and Lgl compete with $freq$.

Advances of the kwic-model:

A kwic-based reduction of the test samples leads to a strong increase of the proportion of SVCs and also to a weaker increase of figurative expressions. Kwic is a good strategy to improve precision, provided the sample selected by means of kwic does not become too large. In the latter case, recall (the number of collocations identified) improves but precision declines.

In the following cases, the kwic-based approach is superior to the lexical association models in identifying SVCs: for set B and C of P.N.VVPP-trigrams, and for set A of base form data. With respect to set A P.N.VVPP-trigrams, the kwic-strategy and the association models perform equally well. $Freq$ $n = 500$ and the kwic-model perform equally well for set B base form data, whereas $freq$ $n = 500$ outperforms the kwic-model with respect to set C of base form data.

The kwic-strategy, however, is not superior to the best association models when employed for identifying figurative expressions from both P.N.VVPP-trigrams and base form data. Nevertheless, the kwic-strategy cannot be completely rejected for figurative expressions as it is among the best models for identifying figurative expressions from sets A and B of P.N.VVPP-trigrams and base forms. Even though recall and precision tend to be worse than in the case of the best association models. In general, the results provide evidence that there is a subset of verbs which are typical for SVCs as well as for figurative expression.

A combination of $freq$ and the kwic-strategy (freq+kwic) performs in all cases significantly better than simply employing $freq$ for identifying SVCs from base form data. The freq+kwic-model also outperforms the model combining I and *kwic* in identifying SVCs from base form data. The advantage of the freq+kwic-model is that recall is similar to $freq$, whereas precision is significantly higher.

Summing up, the kwic-strategy on its own has its clear limitations for identifying SVCs, even though it has been designed for the particular task. A combination with a simple frequency-based approach, however, allows identification accuracy (= precision) of SVCs to improve. This, however, does not hold for recall.

Results employing the entropy model:

Low entropy values of the potential PP-collocates are good indicators for collocativity. With respect to precision, the model is equally well suited for identifying SVCs and figurative expressions from full form data. The entropy model is superior to the best association models in samples with high (sets A) and medium (sets B) occurrence frequency, and it is among the best models for set C. With respect to recall, the entropy model clearly outperforms the respective best association model when identifying SVCs from sets B and C of full form data.

In the case of base form data, the entropy model outperforms, with respect to precision, the best association model in identifying SVCs from highly recurrent data (set A), and in identifying figurative expressions from data containing larger portions of medium frequency data (set B). In the case of set C, however, the precision of entropy is outperformed by *freq* in identifying SVCs and figurative expressions. With respect to recall from base form data, entropy is only slightly better than the respective best association models for identifying SVCs from sets B and C, and substantially better for identifying figurative expressions from set C.

Summing up, with respect to precision entropy is preferable over the lexical association measures for both identifying SVCs and figurative expressions from full form data, and it is also preferable over the association measures in identifying SVCs and figurative expressions from samples containing high (set A) and medium (set B) frequency base form data. With respect to recall, the entropy model is clearly superior to the association measures for identifying SVCs from set C of full form data and figurative expressions from set C of base form data.

Combining the entropy model with the kwic-strategy leads to an improvement of identification accuracy of SVCs from sets A, B and C of base form data compared to simply applying the entropy model. This is also the case for identifying figurative expressions from sets B and C of base form data.

5.7 Control Experiments

In order to evaluate the generality of the results gained by experimenting with the Frankfurter Rundschau corpus, a number of key experiments have been repeated on the basis of a collection of German newsgroup contributions. The corpus has been selected, because newsgroup discussions are a completely different type of text than newspaper articles. While newspapers are typical instances of text with controlled style and orthography, newsgroup contributions are much more spontaneous productions of language, which influences style, wording and orthography. It is for this reasons that such texts are employed for corpus-based approaches to language and grammar checking, see for instance the FLAG project at DFKI, Saarbrücken.[5] As a new extraction corpus, a 10 million word sample has been selected from the corpus of newsgroup messages set up in the FLAG project. The corpus has been made available part-of-speech tagged and syntactically analyzed employing the tools described in section 2.2.1.[6]

---

[5] `http://www.dfki.de/pas/f2w.cgi?ltp/flag-e`

[6] The corpus has been jointly developed at the University of Tübingen and at the DFKI, Saarbrücken.

5.7.1 A Corpus of Newsgroup Contributions

Even though the two extraction corpora differ in size, i.e., 8 million words in the newspaper corpus and 10 million words in the newsgroup corpus, the number of preposition-noun-main-verb combinations extracted is quite similar: 370 013 PNV-triples from the newsgroup corpus, and 372 212 triples from the newspaper corpus. The corpora, however, differ with respect to the distribution of word co-occurrences. The number of P.N.V(full form)-triples that occur only once is slightly smaller in the newsgroup corpus (80 %) than in the newspaper corpus (87 %), the number of recurrent combinations is accordingly higher. There are 14 % word combinations where $c = 2$ in the newsgroup corpus versus 10 % in the newspaper corpus, and 6 % word combinations where $c \geq 3$ in the newsgroup corpus versus 3 % in the newspaper corpus. Accordingly sets A, B, and C derived from the newsgroup corpus are larger than the respective sets taken from the newspaper corpus, see table 5.23.

set	newsgroup corpus	newspaper corpus
A (≥ 10)	1 108	747
B (≥ 5)	5 159	2 864
C (≥ 3)	22 813	10 430
approx. corpus size	10^7	$8 * 10^6$

Table 5.23: Comparison of the frequency distributions in the newsgroup and the newspaper corpus

As word combinations with high co-occurrence frequency are more likely to be collocational than word combinations with low co-occurrence frequency, only data with occurrence frequency $c \geq 10$ will be used in the control experiments. In the following, set A of P.N.V(full form)-triples from the newsgroup corpus is examined with respect to the occurrence of SVCs and figurative expressions. The results are compared to the respective results from the newspaper corpus. The number of SVCs and figurative expressions is larger in set A of the newsgroup corpus than in the according set of the newspaper corpus. The percentage of SVCs, however, is higher in the newspaper corpus. Considering P.N.V(base form)-triples, 1 614 PNV-combinations have been identified where $c \geq 10$ from the newsgroup corpus compared to 1 249 triples from the newspaper corpus. The sets A of the newsgroup corpus already show the typical inversion of the number of SVCs and figurative expressions in the sets of P.N.V(base form)- and -(full form)-triples, whereas this is not yet found in sets A of the newspaper corpus, even though the phenomenon is valid for sets B and C of the newspaper corpus. A summary of the distributions of SVCs and figurative expressions in sets A of the newsgroup and the newspaper corpus is presented in table 5.24.

newsgroup corpus $c \geq 10$		
	P.N.V(base forms)	P.N.V(full forms)
SVC	182 (11.3 %)	190 (17 %)
figur	231 (14.3 %)	149 (13.4 %)
total	1 614	1 108

newspaper corpus $c \geq 10$		
	P.N.V(base forms)	P.N.V(full forms)
SVC	174 (13.9 %)	144 (19.3 %)
figur	150 (12.0 %)	96 (13 %)
total	1 249	747

Table 5.24: Comparison of the occurrences of SVCs and figurative expressions in set A of the newsgroup and newspaper corpus

5.7.2 Comparison of the Newspaper and the Newsgroup Corpus

Collocations in Common

Comparing the PNV-combinations (verbal base forms) with occurrence frequency $c \geq 10$ from the newsgroup corpus and the according combinations with $c \geq 3$ from the newspaper corpus, 146 SVCs and 138 figurative expressions have been found which occur in both samples. In other words, approximately 80 % of the highly frequent SVCs and 60 % of the figurative expressions in the newsgroup corpus also occur in the subset of the newspaper corpus used for collocation identification. If sets A or B of the newspaper corpus are used as bases for comparison, 192 or 254 SVCs and figurative expressions respectively are common to the newsgroup and the newspaper corpus. Comparison of lexical material from different kinds of corpora allows general language collocations to be identified, as well as provide insights into corpus-specific usage of collocations. For illustration, some examples of common SVCs and figurative expressions, i.e., combinations that occur in the newspaper and the newsgroup corpus, are listed in the following, and characteristic differences between the corpora are described.

Examples for Common Support-Verb Constructions
in (den) Griff bekommen (‘get the hang of something’)
in (den) Griff kriegen (‘get the hang of something’)

in Kontakt kommen ('get in contact')
in Mitleidenschaft ziehen ('to inflict damage upon')
unter Kontrolle bringen ('bring under control')
unter Kontrolle halten ('keep something under control')
unter (Det) Schutz stehen ('be under someone's wing')
unter (Det) Schutz stellen ('take someone under one's wing')
zu (Adj) Ergebnissen kommen ('to achieve (Adj) results')
zur Erkenntnis kommen ('to come to the realization that')
zu Fall bringen ('bring about somebody's downfall')
zu Lasten (von jemanden) gehen ('be someone's expense')
zu Rate ziehen ('consult')
zur Verfügung stehen ('be available')
zur Verfügung stellen ('make available')
außer Kraft setzen ('make invalid')
in Kraft treten ('come into force')

Examples for Common Figurative Expressions
in (den) Sternen stehen ('be in the lap of the gods')
an (der) Spitze stehen ('be the head of')
auf (freien) Fuß setzen ('to release from jail')
auf (dem) Programm stehen ('be in the programme')
auf Eis legen ('put on ice')
auf (eine Adj) Grundlage stellen ('put on a (Adj) foundation')
auf (der) Hand liegen ('be obvious')
auf ... Konto gehen ('someone is to blame for')
auf (den) Kopf stellen ('turn things upside down')
auf Nummer (Sicher) gehen ('play it safe')
auf (die) Palme bringen ('to rile someone')
auf (den) Plan rufen ('bring on to the scene')
auf (die) Reihe kriegen ('to get something done')
auf (die) Sprünge helfen ('give someone a leg up')
auf (Det) Standpunkt stellen ('take the view that')

Differences between the Corpora

The following differences are apparent:

1. There is more lexical variation in the newspaper corpus than in the newsgroup corpus. Thus co-occurrence frequency in average is higher in the newsgroup corpus.

2. Compared to the newspaper corpus there is less variation in the group of pseudo-collocations extracted from the newsgroup corpus. A large number of frequently occurring PNV-combinations in the newsgroup corpus, for instance, relate to recipes.

3. The newsgroup corpus contains various colloquial phrases in the word combinations with occurrence frequency $c \geq 10$. Such word combinations could not be found among the PNV-data occurring at least three times in the newspaper corpus. This can be explained by the stylistic difference of the corpora. The newspaper corpus is controlled and stylistically elaborate, whereas the newsgroup corpus is closer to colloquial speech.

Examples for Pseudo-Collocations related to Recipes
in Scheiben schneiden (‘to slice’)
mit Pfeffer würzen (‘season with pepper’)
in (einen) Topf geben (‘put in a pot’)
mit Zitronensaft beträufeln (‘sprinkle with lemon juice’)
zu Teig verarbeiten (‘make it into a dough’)

Note the difference between *in* $\{$*einen, den*$\}$ *Topf geben* and *in einen Topf werfen* (‘lump together’), where the latter is a general language collocation meaning ‘treat two things/persons the same’, and the former is collocational only because of its high frequency in the newsgroup corpus, particularly in recipes.

Examples for Colloquial Figurative Expressions
auf (den) Keks gehen (‘get on one’s wick’),
auf (den) Geist gehen (‘get on someone’s nerves’),
in (die) Hose gehen (‘be a flop’),
in (die) Pfanne hauen (‘to land someone in trouble’),
aus (den) Fingern saugen (‘to make something up’),
um (die) Ohren hauen (‘to throw something back on somebody’),
über (den) Haufen fahren (‘knock someone down’)

5.7.3 Testing of the Models

Based on the results from experimenting with the newspaper corpus the following tasks employing the newsgroup corpus are pursued:

1. Identification of the best statistical association measure for retrieving, on the one hand, SVCs and, on the other hand, figurative expressions.

2. Comparison of the accuracies gained by the best association measures and the ones gained by applying mere co-occurrence frequency.
3. Evaluation of the identification accuracies gained by performing a kwic-based reduction of the collocation candidates.
4. Comparison of the accuracies gained by the best association measures and the entropy model.
5. Comparison of the results achieved employing combined models.

Statistical Association Measures versus Frequency

Set A of P.N.V(full form)-triples is selected from the newsgroup corpus, and the four statistical association measures as well as mere co-occurrence frequency are applied. In order to make the results gained from the two extraction corpora comparable, a similar percentage (approximately 67 %) of highest ranked word combinations is retrieved. Thus instead of retrieving the 500 highest ranking PNV-combinations, as it has been the case for the newspaper corpus, the 742 highest ranking combinations are selected from set A of the newsgroup corpus. The raw data are presented in table 5.25.

newsgroup corpus $c \geq 10$		
P.N.V(full form)-triples		
measure	SVC	figur
MI	124	109
Dice	126	100
I	171	106
Lgl	171	107
freq	152	106

Table 5.25: Results from applying the association measures to set A of the newsgroup corpus

Control Experiment II'a

First of all, it is tested whether differences between the models exist. The research hypotheses to be pursued are thus:

for SVCs:

$H1_{SVC}$**:** The models differ in their feasibility to identify SVCs.

$H0_{SVC}$**:** There are no differences between the models with respect to SVCs.

for figurative expressions:

$H1_{figur}$: The models differ in their feasibility to identify figurative expressions.

$H0_{figur}$: There are no differences between the models with respect to figurative expressions.

As can be seen from the χ^2 values in table 5.26, $H0_{SVC}$ must be rejected, whereas $H0_{figur}$ cannot be rejected.

newsgroup corpus $c \geq 10$				
P.N.V(full form)-triples				
	SVC		figur	
	excl. freq	incl. freq	excl. freq	incl. freq
χ^2	17.88	17.91	0.5	0.5
signif. level	.001	.01	n.s.	n.s.

newspaper corpus $c \geq 10$				
P.N.V(full form)-triples				
	SVC		figur	
	excl. freq	incl. freq.	excl. freq	incl. freq.
χ^2	151.57	150.08	1.34	2.05
signif. level	.001	.001	n.s.	n.s.

Table 5.26: Comparison of the association measures including and excluding frequency; n.s. = not significant; df = 3 (excluding $freq$); df = 4 (including $freq$)

Interpretation

The models differ significantly for the newsgroup corpus when employed for identifying SVCs, but perform equally well for figurative expressions. A similar result has already been found in set A of the newspaper corpus, cf. table 5.26. These results provide evidence for the generality of (i) the dichotomy of the models with respect to identifying SVCs from high frequency data, and (ii) the similarity of the models with respect to identifying figurative expressions from high frequency data.

Control Experiment II'b

It is now examined whether a single best model exists for identifying SVCs from set A of full form data taken from the newsgroup corpus. If the result were the same as in the newspaper corpus, MI and $Dice$ should turn out as best models. This, however, is to be doubted as the frequency distributions differ between the two corpora.

The following research hypothesis is employed:

H_1: There are differences between the models identifying the first and second highest number of SVCs.

H_0: There is no difference between the two best models.

As we see from table 5.25, p. 163, the models with the best recall of SVCs from set A of the newsgroup corpus are: *I/Lgl* (171 SVCs) and *Dice* (126 SVCs) when only the statistical models are considered, or *I/Lgl* (171 SVCs) and *freq* (152 SVCs) otherwise. The according χ^2 values are:

$$\begin{array}{llll} I/Lgl \text{ versus } Dice: & \chi^2 = 8.15 & \alpha = .01 & \text{df} = 1 \\ I/Lgl \text{ versus } freq: & \chi^2 = 1.28 & \text{not significant} & \text{df} = 1 \end{array}$$

Thus H_0 cannot be rejected for *I/Lgl* versus *freq*, but must be rejected for *I/Lgl* versus *Dice*, i.e., there are no significant differences between *I*, *Lgl* and *freq*, but clear differences between *I/Lgl* on the one hand and *Dice* on the other hand.

Interpretation

I and *Lgl* are the best association measures for identifying SVCs from set A of the newsgroup corpus, and a frequency-based approach is equally well suited.

The results differ from those gained from the newspaper corpus, where both *MI* and *Dice* are the highest ranking methods for identifying SVCs from set A. Thus the results from set A of the newsgroup corpus are closer related to the results for set B and C of the newspaper corpus, where *I* and *Lgl* on the whole have proven to be best suited for identification of SVCs, cf. table 5.9. This result is in accordance with the altered frequency distributions in the newsgroup corpus where the set of high frequency data is much larger than in the newspaper corpus,[7] the density of support-verb constructions among the data, however, is clearly smaller.[8] Thus collocation identification from set A of the newsgroup corpus is much more comparable to collocation identification from set B of the newspaper corpus.

---

[7] 1 108 PNV-combinations in the newsgroup versus 747 combinations in the newspaper corpus.

[8] 17 % SVCs in the newsgroup corpus versus 19.3 % in the newspaper corpus.

Kwic-Based Data Reduction

Similar to the newspaper corpus, kwic-based reduction of the collocation candidates results in an increase of the percentage of SVCs and, to a smaller extent, in an increase of the percentage of figurative expressions. See table 5.27 for illustration. Again, the increase of SVCs is overproportional, as support-verbs are used as lexical keys. Thus it is expected that models combined with the kwic-strategy differ significantly for identifying SVCs as well as figurative expressions than the simple models without kwic-based data reduction.

P.N.V(base form)-triples, set A				
	newsgroup corpus		newspaper corpus	
	total	kwic	total	kwic
SVC	182 (11.3 %)	150 (24.2 %)	174 (13.9 %)	147 (32.1 %)
figur	231 (14.3 %)	116 (18.7 %)	150 (12.0 %)	86 (18.8 %)
sample size	1 614	619	1 249	458

Table 5.27: Kwic-based data reduction

Control Experiment IV'a

The hypotheses to be tested are:

for SVCs:

$H1_{SVC}$: Given two models, one being the kwic-based extension of the other one: There are differences between the models in identifying SVCs .

$H0_{SVC}$: Given two models, one being the kwic-based extension of the other one: There is no significant difference between the models in identifying SVCs.

for figurative expressions:

$H1_{figur}$: Given two models, one being the kwic-based extension of the other one: There are differences between the models in identifying figurative expressions.

$H0_{figur}$: Given two models, one being the kwic-based extension of the other one: There is no significant difference between the models in identifying figurative expressions.

The entropy+kwic and the freq+kwic model have shown to significantly improve identification of SVCs from sets A, B, and C of base form data taken from the newspaper corpus. No significant difference between entropy and the entropy+kwic model could be detected for identifying figurative expressions from set A base forms of the newspaper corpus. For the newsgroup corpus, see table 5.28 for the raw data. The percent values represent precision. The total number of PNV-combinations examined by the frequency model is 646. This covers approximately 67 % of the total number of PNV-combinations in set A of P.N.V(base form)-triples of the newsgroup corpus. This strategy is used in order to make the results achieved by the newsgroup corpus comparable to the results from the newspaper corpus.

newsgroup corpus $c \geq 10$			
P.N.V(base form)-triples			
model	SVC	figur	total
frequency	111 (17.2 %)	115 (17.8 %)	646 (100 %)
frequency+kwic	93 (40.8 %)	60 (26.3 %)	228 (100 %)
entropy	63 (18.5 %)	77 (22.6 %)	341 (100 %)
entropy+kwic	53 (39.8 %)	31 (23.3 %)	133 (100 %)

Table 5.28: Comparison of combined and simple models; raw data

model 1	model 2	collocation	χ^2-value	significance level
frequency+kwic	frequency	SVC	51.2	.001
frequency+kwic	frequency	figur	7.1	.01
entropy+kwic	entropy	SVC	22.5	.001
entropy+kwic	entropy	figur	0.0022	n.s.
frequency+kwic	entropy+kwic	SVC	0.0041	n.s.
frequency+kwic	entropy+kwic	figur	0.259	n.s.

Table 5.29: Comparison of the models; in cases where the differences are significant, model 1 is the superior one; n.s. = not significant; df = 1

Considering the significance values in table 5.29, $H0_{SVC}$ must be rejected for both the freq+kwic- and the entropy+kwic-model. In other words, the freq+kwic- and the entropy+kwic-model differ significantly from simply applying the frequency and the entropy model, respectively. In addition, $H0_{figur}$ must also be rejected for the freq+kwic-model, which means that there is a significant difference in employing the combined model versus the simple frequency model. The difference, however is not significant for entropy+kwic and entropy when employed for identifying figurative expressions.

Interpretation

Similar to the results from the newspaper corpus, the results from the newsgroup corpus show that the combined models (freq+kwic and entropy+kwic) are significantly better than the simple models for identifying SVCs. Similarly, as obtained from the newspaper corpus, there is no difference between the entropy+kwic- and the entropy model for identifying figurative expressions in the newsgroup corpus. The finding in the newsgroup corpus that freq+kwic outperforms $freq$ for identifying figurative expressions is in accordance with the overall superiority of freq+kwic over $freq$ in identifying SVCs from the newspaper corpus. Other than in the newspaper corpus, where entropy+kwic has shown to be superior to freq+kwic for identifying SVCs from set A base form data, there is no significant difference ($\chi^2 = 0.0041$) between freq+kwic and entropy+kwic for identifying SVCs from set A base form data taken from the newsgroup corpus.

The Entropy Model

The values of PP-entropy are determined for the PNV-combinations taken from set A of the P.N.V(full form)-triples extracted from the newsgroup corpus. PNV-combinations where PP-entropy > 0.7 are eliminated from the set of collocation candidates. The candidate set reduces to 301 word combinations of which 85 have been manually identified as SVCs and 57 as figurative expressions.

Control Experiment V'a

The following hypotheses are tested:

for SVCs:

$H1_{SVC}$: The entropy model and the best association model differ with respect to the identification of SVCs from sets A of the newsgroup corpus.

$H0_{SVC}$: The entropy model and the best association model do not differ with respect to the identification of SVCs from sets A of the newsgroup corpus.

for figurative expressions:

$H1_{figur}$: The entropy model and the best association model differ with respect to the identification of figurative expressions from sets A of the newsgroup corpus.

$H0_{figur}$: The entropy model and the best association model do not differ with respect to the identification of figurative expressions from sets A of the newsgroup corpus.

The data of research are presented in table 5.30. MI(all) indicates that there is no significant difference between the models identifying figurative expressions, neither between the association measures, nor between the association measures and the frequency-based strategy (cf. tables 5.11 and 5.12). The highest number of figurative expressions, however, is identified by MI. Thus this number is used for comparison with the entropy model. Similarly, $I/Lgl(freq)$ indicates that there is no significant difference between the three models for identifying SVCs; again the highest number of SVCs identified is used for comparison with the entropy model.

As can be seen from table 5.30, $H0_{SVC}$ and $H0_{figur}$ cannot be rejected, i.e., there are no significant differences between the best association models and the entropy model considering precision. Recall is higher for the association models and $freq$.

P.N.V(full form)-triples, set A			
newsgroup corpus	SVC	figur	total
entropy	85	57	301
best assoc. meas.	171 I/Lgl (freq)	109 MI (all)	742
χ^2	2.84	2.58	
signif. level	n.s.	n.s.	
newspaper corpus	SVC	figur	total
entropy	**100**	**49**	212
best assoc. meas.	134 MI	80 MI	500
χ^2	27.08	4.61	
signif. level	.001	.05	

Table 5.30: Comparison of PP-entropy and best association model; set A, full form, newsgroup and newspaper corpus; n.s. = not significant; figures in bold face indicate the superior model; df = 1

Interpretation

I, Lgl, $freq$ and PP-entropy are equally well suited for identifying SVCs from set A of newsgroup corpus. Considering also the result from control experiment I'a, it can be concluded that for the sample at hand these models are significantly better than MI and $Dice$. In the case of figurative expressions, on the opposite, all models I, Lgl, $freq$, MI, $Dice$ and PP-entropy are equally well suited. The results hold with respect to accuracy.

These results differ from the ones gained from set A of the newspaper corpus, where entropy has proven to be significantly better than the best association model for identifying SVCs (MI) as well as figurative expressions (MI). The results from the newsgroup corpus thus are only comparable to the results from set C of the newspaper corpus which is the only case where the differences between entropy and the best association model are not significant.

5.8 Conclusion

The results achieved from the newsgroup corpus confirm, to a large extent, the results gained from examining the newspaper corpus, even though the two corpora differ at various levels. This speaks for the generalizability of the results. The differences between the results can in the first place be attributed to the differences in the frequency distributions between the corpora which is a reflex of the differences in text type. There is less lexical variation in the newsgroup corpus than in the newspaper corpus. Thus collocation identification becomes harder even from highly recurrent word combinations where $c \geq 10$. As a consequence, methods that have been appropriate for medium occurrence frequencies with $c \geq 5$ (set B) in the newspaper corpus are now well suited for collocation identification from high frequency data (set A) extracted from the newsgroup corpus. In the following, the partial results are listed.

The inversion of the number of SVCs and figurative expressions between full and base form data is confirmed by the newsgroup data, providing further evidence that there is in average more variation in verb inflection in figurative expressions than in SVCs.

The results also show that the distribution of collocations differs between corpora. As expected, a broad selection of newspaper text contains more lexical variation than a corpus consisting of contributions to newsgroups. In other words, recurrence is in general larger in the newsgroup than in the newspaper corpus. As a consequence, the differences between the frequency distributions of collocations and noncollocations decline. Thus different statistical measures are appropriate for collocation identification from the two corpora. This is confirmed by the results from applying the statistical association measures to full form data, i.e., association models which have proven to be best for set B of the newspaper corpus are now best for set A of the newsgroup corpus, namely I, Lgl and $freq$.

Similar results are achieved – by means of entropy compared to association measures – for identifying SVCs and figurative expressions from set A of full form data from the newsgroup corpus and set C of full form data from the newspaper corpus.

Summing up, the experimental results provide evidence for a relation between corpus type, the frequency distribution of word combinations in general, and the distribution of collocation classes in particular. Frequency distributions of lexical co-occurrences in a corpus vary depending on factors such as text types and domains represented by the corpus, as well as corpus size. In the work presented, a number of hard and soft criteria influencing the quality of collocation identification could be identified. By soft criteria, we mean restrictions that must be empirically determined on a case by case basis, such as thresholds determining the minimal co-occurrence frequency required for a word combination to be a potential collocation candidate or thresholds determining the entropy value based on which the PPs are divided into collocates and noncollocational phrases in the entropy model.

Hard criteria for corpus-based collocation extraction identified in the present study are summarized below:

A kwic-based selection of collocation candidates increases the accuracy of collocation identification, especially in combination with the entropy model on the one hand, and with the frequency-based approach on the other hand. This effect has been found in the newspaper as well as the newsgroup corpus.

A purely frequency- or statistics-based approach to collocation identification is still improvable because:

- In each corpus, a substantial number of word combinations exists for which no frequency-based distinction between collocations and noncollocations can be made.

- Low frequency collocations cannot be reliably distinguished from other low frequency word combinations by means of statistics.

Chapter 6

A Framework for the Representation of Collocations

6.1 Introduction

A framework is developed for a uniform representation of collocations ranging from grammatically fixed to highly flexible collocation classes. The individual representations contain three classes of information.

1. The lexic of the collocates.
2. The competence base: an underspecified linguistic description of the collocation accounting for morphosyntactic and syntactic properties of the collocates and the collocation phrases.
3. The example base: an extendible selection of actual realizations of collocations identified from corpora.

An integrated representation of linguistic descriptions of collocations and real-world examples is required, as a purely competence-grammatical description of collocations either over- or undergenerates. Corpora provide information on the usage of collocations such as information on the preferred lexical realization of the collocates, on prevalent modification, on actually occurring syntactic realizations, etc. But a purely corpus-driven approach to collocations is insufficient, because of data sparsity, i.e., corpora offer only partial information since they contain just samples of common usage of linguistic constructions, and thus it is rather unlikely that all grammatically possible and pragmatically licensed variants occur. Thus the competence part of the description must be conceived as an outline of the grammatical potential, whereas the corpus examples represent the restrictions in usage. In section 6.2 an outline of the competence part of the representation is given. The example base is described in section 6.3, and the implementation as a relational database is presented in section 6.4. Example queries are given in section 6.5. Facilities for further exploitation of the database output, and for semi-automatic construction of the database entries are sketched in section 6.6.2.

6.2 Competence-Based Representation

6.2.1 Lexical Representation of the Collocates

Collocates are either morphologically fixed or flexible. Fully flexible collocates are represented by their base form, partially flexible collocates are represented by regular expressions. Inflexible collocates are represented by their full forms. Each collocate is associated with part-of-speech information which links the collocates to the (noncollocational) lexicon, and to the standard rules of grammar. See table 6.1 for illustration. The ambiguous pattern "zur?"[1] covers the following two realizations of a preposition : *zu* and *zur*, i.e., without determiner or with determiner included *zur* (APPRART) = *zu* (APPR) + *der* (ART).[2] The form of the noun is fixed, see "Verfügung". As the verbs in PP-verb collocations are usually morphologically flexible, they are represented by their base forms (here: bare infinitive prefixed with ":").

Form	zur?	Verfügung	:stellen
PoS	APPR(ART)?	NN	VV

Table 6.1: Description of the SVC *zur Verfügung stellen* at lexical level

6.2.2 Structural Properties of Collocations

The collocations examined have two components, the PP-collocate and the syntactic structures that are constituted by the collocation and its arguments.

Collocations and Argument Structures

The verb (V_{col}) in a PP-verb collocation syntactically functions as the head of a verbal construction. The PP-collocate (PP_{col}) resembles an obligatory argument which is at least partially determined with respect to its lexical realization. The other arguments (Args) required by the collocation are lexically underspecified. Lexical determination of one PP-argument is a particular property of the

---

[1] More information on the Perl-like notation used here can be found in [Wall *et al.*, 1996].

[2] See [Thielen and Schiller, 1995] for the tagset. NN stands for noun, VV for main verb.

collocations examined in the present study. There are also collocations with verbal syntactic heads exhibiting more than one lexically prespecified argument, e.g., the proverb *die Spreu vom Weizen trennen* ('separate the wheat from the chaff') where the subject is the only lexically unspecified argument. Another example is the proverb *Morgenstund hat Gold im Mund* (morning hour has gold in the mouth, 'the early bird catches the worm') where all arguments are lexically determined. Some examples of PP-verb collocations and related argument structures are provided in table 6.2.

Collocation	V_{col}	PP_{col}	Args
in Betrieb nehmen	nehmen	in Betrieb	NPnom, NPacc
in Betrieb gehen	gehen	in Betrieb	NPnom
vor Augen halten	halten	vor Augen	NPnom, NPdat, {NPacc,S_daß}
aus Hand geben	geben	aus Hand	NPnom, NPacc

Table 6.2: Syntactic structure of PP-verb collocations

Such data are a valuable source of information for further investigations of the argument structure of collocations. In the case of SVCs, for instance, a not yet fully understood relation exists between the argument structure required by the collocation and the argument structures required by the support-verb and the predicative noun in their noncollocational occurrence, information which can be derived from standard lexica. In addition, the information is indispensable for constructing generation and analysis lexica from the representations.

6.2.3 Representation of PP-Collocates

PP-collocates are described with respect to linear precedence, and fixed determination and modification. See for instance *aus den Augen verlieren* ('lose sight of'). In this case, information on the determiner needs to be specified at the collocation entry. However, there are collocates where a mere competence-based description is problematic like in *zur/zu Verfügung*. Even though *zur* suggests the occurrence of the article *der*, *zu der Verfügung* is odd as a predicative phrase. On the other hand, the use of a possessive pronoun – *zu seiner Verfügung* (at his disposal) – is acceptable, but rare. Thus corpus data shed light on the actual usage.

With respect to syntactic structure, the majority of PPs discussed in the present study only consist of a preposition and a noun. Depending on the degree of lexicalization, the PP-collocate may be modifiable. According to competence grammar, possible prenominal modifiers in German NPs or PPs are genitive NPs (NP_{gen}) and adjective phrases (ADJP), postnominal modifiers are NP_{gen}, PP and adverbial phrases (ADVP). This general modification pattern is clearly restricted in most collocations. A number of the PP-collocates permit modification with attributive adjectives leading to structures such as APPR-ADJA-NN, *mit offenen Augen (sehen)* (see something with open eyes, 'be fully aware of something'), or APPR-ART-ADJA-NN *in den allerersten Anfängen (stecken)* (in the very first beginnings stick, 'be in the very beginning'). Occurrences of other kinds of modifiers are rare. Examples can be found in idiomatic expressions such as *in Teufels Küche kommen, Nägel mit Köpfen machen* where *Teufels* is a prenominal genitive, and *mit Köpfen* is a PP. Even though the constituents are in typical modifier position, they are lexically fixed parts of the collocations, and thus obligatory. Attributive adjectives, on the other hand, usually are optional in collocations. Thus the standard representation of the PP-collocate will be that of an underspecified kernel PP consisting of a preposition and a noun. While obligatory modification is fully specified in the competence part of the representation, optional modification is represented in the realization part. See table 6.3 for competence-based representations of linear precedence and hierarchical structure in the PP-collocate, and table 6.4 for examples with fixed determination, where 'def' stands for definite determination, i.e., *aus* **den** *Augen verlieren* ('to lose sight'), *in* **die** *Hände fallen* ('fall into someone's hands'), 'incorp' indicates fusion of determiner and preposition which is just a reduplication of information already available from the part-of-speech tag APPRART; 'nil' indicates that determination is blocked.

Form	Category	Precedence
zur?	APPRART	0 - 1
Verfügung	NN	1 - 2
zur? Verfügung	PPcol	0 - 2
in	APPR	0 - 1
Teufels	NPgen	1 - 2
Küche	NN	2 - 3
in Teufels Küche	PPcol	0 - 3

Table 6.3: Syntactic structure of the PP-collocate

Summing up, the examples on modification and determination in the PP collocate demonstrate that corpus data and native speaker competence need to be combined for an adequate description. In the competence part of the representation, modification and determination is explicitly blocked, spelled out with respect to dominant variants, and left unspecified in the case of flexible collocation phrases. Especially for the latter, corpus data are important for providing information on the actual usage. Regularities in the corpus data, on the other

collocation	determination
aus Augen verlieren	def
in Hände fallen	def
ins Auge fassen	incorp
zu Felde ziehen	nil

Table 6.4: Determination in the PP-collocate

hand, are candidates for being represented in the competence base.

6.2.4 Collocation-Specific Properties

Support-verb constructions are a good example for collocations with very particular collocation-type-specific properties, cf. section 3.4.3. For the reader's convenience, the main properties are repeated. The main function of SVCs is to express various aspects of a predicate. The core meaning of the predicate is determined by the predicative noun which in many cases is morphologically derived from a verb. Exchanging the support verbs is a means for variation of the thematic structure of the predicate, and to vary Aktionsart. Consider the following example: *in Betrieb nehmen* ('to set into operation') and *betreiben* ('to operate'), the verb underlying *Betrieb*, have causative reading, which means an agent exists who causes something to be in operation. In order to eliminate the causer, the construction can be passivized (sentences 6.1) or the verb in the SVC can be exchanged (sentence 6.2). Both methods lead to similar results: the former object (NP_{acc}) has become subject and the old subject (NP_{nom}, the causer) has been deleted. Examples of the variation of Aktionsart can be found in tables 3.7 on page 76 and 3.8 on page 77.

(6.1) a. die Anlage wurde betrieben
('the plant was operated by ...')

b. die Anlage wurde in Betrieb genommen
('the plant was put into operation')

(6.2) die Anlage ist in Betrieb gegangen
('the system went into operation')

6.3 Collection of Real World Data

In this section, the relevance of real-world data for the description of collocations will be discussed.

6.3.1 Typical Lexical Realizations

Collocations may vary with respect to the morphological realization of individual collocates, as well as with respect to the lexical items used. An already discussed example of the latter case is the variation of support verbs in SVCs. Another example of lexical variation is given in the following: *Beine* (legs) and *Füße* (feet), on the one hand, denominate different body parts, on the other hand the lexical items are regional variants meaning 'legs'. Interestingly, the two words may occur with the same collocates in PP-verb constructions, i.e., *auf* {*Beine, Füße*} *stellen* (at {legs, feet} put). Corpus data, in this case, can provide information on the frequency of a particular realization, and on possible differences in usage or interpretation. In the newspaper corpus, there are 55 instances of *auf Beine stellen*. The PP-collocates in all examples require definite determination – *auf die Beine*. These are opposed by 7 instances of *auf Füße stellen* with highly flexible prenominal modification. See the examples below, where the PP-collocates are printed in bold face. Note, the examples are automatically extracted from the collocation database. As punctuation marks are treated as individual tokens, punctuation marks thus are surrounded by blanks in the examples below. The context-dependent translation of the PP-verb collocations are printed in bold face.

(6.3) " Wir suchen weitere Sponsoren , um uns **auf mehrere Füße** zu stellen " , hofft Leonhardt auf gesteigertes Interesse in der heimischen Wirtschaft
('We are looking for more sponsors, in order to diversify our income, hopes Leonhardt for increasing interest in the local economy')

(6.4) Der Bürgermeister von Glashütten , Helmut Diehl (CDU) , will dagegen die Stromversorgung seiner Gemeinde jetzt **" auf sichere Füße** stellen "
('The mayor of Glashütten, Helmut Diehl (CDU), wants to secure the electricity supply of his community')

(6.5) Die Obdachlosen wieder **" auf eigene Füße** zu stellen " , das scheint auch in Egelsbach das größte Problem zu sein
('The greatest problem in Egelsbach seems to be making the homeless stand on their own two feet again')

(6.6) Doch Karin Oster vom BBJ hofft , irgendwann die Kooperative **auf sichere finanzielle Füße** stellen zu können
('Nevertheless Karin Oster of BBJ hopes one day to be able to give the co-operative a sound financial base')

(6.7) Eine Frau , die zehn oder zwanzig Jahre von den Einkünften ihres Mannes gelebt hat und in dieser Zeit sorgfältig alle Anstrengungen vermied , sich

wirtschaftlich **auf eigene Füße** zu stellen , darf nicht auch noch belohnt werden
('A woman who has been living for ten to twenty years off her husband's income, and during this time has been carefully avoiding any effort to become economically independent should not then be rewarded for it')

(6.8) Und damit sich einmal Autos den Bäumen **auf die Füße** stellen können , werde ja erst einmal abgegraben , aufgeschüttet und verdichtet
('And so that cars can one day tread on trees' feet, first of all they get dug up, then gravel gets strewn and then it is sealed over.')

(6.9) Walter Kempowski hat dieses Klischee vom Kopf **auf die Füße** gestellt
(Walter Kempowski turned the cliché the right way up')

While *auf Füße stellen* has different readings like 'to provide security for' (e.g. 6.3), 'dominate' (6.8), 'turn something right' (6.9), 'make independent' (6.5), *auf Beine stellen* in the realization *auf die Beine stellen* has an uniform interpretation meaning 'to set up or organize something'. A few examples are given below.

(6.10)Entgegen der Absprache mit dem Vereinsring habe Roth parallel zu dem Stadtteilfest seine eigene Fete auf die Beine gestellt
('Contrary to the agreement with the organization, Roth is said to have organized his own celebrations parallel to the district festival')

(6.11)Denn nach ihrem durchschlagenden Erfolg vom vorigen Jahr stellten die Stadtjugendpflegerinnen Petra Bliedtner und Petra Vogel-Jones wieder eine Mädchen-Aktionswoche für Zehn- bis Fünfzehnjährige auf die Beine
('Since their striking success the year before the youth organizers Petra Bliedtner and Petra Vogel-Jones organized another girls' action week for 10 to 19 year olds')

(6.12)Vielleicht mit anderen Tänzern , die einmal bei Forsythe gearbeitet haben - viele leben im Raum Frankfurt - etwas auf die Beine stellen , oder therapeutisch arbeiten
('Perhaps setting something up with other dancers who have worked with Forsythe – any of them live in the Frankfurt district – or doing some therapy')

(6.13)" Ohne dieses Netzwerk " , sagt Negel , " könnten wir als ehrenamtlich tätiges Organisationskomitee einen solchen Kongreß gar nicht auf die Beine stellen
('Without this network, said Negel, we could not, as a voluntary organization, organize such a congress')

In the following, examples for morphological variation of preposition, noun, and verb are presented.

Morphological variation of the preposition: Usually there is little variation with respect to the preposition in the PP-collocate. For some collocations variation between plain preposition and preposition with incorporated determiner exists, see for instance {*zur, zu*} *Verfügung* (at (the) disposal), {*zum, zu*} *Ergebnis* (to (the) result). In these cases, corpus data give information about preferred usage. In both, the newspaper and the newsgroup corpus, the variant *zu Verfügung* is rare. There over 900 instances of *zur Verfügung* versus 6 instances of *zu Verfügung* in the newspaper corpus, and 1005 instances of *zur Verfügung* versus 28 instances of *zu Verfügung* in the newsgroup corpus, providing strong evidence that *zur Verfügung* is the more common variant. In contrast, there are 8 instances of *zum Ergebnis* versus over 40 instances of *zu (Det) Ergebnis* in the newspaper corpus, indicating that the variant where preposition and determiner are separated is more common. This is also supported by the newsgroup corpus, where 61 occurrences of *zu (Det) Ergebnis* are opposed to 21 occurrences of *zum Ergebnis*.

Morphological variation of the nominal collocate like *zum* {*Zuge, Zug*} *kommen* where *Zuge* is an archaic strong declension form and more likely to be part of a collocation than *Zug*. There are 303 instances of *Zug* and 209 instances of *Zuge* in the newspaper corpus. All *Zuge*-instances take part in collocations. There are 166 instances of *im Zuge* immediately followed by a genitive, 6 of which are pseudo-genitives realized as PP_{von}. *Im Zuge*, in this case, is a word level collocation meaning ‘during’. In addition, there are 33 instances of *zum Zuge kommen* (‘get an opportunity’), and 9 instances of *am Zuge sein* (‘have an opportunity’). In comparison, there are 14 instances of *im Zug*, but only 8 are followed by a genitive, 2 of which do not allow for the collocational reading ‘during’. There 19 instances of *zum Zug*, 18 of which are part of the collocation *zum Zug kommen*, and there are 16 instances of *am Zug* where 14 collocate with *sein* (be)[3]. All in all, the corpus examples confirm that the archaic form *Zuge* is a good indicator for collocativity, whereas the form *Zug* cannot be used for distinguishing between collocativity and noncollocativity.

Morphological variation of the verbal collocate: In general, variation of preposition and noun in PP-verb collocations is either impossible or strongly restricted, variation of the verbal collocate, on the other hand, is free,

---

[3] *jemand ist am Zug* (‘it is someone’s goal’)

even though the corpus reveals usage preferences. In the case of *im Alter (von . . .) sterben* (die at the age (of . . .)), for instance, the verb exclusively occurs in past tense in the newspaper corpus either as a past participle – *gestorben* ('has died') – or a finite verb, third person, singular – *starb*, (died). Similarly, the collocations *unter Berufung (auf . . .) berichtete* (referring (to . . .) reported) (past, third person singular), *nach Angaben getötet*, 'according to . . . killed' (past participle) occur only in the particular realizations in the extraction corpus.

6.3.2 Modification Patterns

There are two possibilities for modifying PP-verb collocations, namely modification in the PP, and modification at clause level. Modification, if not fixed or blocked, is open to variation. The crucial point with respect to the latter case is that from a competence-based view modification may be rather flexible while actually occurring examples will be much more restricted. In order to cope with this discrepancy, information on modification is extracted from corpora, and stored in the realization part of the collocation database.

Modification of the Collocation as a whole

A particular class of collocation-specific modifiers, namely adverbs and predicative adjectives, can be found adjacent to the PP-collocate. Thus for each collocation, a set of modifiers can be automatically accessed from corpus data, which is particularly important for collocations with flexible modification. In this case, corpora provide information on the typicality of particular modifiers for a certain collocation. This information can be utilized for elaborating natural language generation components.

The following examples have been taken from the newspaper corpus.
{*automatisch, endlich, gerade, sofort, sogar, später*} *in Kraft*
({automatically, finally, just, immediately, even, later} into force),
{*derzeit, gestern*} *im Gespräch*
({at present, yesterday} in a conversation),
{*nicht, nicht mehr, schon, wieder*} *in Betrieb*
({ not, no longer, already, again} in operation),
{*bedingt, kostenlos, nicht, noch, voll, vorübergehend, wieder*} *zur Verfügung*
({conditionally, free of charge, not, still, completely, temporarily, again} at the disposal of),
{*nicht, noch, nur, wieder*} *in Frage*
({not, still, only, again} into question),
{*noch*} *in den Anfängen*

({still} at the beginning).

Besides being useful for lexical selection in generation and machine translation, corpus data on modification can also be employed for restricting the prediction of collocation partners, see for instance {*aber, fest*} *ins Auge fassen* ('contemplate doing something', 'plan something'), {*besonders, ohnehin zu sehr*} *ins Auge fallen* ('attract ones attention', 'attract ones attention too much'), {*direkt*} *ins Auge stechen* ('catch one's eye'), which are examples of collocations that contain the PP-collocate *ins Auge*. There was no overlap found in the extraction corpus with respect to modification between the PP-collocates of the three collocations, even though the modifiers are interchangeable among the collocations from a competence-based point of view, except for *fest* which does semantically not very well combine with *ins Auge fallen* and *ins Auge stechen*.

Modification in the PP-Collocate

In general, modification of PP-collocates is either blocked or strongly restricted. Considering randomly selected SVCs from the extraction corpus, there is strong evidence that predicative nouns typically occur without modification. An example is *zur Verfügung*. Among 423 instances of *zur Verfügung stehen* occurring in the newspaper corpus, there is only one which is modified, i.e., *zur sprachlichen Verfügung stehen* ('have available a wide variety of expressions').

Reverse cases identified from the newspaper corpus are:

auf den {*neuesten, neuesten technischen, neuesten ökologischen, modernsten*} *Stand (bringen)*
('bring up to date', 'bring technically up to date', 'bring environmentally up to date', 'bring up to date')
auf einen {*knappen, kurzen, gemeinsamen, einfachen*} *Nenner (bringen)*
('reduce to a {concise, concise, common, simple} denominator')
auf {*finanziell dünnen, schwachen, wackligen, eigenen*} *Beinen (stehen)*
(on {financially thin, weak, shaky, ones own} legs stand)
('be financially weak', 'be shaky', 'be shaky', 'be one's own boss')
auf {*gesunden, wackligen, eigenen, mehreren*} *Füßen (stehen)*
(on {healthy, shaky, ones own, several} feet stand)
('rest on a healthy foundation', 'rest on a shaky foundation', 'to stand on one's own two feet', 'to have a broad base')

In all of these cases, except for *auf einen Nenner bringen*, only the modified variant exists. The examples also illustrate that there is a lexical relation between the prenominal modifiers and the verbs. In other words, the semantics as well as the morphosyntactic appearance[4] of the adjectives in the particular

---

[4] See for instance *auf* {*gesunden, wackligen, eigenen, mehreren*} *Füßen (stehen)* (dative)

PP-collocates create expectations about the verbs to come.

6.3.3 Recurrent Syntactic Realizations

Corpus data can also be utilized for detecting preferred linear order, and for supporting determination of attachment sites. The two aspects will be illustrated using the PP-verb collocation *zu Felde ziehen* (‘to act against something/someone’, ‘to campaign against something/someone’). In the majority of occurrences in the newspaper corpus, the PP-argument against what or whom the action is directed comes immediately to the left of the PP-collocate. There is only one exception, example (6.14)h. where the verbal collocate precedes the PP-collocate. See (6.14) for the corpus realizations of *zu Felde ziehen*. Translations will be given for the collocation and its PP-argument.

On the other hand, the knowledge on the collocativity of *zu Felde ziehen* can be used for predicting PP-attachment in parsing. This is particularly useful when parsing a sentence like j. Here high attachment (i.e. attachment to the main verb) of the three PPs *an der Rhönstraße (‘in Rhönstraße’), im Stadtteil Bischofsheim (in the Bischofsheim district), mit 18 Sozialwohnungen* (with 18 council flats) can be ruled out. In general, knowledge about PP-verb collocations allows ruling out attachment of the PP-collocate to a preceding noun, or attachment of PPs to the PP-collocate.

(6.14)a. früher ” **gegen** die Überfremdung von Volk und Heimat ” **zu Felde zog** (‘campaigned against infiltration of foreigners into one’s own land’)

b. verbal **gegen** eine Verwarnung **zu Felde gezogen** war (‘speak out against a warning’)

c. seit vielen Jahren mit Information und Aktion **gegen** alle Formen von Intoleranz **zu Felde ziehen** (acts against all forms of intolerance)

d. doch nicht **gegen** die Ost-Trainer **zu Felde ziehen** (‘campaign against the trainers from the East’)

e. die seit den sechziger Jahren **gegen** eine romantisierend-verklärende Volkskunde **zu Felde zieht** (‘campaign against a blissfully romanticised folklore’)

f. **zog** seinerzeit ” nur ” **gegen** ” Republikflüchtlinge ” **zu Felde** (‘campaigned against deserters from the Republic’)

g. **gegen** die Chlorchemie **zu Felde** gezogen (‘campaigned against chlorine industry’)

---

but
auf {gesunde, wacklige, eigene, mehrere} Füßen (stellen) (accusative)

h. Aber nicht nur **gegen** den gestrauchelten Ex-Präsidenten **zieht** Rosa e Silvas Buch **zu Felde** (‘camaign against the fallen ex-president’)

i. vehement **gegen** die Überbauung und Zerstörung der Landschaft **zu Felde gezogen** ist (‘campaigned against the development and destruction of the landscape’)

j. seit gut einem Jahr **gegen** die - inzwischen von der rot-grünen Koalition im Maintaler Stadtparlament beschlossene - Bebauung einer Grünfläche an der Rhönstraße im Stadtteil Bischofsheim mit 18 Sozialwohnungen **zu Felde zieht** (‘campaign against the building of 18 council flats in the green belt area in Rhönstraße in the Bischofsheim district which has been decided by the red-green coalition in Maintal city parliament’)

6.4 CDB – The Collocation Database

Two kinds of data need to be represented in a collocation database which aims at accounting for both, generative and rigid, aspects of collocations, these are

- linguistic descriptions of collocation types, and
- example instances derived from various corpora.

The linking of linguistic descriptions and corpus examples is of particular interest. Book-keeping information such as corpus name and sentence number needs to be stored, in order to be able to trace back the origin of a particular example, and to access larger contexts. The database is required to be extendible with respect to linguistic descriptions and corpus data. The information needs to be represented in such a way that flexible views on the data can easily be provided. This is particularly important as the database on the one hand is conceived as a research tool which supports the development of collocation theories, and on the other hand, it is intended to function as a collocation lexicon supporting parsers and generators. The previously stated requirements are best met by a relational database. See section 2.4.1 for a brief introduction to concept and basic terminology.

6.4.1 The Entity-Relationship Model

The relational model of CBD is defined by six base relations or entities which are linked via keys. The conceptual structure of the collocation database is illustrated in figure 6.1. The individual attributes are described in section 6.4.2.

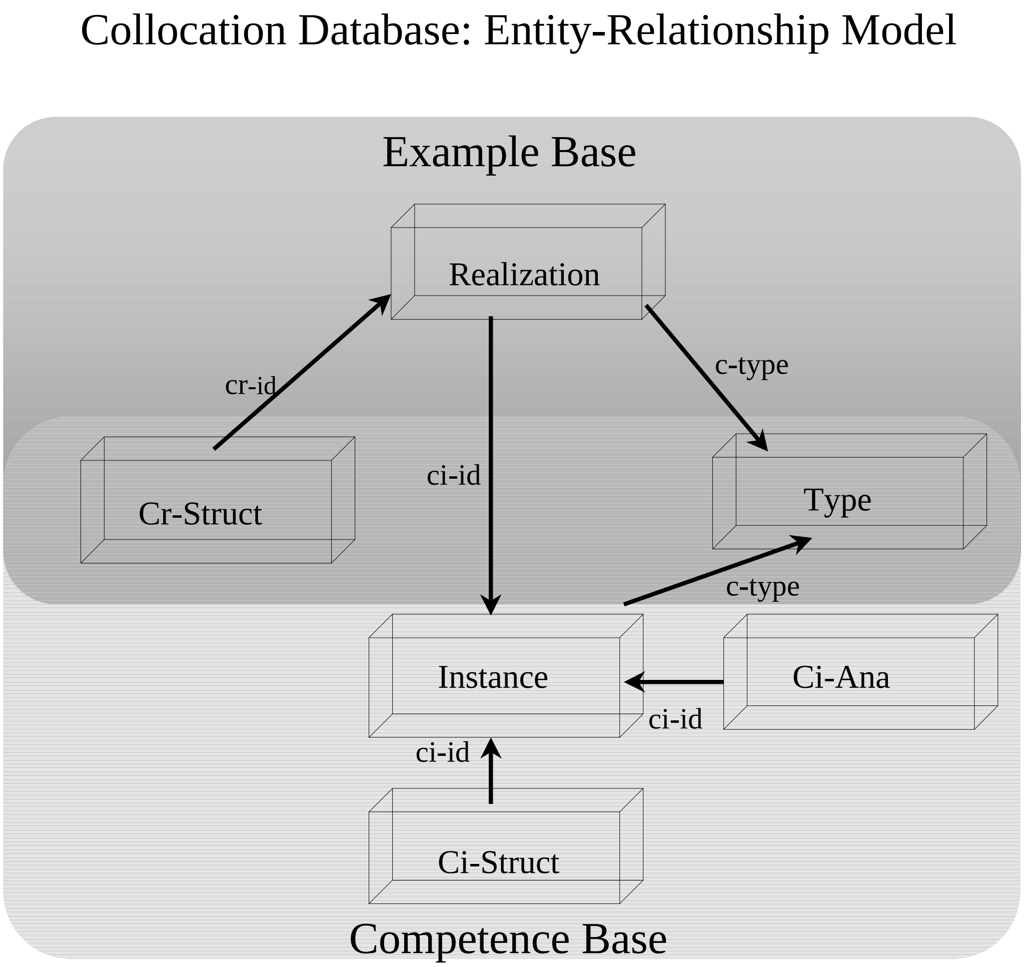


Figure 6.1: The entity-relationship model of CDB

The figure depicts the distinction of competence base and example base, and how linguistic description and performance data relate to each other. The distinction between competence and example base is represented by the two relations COLLOCATION-INSTANCE (Instance) and COLLOCATION-REALIZATION (Realization) with the former being the core of the competence base and the latter being the one of the performance base. In this vein, a collocation instance in CDB is a generalized representation of a collocation consisting of a preposition, a noun and a verb, whereas a collocation realization is an actually occurring surface form of a collocation within its sentential context. The competence base contains two additional core entities, which are CI-STRUCTURE (Ci-Struct) and CI-ANALYSIS (Ci-Ana). While the lexical properties of the collocates of a collocation are represented in COLLOCATION-INSTANCE, the syntactic and collocation-type-specific properties of collocations are represented in CI-STRUCTURE and CI-

ANALYSIS, respectively. In order to connect competence and performance base, COLLOCATION-INSTANCE and COLLOCATION-REALIZATION are linked, i.e., each corpus example or realization of a collocation is associated with exactly one tuple in COLLOCATION-INSTANCE. As the tuples in CI-STRUCTURE and CI-ANALYSIS are related to COLLOCATION-INSTANCE, a competence-based model for each realization of a collocation stored in the database exists. The competence part is what current representations of collocations are all about. The novelty of the representation presented in this work is that, other than in the existing approaches, real-world occurrences of collocations are used in a large scale to constrain the overgenerating competence-based descriptions. Thus corpus-based methods are employed to cope with the nongenerative aspects of collocations. In particular, each realization is associated with a structural representation on its own which is represented by the entity CR-STRUCTURE. The specialty is that the attributes are similar to those in CI-STRUCTURE, which is important for systematic investigations of the relationship between generative and static aspects in collocations, and a further step towards a theory of grammar where a model of collocational aspects of language is an integrative part. COLLOCATION-TYPE is the other relation which is conceived as being located in the intersection of competence and performance base.

6.4.2 Relations and Attributes

In the following, the relations and their attributes will be described in more detail.

COLLOCATION-TYPE

COLLOCATION-TYPE is the most abstract level of representation where collocations are grouped into larger classes, currently SVC, figurative expression and pseudo-collocations. This is only a coarse classification which in the long run shall be exchanged by a classification along various dimensions such as syntactic flexibility versus rigidity, semantic interpretability versus opacity, kind of interpretation available such as literal, figurative or metaphoric, domain-specifity of a collocation, and pragmatic function. In the current database, the closest approximation to domain-specificity is the name of the corpus within which the collocation realization has been found. Such an approach, however, is only promising when the corpora differ from each other with respect to the domains contained.

Overview of the currently specified attributes:

c-type type of collocation such as support-verb construction, figurative expression, etc.;

ct-domain language domain(s) to which the collocation belongs;

ct-comment user-defined comment related to the entity COLLOCATION-TYPE.

COLLOCATION-INSTANCE

Collocation instances are generalized representations of the major collocates of a collocation. For most cases of the PNV-data, this is a triple comprising the respective full form of preposition and noun, and a base form of the verb, here bare infinitive. See table 6.5 for illustration.

ci-id	c-type	ci-string
2012	SVC	zu Verfügung stellen
2013	SVC	zur Verfügung stellen
2014	SVC	zur Verfügung stehen
2015	SVC	zur Verfügung haben
1745	SVC	in Betrieb gehen
1746	SVC	in Betrieb nehmen
1751	SVC	außer Betrieb setzen
1752	SVC	außer Betrieb gehen
2802	figur	unter Lupe nehmen
2823	figur	am Herzen liegen
28113	figur	in Teufels Küche kommen

Table 6.5: The relation COLLOCATION-INSTANCE and its attributes

The collocates are represented in the attribute ci-string. In cases where the PNV-combination only partially covers the collocation such as *in Teufels Küche kommen*, ci-string is correspondingly larger. Each collocation instance is currently related to exactly one collocation type by means of the attribute c-type. This kind of representation contains a strong potential for generalizations, for example, the variants of PP-collocates containing either a simple preposition or a preposition fused with a determiner can be collapsed; it is also informative to group collocations along predicative nouns, or along dichotomous verb pairs like *stehen - stellen, setzen - sitzen, legen - liegen*, etc.

Overview of the attributes:

ci-id identification number of collocation instance;

c-type the collocation type the collocation instance is related to;

ci-string generalized representation of the collocates, i.e. full forms, base forms or regular patterns;

ci-comment user defined comment related to the entity COLLOCATION-INSTANCE.

COLLOCATION-REALIZATION

The relation is defined for storing sentences identified from corpora which contain occurrences of a collocation-instance. For each example sentence, the following information is stored: the surface realization cr-string, a unique identification number cr-id, the number of the collocation instance ci-id the sentence is an example of, the corpus cr-source it has been retrieved from, and the number the sentence cr-number relative to the other sentences in the corpus from which the collocation example has been identified. See table 6.6.

cr-id	ci-id	cr-sent	cr-source	cr-s-num	c-type
508	2	3800 Quadratmeter Fläche auf drei Etagen stehen in dem Neubau nun dort zur Verfügung , wo einst Kühe in Ställen untergebracht waren	ger03fi	409585	SVC

Table 6.6: The relation COLLOCATION-REALIZATION and its attributes

Overview of the attributes:

cr-id identification number of a particular realization of a collocation;

ci-id the identification number of the collocation instance to which the collocation example (realization) is related;

c-type the collocation type the collocation instance is related to;

cr-source the corpus in which the collocation has been found;

cr-s-num the sentence number relative to the corpus;

cr-sent a sentence from a corpus that contains the realization of a particular collocation instance;

cr-comment user defined comment related to the entity COLLOCATION-REALIZATION.

CI-STRUCTURE and CR-STRUCTURE

The two relations store information on the basic syntactic structure of a collocation instance or collocation example. Table 6.7 shows the structural description of the collocation instance *zu Verfügung stellen*. In the first three lines the canonical positions of the collocates and their function in the collocation are described.

ci-id	ci-position	ci-substring	ci-category	ci-function
1	0-1	zu	APPR	prep
1	1-2	Verfügung	NN	N_{pred}
1	2-3	stellen	VV	V_{sup}
1	0-2	zu Verfügung	PP	P_{pred}

Table 6.7: The relation CI-STRUCTURE and its attributes

In the fourth line it is stated that the preposition and the noun constitute a PP which is the predicative phrase (P_{pred}) of the collocation, i.e., the phrase constituted by the predicative noun (N_{pred}) of the SVC. V_{sup} stands for support-verb, and 'prep' for prepositional collocate. As the representation of *zu* and *zur Verfügung* just differ in the morphosyntactic realization of the preposition, the descriptions are promising candidates for being merged.

ci-id	cr-id	cr-position	cr-substring	cr-category	cr-function
2	508	12-13	zur	APPRART	prep
2	508	13-14	Verfügung	NN	N_{pred}
2	508	12-14	zur Verfügung	PP	P_{pred}
2	508	6-7	stehen	VVFIN	V_{sup}
2	520	9-10	zur	APPRART	prep
2	520	11-12	Verfügung	NN	N_{pred}
2	520	9-12	zur sprachlichen Verfügung	PP	P_{pred}
2	520	12-13	stand	VVFIN	V_{sup}
1	532	20-21	zu	APPR	prep
1	532	21-22	Verfügung	NN	N_{pred}
1	532	20-22	zu Verfügung	PP	P_{pred}
1	532	8-9	stehen	VVFIN	V_{sup}

Table 6.8: The relation CR-STRUCTURE and its attributes

Table 6.8 presents structural information related to collocation realizations. The particular sentences are[5]

3800 Quadratmeter Fläche auf drei Etagen stehen in dem Neubau nun dort zur Verfügung , wo einst Kühe in Ställen untergebracht waren.

---

[5] Translations are omitted, as they are not relevant in this context.

Wer sich von der Grazie , die dem Literaten Fénéon zur sprachlichen Verfügung stand , ein optisches Bild machen will , das seiner brillanten und geheimnisvollen Lakonie entspricht , der kann sich in den Filmen , die der Georgier Otar Jiosseliani - z. B. den Günstlingen des Mondes - in Frankreich gedreht hat , eine Vorstellung davon machen.

In den drei Kindergarten- und den zwei Hortgruppen stehen dem Nachwuchs 60 Quadratmeter in Gruppen- und 20 Quadratmeter in Kleingruppenräumen zu Verfügung.

Collocates and collocation phrases are associated with position, part-of-speech and collocation-internal function labels. Each other word in the example sentence is associated with position and part-of-speech information. Table 6.8 illustrates the following kinds of differences between the particular realizations. While in examples 508 and 532 the support-verb precedes the predicative phrase, the opposite is the case in example 520. In the former examples preposition and noun collocate are adjacent whereas they are interleaved by *sprachlichen* in the latter example. In all examples, the surface realizations of the nouns are identical, prepositions and verbs vary. The realizations are related to the instances *zur verfügung stellen* and *zu verfügung stellen* by means of the attribute ci-id.

Overview of the attributes of CI-STRUCTURE:

ci-id the identification number of the collocation instance;

ci-position the position of the substring;

ci-substring the substrings constituting the collocation, i.e. the individual collocates and minimal collocation phrases;

ci-category the syntactic categories of the substrings;

ci-function the collocation-specific functions of the collocates and collocation phrases;

ci-str-comment user defined comment related to the entity CI-STRUCTURE.

Overview of the attributes of CR-STRUCTURE:

ci-id the identification number of the collocation instance;

cr-id the identification number of a particular realization of a collocation instance;

cr-position the position of a collocation phrase (i.e., a phrase containing a collocate) within a particular example sentence;

cr-substring the collocation phrase;

cr-category the syntactic category of the collocation phrase;

cr-function the collocation-internal function of the collocate or collocation phrase;

cr-str-comment user defined comment related to the entity CR-STRUCTURE.

CI-ANALYSIS

The relation is designed for representing collocation specific properties leading to flexible and extendible collocation-specific descriptions. The relation ci-analysis is strongly underspecified with its three attributes ci-id, ci-attrib, and ci-value.

The values of ci-attrib and ci-value are pairwise defined for each data-record allowing the definition of analysis schemes of different classes of collocations. Appropriateness checks on the pairs, however, are outside of the scope of TSDB, as the database technology does not support consistency checks between values of different attributes. This can be easily achieved by an extra program which operates on the ASCII-file storing the relation CI-ANALYSIS. In table 6.9 the attribute-value pairs for the description of SVCs are listed representing the linguistic analysis given in section 3.4.3.

ci-attrib	ci-value
caus	{+, - }
a-art	{incho, contin, term, neut}
reciproc	<ci-id>
args	<subcategorization frame of SVC>
p-det	{-, u, <realization>}
p-modpre	{-, u, <realization>}
p-modpost	{-, u, <realization>}
mods	{-, u, <realization>}

Table 6.9: Attribute-value pairs for CI-ANALYSIS of SVCs

The SVC-specific information related to *zur Verfügung stellen* is presented in table 6.10. The collocation instances for which the description is valid are identified by ci-id. In the current example this is the collocation instance with ci-id = 1. The collocation is causative, and has inchoative Aktionsart. With respect to argument structure at least a causer and a theme realized as NPnom and NPacc respectively are required. The surface realization of a dative object is optional as far as competence grammar is concerned. The availability of example sentences from corpora allows insights into the realization of competence-grammatically optional arguments. **P-det, p-modpre,** and **p-modpost** specify properties of the predicative phrase, i.e., determination is underspecified. Thus no particular alternatives are listed, and information on the actual realization of determination needs to be derived from the corpus data. Similarly, information on prenominal modification (p-modpre = u) and modification of the whole SVC (mods = u) is underspecified. For information on restrictions, again corpus examples need to be accessed. In contrast, postnominal modification in the predicative phrase is an example for a feature which is blocked, i.e., p-modpost = -.

ci-id	ci-attrib	ci-value
1	caus	+
1	a-art	incho
1	reciproc	2
1	args	NPnom (NPdat) NPacc
1	p-det	u
1	p-modpre	u
1	p-modpost	-
1	mods	u

Table 6.10: The relation CI-ANALYSIS specified for the SVC *zur Verfügung stellen*

Overview of the attributes of CI-ANALYSIS:

ci-id the identification number of the collocation instance;

ci-attrib dummy attribute the values of which are defined according to a particular collocation-type-specific analysis;

ci-value dummy attribute the values of which are defined such that a attribute-value relation with the values of ci-attrib is established;

ci-ana-comment user defined comment related to the entity CI-ANALYSIS.

6.5 Example Queries

As stated earlier, query results are unnamed derived relations. Tsdb allows writing a query result to a user-defined plain ASCII file, which eases further processing. In the example results given below, the table fields are separated by |.

```
retrieve ci-string where c-type = ‘‘SVC’’.
```

This is a simple query for retrieving all SVCs from the Database. The output is a list of collocation instances like the one below.

```
am anfang stehen
am beginn stehen
am ende stehen
am leben bleiben
am leben erhalten
am leben halten
an arbeit gehen
an arbeit machen
an bedeutung verlieren
an land gehen
an land kommen
an macht bleiben
zur ruhe kommen
zur sache gehen
zur sache kommen
zur schau stellen
zur sprache bringen
zur sprache kommen
zur verantwortung ziehen
zur verfügung bekommen
zur verfügung stehen
zur verfügung stellen
zur vernunft bringen
zur vernunft kommen
zur verzweiflung bringen
...
```

```
retrieve ci-string where ci-string ∼ "stehen" | ci-string ∼ "stellen".
```

The query allows retrieving all collocation instances that contain the collocates *stehen* or *stellen*.

```
außer frage stehen
in frage stehen
in frage stellen
vor frage stehen
vor frage stellen
```

```
auf füße stellen
auf füßen stehen
vor gericht stehen
vor gericht stellen
im mittelfeld stehen
im mittelpunkt stehen
in mittelpunkt stehen
in mittelpunkt stellen
im raum stehen
in raum stellen
in rechnung stellen
vor schließung stehen
unter schutz stehen
unter schutz stellen
im regen stehen
im schatten stehen
in schatten stellen
```

```
retrieve cr-sent cr-source cr-s-num where ci-string = "zur verfüg-
ung bekommen".
```

Retrieve all sentences (cr-sent) from the database that contain an instance of the collocation (ci-string) *zur verfügung bekommen*. Also retrieve the sentence number (cr-s-num) and the name of the corpus (cr-source) from which the example originated. The two examples come from the same corpus, namely ger03f-i which is a 8 million portion of the Frankfurter Rundschau Corpus. In order to access a broader context, the sentence numbers are required.

```
Der Mörfelder bekam einen 40-Tonnen-Laster zur Verfügung , dazu
einen Schiffscontainer , der auf dem Seeweg nach Sankt Petersburg
kommt | ger03f-i | 227762

Mehrmals haben die Juz-Betreiber im vergangenen Jahr beim zuständigen
Ortsbeirat 2 und beim Sozialdezernenten Martin Berg versucht , mehr
Mittel zur Verfügung zu bekommen | ger03f-i | 500877
```

```
retrieve ci-string cr-sent where ci-id = 2.
```

Retrieve the collocation (ci-string) which has identification number 2 (ci-id = 2) and related example sentences (cr-sent).

```
zur verfügung stehen | " 1992 standen uns für den Bezirk Gießen 17
Millionen zur Verfügung " , sagt der stellvertretende Arbeitsamts-
direktor Schäfer , " für das laufende Jahr hatten wir eine

Zuteilung von 500 000 Mark , dann kam der Stopp dazwischen

zur verfügung stehen | " Dann steht kein Geld mehr zur Verfügung

zur verfügung stehen | " Der hat doch dadurch nicht mehr Rechte
gehabt als irgend jemand anderer " , sagt Albert Burkhardt , der
noch vor zwei Wochen versichert hatte , daß Hofmann für den
Posten wieder zur Verfügung stehen werde

zur verfügung stehen | " Ein bißchen " sauer ist Hofmann allerdings
auch auf Burkhart , weil der ein " bißchen zu optimistisch " gewe-
sen sei und voreilig gesagt habe , daß er , Hofmann , wieder zur
Verfügung stehen werde

zur verfügung stehen | " Erhebliche Mittel aus Bonn und Brüssel "
stünden jedoch zur Verfügung

zur verfügung stehen | " Es gibt keine Grundsatzerklärung des Magis-
trats , daß das Osthafenareal nicht zur Verfügung steht

zur verfügung stehen | " Ich stehe für öffentliche oder nicht-
öffentliche Schlammschlachten nicht zur Verfügung " , hatte
Kassierer Peter Oelschläger dem Grünen-Kreisvorstand geschrieben
und wissen lassen , daß auch er " kein Interesse mehr an einer
Zusammenarbeit " habe

zur verfügung stehen | " Ich stehe im Herbst zur Verfügung

zur verfügung stehen | " In nicht allzuferner Zukunft werden hier
weitere Wohnungen zur Verfügung stehen
```

```
retrieve ci-id ci-string ct-type ci-position ci-substruct ci-category.
```

Values for the attributes ci-id, ci-string, ct-type, ci-position, ci-substruct and ci-category are retrieved. The second line of the query output says that *in anfängen stecken* has identification number 1001, ist is an SVC and has a substring of length 2 ranging from position 0 to position 2. The abstract lexical realization of the substring is *in anfängen*, i.e. no morphological variation of preposition and noun collocate is allowed. The syntactic category of the substring is PP. Examples for lexical variability are the verbs, e.g *:stecken* and the noun *sande?*. While no collocation-specific restriction applies to the former, the latter may occur in two realizations, namely *sand* and *sande*. This kind of information is part of the competence-based description of the collocation entries. Information about the commonness of theoretically assumed variants can be derived from the related corpus data.

```
1001 | in anfängen stecken | SVC | 0-1 | in | APPR
1001 | in anfängen stecken | SVC | 0-2 | in anfängen | PP
1001 | in anfängen stecken | SVC | 1-2 | anfängen | NN
1001 | in anfängen stecken | SVC | 2-3 | :stecken | VV
1006 | im schatten stehen | figur | 0-1 | im | APPRARTd
1006 | im schatten stehen | figur | 0-2 | im schatten | PP
1006 | im schatten stehen | figur | 1-2 | schatten | NN
1006 | im schatten stehen | figur | 2-3 | :stehen | VV
1007 | im sande verlaufen | figur | 0-1 | im | APPRARTd
1007 | im sande verlaufen | figur | 0-2 | im sande? | PP
1007 | im sande verlaufen | figur | 1-2 | sande? | NN
1007 | im sande verlaufen | figur | 2-3 | :verlaufen | VV
```

```
retrieve cr-s-num cr-id ci-string cr-substring cr-function cr-sent
where ci-string ∼ "ins auge" & cr-function ∼"Col".
```

Sentences which contain a collocation with a PP-collocate *ins Auge* are retrieved. Apart from the collocation instance (ci-string), the actual realization of the PP-collocate (cr-substring where cr-function ∼ "Col") in a particular sentence and the sentence itself (cr-sent) are retrieved.

```
110713 | 8 | ins auge fassen | ins Auge | Unter diesem Dachverband
haben sich 1965 die landwirtschaftlichen Betriebe national organi-
siert , die einerseits den Tourismus als Einnahmequelle ins Auge
fassen , andererseits aber auch zur Erhaltung und Vermittlung
der ländlichen Kultur beitragen wollen

115684 | 9 | ins auge fassen | ins Auge | Es gibt im Bereich der
Straße am Alten Bach noch Flächen , die man dafür mal ins Auge
fassen sollte

1345 | 1 | ins auge fassen | ins Auge | Oder die gezielte Züchtung
```

```
anenzephaler Föten könnte ins Auge gefaßt werden , deren Organe
bis weit über dieses Datum hinaus unbedenklich
entnommen werden könnten

154913 | 10 | ins auge stechen | ins Auge | Der Platz im Regal ist
so groß , da dieser beim Vorbeigehen direkt ins Auge sticht

17377 | 3 | ins auge fassen | ins Auge | das Entnazifizierungs-
komitee des Literaturbetriebs faßte ihn scharf ins Auge , und
das war , wie Assouline enthüllt , der wahre Grund für
die rasche Abreise des Schriftstellers nach Amerika

191820 | 11 | ins auge fallen | ins Auge | Die Partei mit der großen
Mehrheit findet , die Fülle der Plakate verunstalte das Ortsbild ,
zumal sie gerade dort aufgestellt werden , wo sie ins Auge fallen
und damit stören würden

254342 | 13 | ins auge fassen | ins Auge | Für 1993 solle das Haus
Schlesinger darüber hin aus eine Senkung der Leitzinsen ins Auge
fassen und " nicht erst das Licht anmachen , wenn die Konjunktur
die Kellertreppe heruntergefallen ist "

539704 | 30 | ins auge fassen | ins Auge | " Wenn die Grünen bei
den Koalitionsverhandlungen Forderungen stellten , " die an die
Substanz unserer Vorstellungen gehen " , müßten auch andere
Konstellationen ins Auge gefaßt werden , obwohl es eine Präferenz
für die Fortsetzung der bisherigen Koalition gebe
```

6.6 Additional Facilities

6.6.1 Exploitation of the Database Output

While the database enables flexible views on the data, additional facilities are required for further exploitation of the database output. An important task is extraction of frequency information from the tables resulting from queries to the example base.

The following information is of interest:

- average number of words in the PP-collocate;
- average distance between PP-collocate and verb collocate, measured in words or phrases;

- typical modification in the PP-collocate, such as statistics over positions, words, parts-of-speech and syntactic structure of the modifiers;
- material between the PP-collocate and the verb collocate, such as statistics on syntactic structure and lexical realization of the intervening material.

The information is of particular interest for the following tasks:

- development of a theory of collocations;
- decisions on which information shall be incorporated into the competence base;
- construction of specialized collocation lexica for natural language parsing, generation and machine translation.

Already existing tools can be used for exploiting the query results. Corset, the predecessor of Gsearch , for instance, allows specifying n-gram frequencies over words and/or tags. In Gsearch [Keller *et al.*, 1999] context-free grammars can be written based on which the database output is examined with respect to the frequency of user-defined syntactic structures. Additional programs are required for processing distance information, either operating on database output or output from Corset and Gsearch. This can be easily achieved, as all output is written to plain ASCII-files.

6.6.2 Automation of Database Construction

The relation files constituting the current collocation database are to a large extent generated automatically. With respect to database construction, as many data records as possible are generated automatically. Thus implementation and extension of the collocation database is supported by the availability of means for syntactic preprocessing of the extraction corpora, and the automation of identification of certain collocation types. For the time being, the following aspects of database constructions are automated:

Representation of Collocation Instances and Realizations: The data structures in COLLOCATION-INSTANCE and COLLOCATION-REALIZATION are well suited for automatic construction, as they mainly contain bookkeeping information and plain lexical data. The only piece of information which currently needs to be hand corrected is collocation type. Empty comment fields are generated, as the particular fields are reserved for user defined comments. This holds for the comment fields in all relations.

Representation of Collocation Structures: Due to syntactic preprocessing of the extraction corpora, the data for CR-STRUCTURE can be automatically generated. The accuracy of the data obviously depends on the accuracy of the preprocessing tools applied. Systematic errors, however, are most likely to be detected in the output resulting from queries to the database, and can be easily corrected by manipulating the according entries in the data files. The entries in CI-STRUCTURE can be created automatically as well, as this information is perfectly regular.

6.7 Conclusion

Summing up, collocations are represented in the database at three levels of abstraction: (1) Collocation types: Currently three types are distinguished which are support-verb construction, figurative expression and pseudo-collocation. (2) Collocation instances: These are preposition-noun-verb triples where the verbs are reduced to base forms. (3) Collocation realizations: For each collocation instance, a number of realizations is stored which have been identified from text corpora. Collocation instances and realizations are described at morphosyntactic, lexical and structural level. A characteristic of the descriptions is that the same representations are used for instances and realizations, which allows integrating competence and performance aspects of collocations. In addition, for each collocation instance a collocation-type-specific analysis is given. As each realization is linked to a collocation instance, the analysis is also accessible via the realization. This way, linguistic analysis and actually occurring data complement each other, whereby competence-based linguistic description and analysis of the collocation instances are a means to cope with the incompleteness of corpus data, and the base of collocation realizations, on the other hand, is a means to account for seemingly nongenerative aspects of collocations. To achieve this task, it is important that large numbers of realizations originating from different domains are accessible via the database. With the availability of large numbers and a broad variety of examples, the linguistically annotated collocation instances are used for generating new collocation instances which account for prevalent regularities in the corpus data. Thus information on the usage of collocations is introduced into a higher level of abstraction, and the database functions as a resource for theory building, and as well as a basis from which input structures for collocation analyzers and generators can be built. In addition, the linguistically annotated example sets can be used as training material for inducing stochastic models of individual collocations. Such a statistical approach is expected to be an alternative to a principled account of collocations.

Chapter 7

Summary and Outlook

7.1 Summary

In the work presented, two major problems related to lexical collocation phenomena are addressed:

1. insufficiency of merely frequency-based or statistics-based approaches to collocation identification, and
2. inappropriateness of competence grammatical analyses and descriptions of collocations.

In order to account for the former, an approach to collocation identification has been devised, where statistical techniques and knowledge on distinctive linguistic properties of collocations have been combined. As a step towards a solution of the latter, a representation model and database for collocations have been developed and implemented, where linguistic descriptions of collocations and data on real-world occurrences are combined.

The overall conclusion to be drawn from the identification part of the present study is that a purely statistics-based approach to collocation identification needs further improvement by incorporating linguistic information.

There are two essential problems of a purely statistics-based approach: First, in many cases, collocational and noncollocational word combinations do not differ in their frequency distributions. Secondly, statistical measures tend to overestimate low frequency data. This is particularly the case for measures that do not account for the significance of the data examined. Based on a variety of experiments on corpus-based collocation identification it could be shown that linguistic information is optimally used when employed at different stages of the collocation identification process.

The following strategy has proven successful: First of all, the corpus used for collocation identification is automatically part-of-speech tagged and annotated with rudimentary syntactic structure. From a thus annotated corpus, collocation candidates are selected applying collocation-specific syntactic constraints. For identification of PP-verb collocations these constraints are: preposition and noun must be constituents of a single PP, PP and verb must co-occur in a sentence which may consist of more than one clause. The latter constraint has been kept this weak, because automatic PP-attachment is highly unreliable when only information on part-of-speech and phrasal category is available. On the contrary, knowledge on collocability is envisaged to be employed for deciding on PP-attachment. Experiments using different syntactic constraints have shown that those PNV-combinations are most likely SVCs where the main verb is a past participle, and the preposition and the noun are constituents of the immediately preceding PP.

Secondly, morphosyntactic constraints are applied. Full forms of preposition and noun are used for constructing the PNV-triple, whereas the verb is reduced to a base form; thus accounting, on the one hand, for the morphosyntactic rigidity of PP-collocates, and on the other hand for the flexibility of verbal collocates.

Third, information on collocation-specific linguistic restrictions is not only utilized for constructing the candidate data, but is also used for selecting collocates from the candidate set. Two approaches have been pursued

1. a statistical one, where the entropy value of the PPs constituted by the preposition and the noun collocate is used to distinguish collocational from non-collocational PNV-combinations;
2. a lexical one, where words are used as lexical keys to identify classes of collocations, in particular, typical support-verbs are employed for selecting SVCs from the candidate data.

Another important result of the work is the insight that there exists no single best model for collocation identification.

The quality of the identification models is influenced by the following factors:

1. The linguistic constraints applied for selecting the candidate word combinations.
2. The ability of a statistical model employed to account for the significance of an individual word combination within a sample of research.
3. The feasibility of both statistical and linguistically-motivated strategies to model distinctive collocation-specific properties.

Moreover, the decision which strategies should be combined depends on the tasks to be pursued, and the applications of interest. If the identification component is intended as an automatic collocation learner, models maximizing identification accuracy will be employed. On the other hand, models leading to high recall are preferable when the identification results are hand-corrected. Precision, however, still needs to reach a certain level, as otherwise hand-correction would require too much effort.

Taking lexical keys into account is a general means for increasing identification accuracy. In the current work, it could be shown that high accuracy in identifying SVCs is achieved when typical support-verbs are used as lexical keys.

Identification accuracy also depends on the linguistic constraints applied during construction of the collocation candidates. Accuracy for identification of SVCs, as already stated, is best when the set of candidate word combinations comprises triples of preposition, noun and past participle, where noun and past participle need to be adjacent, and preposition and noun are constituents of the same PP.

Relaxation of the syntactic constraints applied leads to an increase of the total number of collocations covered by the candidate data. Nevertheless it is important that linguistic constraints are not completely abandoned, which would be the case if employing numerical spans, as it would lead to unnecessary increase of noise among the collocation candidates. All in all, there is a trade-off between recall and precision, insofar as the more constrained the selection of collocation candidates is with respect to certain collocation-type-specific properties, the higher is the precision of identification, but the lower is recall, because the broad range of collocations is syntactically flexible. When the only restriction for constructing the PNV-triples is that PP and verb co-occur in the same sentence, the number of true collocations among the data is much higher than in the set consisting of PNV-sequences. For example, there are two preferable strategies for maximizing SVC-identification accuracy with respect to syntactically less restricted sets.

> Either SVC-candidates are selected by means of the kwic-strategy from a small subset of highly recurrent PNV-combinations, or
>
> a subset of collocation candidates is identified from the initial set by employing the entropy model in combination with the kwic-model.

There are two preferable strategies for increasing SVC-recall by keeping the effort of hand-correction still feasible.

> The kwic-model is used for selecting SVCs from the set where co-occurrence frequency is greater than or equal to 3.

> A combination of relative entropy (I) and kwic-based selection is applied to the set where co-occurrence frequency is greater than or equal to 5.

In both cases, recall is approximately 76 %, precision is about 11 %. In the former case, some 3 000 word combinations need to be looked at, in the latter case the number is 2 000. While the disadvantage of the kwic-model is that only collocations containing typical support-verbs are selected, which speaks for applying statistical methods only, the drawback of statistical models is their lower accuracy. Considering purely statistical models, the best results with respect to recall and precision are achieved employing relative entropy (I), log-likelihood statistics (Lgl), or the entropy-model.

An architecture for the construction of the candidate data has been presented which makes candidate construction without hand-correction feasible. Thus any arbitrary text can be employed for collocation identification. This is important for statistics-based induction of lexical models for arbitrary domains, as well as for identifying appropriate material for developing and testing theories on lexicalization. As only a small percentage of the lexical material in a corpus[1] can be used for frequency-based or statistics-based collocation identification, large amounts of data need to be processed, thus collocation identification from manually annotated corpora such as Penn Treebank or Negra Corpus would be inappropriate, even if the treebanks become larger. If human annotated or corrected data on collocations are required, it is much more appropriate and time saving to work on collocation examples stored in a collocation database like the one developed in the work presented, because in this case only collocation relevant data are annotated. As collocation examples are linked to their position of occurrence in the original corpus arbitrary contexts are accessible for further annotation and manual correction.

7.2 Outlook

In the following a number of open questions with respect to the identification and description of collocations will be discussed. In addition, two strands of research which have evolved from the current work will be outlined briefly.

---

[1] 3 % of the PNV-combinations in the 8 million word newspaper corpus occur three times or more.

7.2.1 Collocation Identification

Statistics for High and Low Frequency Words

In linguistics, there is a well known dichotomy of high and low frequency words, i.e., a small set of unproductive but frequently occurring function words is opposed to a large set of productive but less frequently occurring content words. Considering the distribution of words and word combinations in text corpora, a dichotomy of low and high frequency occurrences can be found as well, but linguistic classification is less clearly tied to occurrence frequency. This is particularly the case with respect to collocations which on the one hand are frequent among highly recurrent word combinations, but which also occur among low frequency data. Collocation density, however, is high in sets of word combinations with high occurrence frequency and low in sets with low occurrence frequency.

On the statistics side, there are models that preferably select high frequency data, in our case I and Lgl, and there are measures that select for low frequency data, in our case MI and $Dice$. Thus high and low frequency data need to be examined separately unless statistical models can be found which work similarly well for both kinds of data, which is highly questionable. Due to little collocation density among low frequency data, collocation identification from this source is considerably hard, and the feasibility of a statistics-based account still awaits close examination.

An Account of "Commonness" of Word Combinations

Closely related to the previous discussion is the following assumption: collocations are distinguished from noncollocations by the native speaker because of their commonness, i.e., their acceptability and thus frequency within a certain communicative situation. This kind of information, however, cannot be counted in the corpora available, as there are neither corpora which are annotated with communicative situations, nor do tools exist which allow texts to be automatically annotated with it. Occurrence frequency is a highly provisional approximation to modeling commonness in corpora, because on the one hand no information on the situatedness of word combinations is accounted for, and on the other hand only high frequency word combinations are considered, whereas low frequency data in the corpus are left unaddressed. A means to cope with this situation is employing psycholinguistic acceptability tests for judging the commonness of a word combination. A first step in this direction has already been made by correlating lexical co-occurrence frequencies found in corpora with human acceptability ratings, cf. [Lapata *et al.*, 1999].

Relations between Recurrence Patterns and Distribution of Collocations in a Corpus

The major question here is whether it is possible to deduce the approximate distribution of particular collocations from the frequency distributions of all word combinations with related syntactic structure occurring in a certain corpus. Knowledge about typical distributions of collocations in corpora representing certain domains would guide the decision about which statistical models should be applied for collocation identification. To obtain such information, it would be necessary to investigate a number of corpora from various domains, including to a large extent manual inspection of the data in order to decide which word combinations are collocational.

Another open question is whether corpus size approximates a maximum above which the gain of new collocations is marginal. If such a saturation with collocations is the case, it is expected that the level of such a saturation differs between domains and between domain-specific and general language collocations, i.e., collocations which belong to the general lexicon as opposed to word combinations which are collocational only with respect to a certain domain or just a particular corpus as this is the case with pseudo-collocations.

It is important to note that answers to these questions are strongly influenced by the collections of texts constituting a corpus.

In-Depth Empirical Studies on the Differences between Models for Collocation Identification

The work presented has provided a range of evidence that the goodness of a particular model for collocation identification is influenced by the particular class of collocations to be identified, the threshold determining the minimum occurrence frequency required for a word combination to be part of the candidate sample, by the syntactic constraints employed for candidate selection from the extraction corpus, as well as by the extraction corpus itself. In order to obtain a clearer picture about the interrelation of these features, a variety of in-depth studies is required building upon the results from the experiments conducted in the thesis.

7.2.2 Additional Levels of Description

Collocations in the current work have been examined from a mainly syntax-based view. The reason has been that even though the co-occurrence of syntactic generativity and collocation-specific rigidity in collocations is apparent, a principled approach is still out of sight. A step towards an understanding of this kind of interrelation has been made in the work presented by specifying a representation scheme and implementing a database, both of which accounting for generative and static aspects in an integrative way combining competence-based syntactic description and real-world data in a large scale. This has become feasible, because of the availability of efficient tools for shallow syntactic processing and the existence of respective training corpora.

Semantic Tagging

Semantic tagging is another crucial step towards a theory of collocations. This is especially the case as it is assumed that collocations are a phenomenon of semantics and pragmatics, and particularities in syntactic structure are no more than a reflex of underlying semantics- and pragmatics-driven processes. Automation of semantic tagging is indispensable for large scale annotation. The ground for such a task is already set with the availability of semantic databases like WordNet[2], and preliminary studies on semantic taggers such as [Segond *et al.*, 1997].

Pragmatic Aspects of Collocations

Description at pragmatic level is necessary, in order to account for the commonness of a word combination; in particular, for investigating the pragmatic function of a collocation and the stylistic implications of its usage. The current database already contains some information of this kind, such as information on the origin of a particular collocation realization (cf. the attribute cr-source), and the encoding of Aktionsart and causativity at SVCs. With respect to the former, more data and an enlargement of the pool of corpora used for collocation identification is necessary. With respect to the latter, strategies for automating the assignment of Aktionsart and causativity need to be defined, and methods developed which enable systematic comparison of utterances where SVCs are used, and cases where verbal equivalents are employed.

7.2.3 Follow-up Projects

In the following, two projects will be outlined briefly which have emerged from the work on collocation identification. The one is a research project that aims at improving lexicalized stochastic parsing. The other one is a pilot study employing psycholinguistic acceptability tests for classifying PNV-collocations.

---

[2] See for instance `http://www.ilc.pi.cnr.it/EAGLES96/rep2/node20.html` for links to the Princeton WordNet and EuroWordNet.

Stochastic Lexical Models for Parse Pruning

Insights gained in the present work on developing models for automatic collocation identification will be utilized for learning lexical models that will then be applied for pruning in stochastic parsing. A lexical approach to structural disambiguation is expected to be particularly well suited to improve PP-attachment. While the methods developed in this study allow lexical models to be learned from arbitrary raw text, syntactic models are best learned from fully annotated, hand corrected treebanks.[3] As available treebanks[4] are far too small for inducing reliable lexical generalizations, syntactic and lexical models need to be trained from different sources. Thus a main task of the project is to combine the syntactic and the lexical model within a stochastic parser. Another advantage of separate training of syntactic and lexical model is that it allows the lexical model to be better adapted to the text domain which shall be parsed. Such adaptation would be desirable for the syntactic model as well, but would require unsupervised learning.

Psycholinguistically Motivated Classification of Collocations

Collocations can be described at various linguistic levels, such as syntactic structure constituted by the collocates, semantic interpretation(s) available, syntactic rigidity, semantic opacity, and pragmatic function. The problem of any such classification is that collocations tend to divide into prototypical cases and borderline cases of a class. Thus the distinction between collocations and noncollocations is controversial in the literature. Frequency-based approaches are also in many cases infeasible for a distinction. On the other hand, native speakers have good intuitions on the usage of collocations. This ability shall be employed in a controlled way for testing and grouping collocations by means of psycholinguistic acceptability tests conducted with a large number of subjects. A software[5] is employed which is particularly designed for running experiments over the world wide web. Thus experiments on a large scale become feasible, and moreover a large and heterogeneous pool of subjects can be accessed this way, which is crucial for studying collocability.

---

[3] Unsupervised learning is still less accurate than supervised learning.

[4] The Penn Treebank which is the reference treebank for English contains approximately 40 000 structurally annotated sentences, the Negra corpus which is currently the only publicly available treebank for German covers approximately 20 000 sentences.

[5] See `http://surf.to/experiments`.

Bibliography

[Bar-Hillel, 1955] Yehoshuah Bar-Hillel. Idioms. In W. N. Locks and A. D. Booth (eds.), *Machine translation of languages*, pages 47 – 55. MIT Press, 1955.

[Blochwitz, 1980] Werner Blochwitz. Zur Frage der semantischen Relationen zwischen Verb und verbaler Periphrase im Französischen in Konfrontation mit dem Deutschen. *Linguistische Studien*, 69/II:1 – 121, 1980.

[Bolinger, 1976] Dwight Bolinger. Meaning and Memory. *Forum linguisticum*, 1, 1976.

[Bortz, 1985] Jürgen Bortz. *Lehrbuch der Statistik. Für Sozialwissenschaftler. 2. Ausgabe.* Springer, 1985.

[Brants, 1996] Thorsten Brants. TnT – A Statistical Part-of-Speech Tagger. Technical Report, Universität des Saarlandes, Computational Linguistics, 1996.

[Brants, 1999] Thorsten Brants. *Tagging and Parsing with Cascaded Markov Models – Automation of Corpus Annotation.* Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 6. German Research Center for Artificial Intelligence and Saarland University, Saarbrücken, Germany, 1999.

[Breidt *et al.*, 1996] E. Breidt, F. Segond and G. Valetto. Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 1036 – 1040, Copenhagen, Denmark, 1996.

[Breidt, 1993] Elisabeth Breidt. Extraction of N-V-Collocations from Text Corpora: A Feasibility Study for German. In *Proceedings of the 1st ACL-Workshop on Very Large Corpora*, 1993.

[Bresnan, 1982] Joan Bresnan. *The mental representation of grammatical relations.* MIT Press, Cambridge, MA., 1982.

[Burger *et al.*, 1982] Harald Burger, Annelies Buhofer and Ambros Sialm. *Handbuch der Phraseologie.* de Gruyter, Berlin, New York, 1982.

[Bußmann, 1990] Hadumod Bußmann. *Lexikon der Sprachwissenschaft.* Kröner, 2nd edition, 1990.

[Carrol *et al.*, 1999] John Carrol, Guido Minnen and Ted Briscoe. Corpus Annotation for Parser Evaluation. In *Proceedings of the Workshop on Linguistically Interpreted Corpora*, Bergen, Norway, 1999.

[Christ, 1994] Oliver Christ. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94*, Budapest, 1994.

[Church and Hanks, 1989] K.W. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76 – 83, Vancouver, Canada, 1989.

[Col, 1996] Collins. Plain English Dictionary, 1996.

[Collins, 1997] Michael Collins. Three Generative, Lexicalized Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, 1997.

[Cover and Thomas, 1991] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 1991.

[Daille *et al.*, 1994] B. Daille, E. Gaussier and J.-M. Lange. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan, 1994.

[Date, 1995] C.J. Date. *An introduction to database systems.* (6th ed.) [World Student Series Edition]. Addison-Wesley, 1995.

[Docherty *et al.*, 1997] Vincent J. Docherty, Ulrich Heid and Judith Eckle-Kohler. Computational linguistic support for the revision of a commercial dictionary – corpus-based updates of entries and of collocation information. In *Beiträge zur 6. Fachtagung der Sektion Computerlinguistik DGfS-CL*, Heidelberg, Germany, October 1997.

[Drosdowski et al., 1989] Günther Drosdowski et al. *Duden Deutsches Universalwörterbuch. 2. Aufl.* Dudenverlag, 1989.

[Dufour, 1998] Nicolas Dufour. A database for computerized multi-word unit recognition. In *Proceedings of ISP-3*, Stuttgart, Germany, 1998.

[Dunning, 1993] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61 – 74, 1993.

[Eisner, 1996] Jason Eisner. Three New Probabilistic Models for Dependency Parsing: An Exploration. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 340 – 345, Copenhagen, Denmark, 1996.

[Fano, 1961] R. Fano. *Transmission of Information.* MIT Press, Cambridge, Massachusetts, 1961.

[Firth, 1957] J. R. Firth. *Papers in Linguistics 1934 - 1951.* Oxford University Press, London, 1957.

[Fleischer, 1982] Wolfgang Fleischer. *Phraseologie der Deutschen Gegenwartssprache.* VEB Bibliographisches Institut Leipzig, Germany, 1982.

[Flickinger *et al.*, 1998] Daniel P. Flickinger, Ivan A. Sag and Ann Copestake. A grammar of English in HPSG. Design and Implementation. 1998.

[Frantzi and Ananiadou, 1996] Katerina T. Frantzi and Sophia Ananiadou. Extracting Nested Collocations. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 41 – 46, Copenhagen, Denmark, 1996.

[Grefenstette and Teufel, 1995] Gregory Grefenstette and Simone Teufel. A corpus-based method for Automatic Identification of Support Verbs for Nominalisations. In *Proceedings of the EACL*, Dublin, Ireland, 1995.

[Günther and Pape, 1976] Heide Günther and Sabine Pape. Funktionsverbgefüge als Problem der Beschreibung komplexer Verben in der Valenztheorie. In Helmut Schumacher (ed.), *Untersuchungen zur Verbvalenz. Forschungsberichte des Instituts für deutsche Sprache*, pages 92 – 128. Tübingen, 1976.

[Haruno *et al.*, 1996] Masahiko Haruno, Satoru Ikehara and Takefumi Yamazaki. Learning Bilingual Collocations by Word-Level Sorting. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 525 – 530, Copenhagen, Denmark, 1996.

[Healey, 1968] Alan Healey. English idioms. *Kivung*, 1(2):71 – 108, 1968.

[Helbig and Buscha, 1980] Gerhard Helbig and Joachim Buscha. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.* VEB Verlag Enziklopädie, Leipzig, 1980.

[Helbig, 1979] Gerhard Helbig. Probleme der Beschreibung von Funktionsverbgefügen im Deutschen. *Deutsch als Fremdsprache*, (16):273 – 285, 1979.

[Heringer, 1968] Hans Jürgen Heringer. *Die Opposition von "kommen" und "bringen" als Funktionsverben.* Pädagogischer Verlag Schwann, 1968.

[Herrlitz, 1973] Wolfgang Herrlitz. Funktionsverbgefüge vom Typ "in Erfahrung bringen". Ein Beitrag zur generativ-transformationellen Grammatik des Deutschen. *Linguistische Arbeiten*, (1), 1973.

[Hoberg, 1981] Ursula Hoberg. *Die Wortstellung in der geschriebenen deutschen Gegenwartssprache.* Heutiges Deutsch. Bd. 10. Linguisitische Grundagen. Forschungen des Instituts für deutsche Sprache. Max Huber Verlag, München, 1981.

[Hockett, 1958] Charles F. Hockett. *A course in modern linguistics.* MacMillan, 1958.

[Ikehara *et al.*, 1996] Saturo Ikehara, Satoshi Shirai and Hajime Uchino. A Statistical Method for Extracting Uninterrupted and Interrupted Collocations from Very Large Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 41 – 46, Copenhagen, Denmark, 1996.

[Joshi and Schabes, 1991] Aravind Joshi and Yves Schabes. Tree-adjoining Grammars and Lexicalized Grammar. In M. Nivat and A. Podelski (eds.), *Definability and Recognizability of Sets of Trees.* Elsevier, 1991.

[Katz and Postal, 1963] Jerrold J. Katz and Paul M. Postal. Semantic interpretation of idioms and sentences containing them. Technical Report, MIT Research Laboratory of Electronics, Quarterly Progress Report No. 70, 1963.

[Keil, 1997] Martina Keil. *Wort für Wort - Repräsentation und Verarbeitung verbaler Phraseologismen.* [Sprache + Information.] Niemeyer, Tübingen, 1997.

[Keller *et al.*, 1999] Frank Keller, Martin Corley, Steffan Corley, Matthew Crocker and Shari Trewin. Gsearch: A Tool for Syntactic Investigation of Unparsed Corpora. In *Proceedings of LINC-99. Linguistically Interpreted Corpora*, Bergen, Norway, 1999.

[Kim and Cho, 1993] P.K. Kim and Y.K. Cho. Indexing Compound Words from Korean Text Using Mutual Information. In *Proceedings of NLPRS*, pages 85 – 92, 1993.

[Krenn and Erbach, 1993] Brigitte Krenn and Gregor Erbach. Idioms and Support Verb Constructions. In J. Nerbonne, K. Netter and C. Pollard (eds.), *German Grammar in HPSG*, CLSI Lecture Notes. 1993.

[Krenn and Samuelsson, 1996] Brigitte Krenn and Christer Samuelsson. A Linguist's Guide to Statistics. Technical Report, University of the Saarland, Dept. of Computational Linguistics, 1996.

[Krenn and Volk, 1993] Brigitte Krenn and Martin Volk. DiTo-Datenbank. Datendokumentation zu Funktionsverbgefügen und Relativsätzen. Technical Report, Dokument D-93-24, Deutsches Forschungszentrum für Künstliche Intelligenz, 1993.

[Lakoff and Johnson, 1981] George Lakoff and Mark Johnson. *Metaphors we live by.* University of Chicago Press, Chicago, London, 1981.

[Lapata *et al.*, 1999] Maria Lapata, Scott McDonald and Frank Keller. Determinants of Adjective-Noun Plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 1999.

[Magerman and Marcus, 1990] David M. Magerman and Mitchell P. Marcus. Parsing a Natural Language Using Mutual Information Statistics. In *AAAI*, pages 984 – 989, Boston, MA, 1990.

[Makkai, 1972] Adam Makkai. *Idiom Structure in English.* Mouton, 1972.

[Manning and Schütze, 1999] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA., 1999.

[Melamed, 1997] I. Dan Melamed. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*, Providence, RI, 1997.

[Mesli, 1989] Nadia Mesli. *Les Locutions Verbales dans les Ecrits de M. Luther.* PhD thesis, Université d'Aix-Marseille I, 1989.

[Mesli, 1991] Nadia Mesli. Funktionsverbgefüge in der maschinellen Analyse und Übersetzung: linguistische Beschreibung und Implementierung im CAT2-Formalismus. Technical Report, Eurotra-D Working Papers, No. 19, 1991.

[Morrill, 1994] Glyn Morrill. *Type Logical Grammar: Categorial Logic of Signs.* Kluwer Academic Publishers, 1994.

[Müller, 1999] Stefan Müller. *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche.* Linguistische Arbeiten, Nr. 397, Niemeyer, Tübingen, 1999.

[Netter et al., 1998] Klaus Netter et al. DiET – diagnostic and evaluation tools for natural language processing applications. In *Proceedings of the first International Conference on Language Resources & Evaluation*, Granada, Spain, 1998.

[Nunberg *et al.*, 1994] Geoffrey Nunberg, Ivan Sag and Tom Wasow. Idioms. *Language*, 70:491 – 538, 1994.

[Oepen *et al.*, 1998] Stephan Oepen, Klaus Netter and Judith Klein. TSNLP — Test Suites for Natural Language Processing. In John Nerbonne (ed.), *Linguistic Databases*, CSLI Lecture Notes 77. Center for the Study of Language and Information, 1998.

[Persson, 1975] Ingemar Persson. *Das System der kausativen Funktionsverbgefüge.* Liber, Malmö, 1975.

[Polenz, 1963] Peter von Polenz. Funktionsverben im heutigen Deutsch. Sprache in der rationalisierten Welt. *Wirkendes Wort. Beiheft*, (5), 1963.

[Pollard and Sag, 1994] C. Pollard and I. Sag. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago, 1994.

[Rabiner, 1989] Lawrence R. Rabiner. Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE 77(2)*, pages 257–295, 1989.

[Riehemann, 1997] Susanne Riehemann. Idiomatic Constructions in HPSG. Paper presented at the 1997 HPSG conference, Ithaca, 1997.

[Segond and Tapanainen, 1995] Frédérique Segond and Pasi Tapanainen. Using a finite-state based formalism to identify and generate multiword expressions. Technical Report, Technical Report MLTT-019 Rank Xerox Research Centre, Grenoble, 1995.

[Segond *et al.*, 1997] Frédérique Segond, Anne Schiller, Gregory Grefenstette and Jean-Pierre Chanod. An Experiment in Semantic Tagging using Hidden Markov Model Tagging. In *ACL'97 Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain, 1997.

[Shimohata *et al.*, 1997] Sayori Shimohata, Toshiyuki Sugio and Nunji Nagata. Retrieving Collocations by Co-Occurrences and Word Order Constraints. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, 1997.

[Siegel, 1956] Sidney Siegel. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill Kogakusha Ltd., Tokyo, 1956.

[Skut and Brants, 1998] Wojciech Skut and Thorsten Brants. Chunk Tagger. Stochastic Recognition of Noun Phrases. In *ESSLI Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrücken, Germany, August 1998.

[Skut *et al.*, 1997] Wojciech Skut, Brigitte Krenn, Thorsten Brants and Hans Uszkoreit. An Annotation Scheme for Free Word Order Languages. In *Proceedings of ANLP-97*, Washington, 1997.

[Skut *et al.*, 1998] Wojciech Skut, Thorsten Brants, Brigitte Krenn and Hans Uszkoreit. A Linguistically Interpreted Corpus of German Newspaper Text. In *Proceedings of the first International Conference on Language Resources & Evaluation*, Granada, Spain, 1998.

[Skut, forthcoming] Wojciech Skut. *Partial Parsing for Corpus Annotation and Text Processing.* Saarbrücken Dissertations in Computational Linguistics and Language Technology. German Research Center for Artificial Intelligence and Saarland University, Saarbrücken, Germany, forthcoming.

[Smadja *et al.*, 1996] Frank Smadja, Kathleen R. McKeown and Vasileios Hatzivassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1):3 – 38, 1996.

[Smadja, 1993] Frank Smadja. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19:143 – 177, 1993.

[Thielen and Schiller, 1995] Christine Thielen and Anne Schiller. Ein kleines und erweitertes Tagset fürs Deutsche. In *Tagungsberichte des Arbeitstreffens Lexikon + Text 17./18. Februar 1994, Schloß Hohentübingen. Lexicographica Series Maior*, Tübingen, 1995. Niemeyer.

[Tschichold and Hacken, 1998] Cornelia Tschichold and Pius Ten Hacken. English Phraseology in Word Manager. In *Proceedings of ISP-3*, Stuttgart, Germany, 1998.

[Tschichold, 1997] Cornelia Tschichold. English idioms in a computational lexicon. In *Beiträge zur 6. Fachtagung der Sektion Computerlinguistik DGfS-CL*, Heidelberg, Germany, October 1997.

[Uszkoreit *et al.*, 1994] Hans Uszkoreit, Rolf Backofen, Stephan Busemann, Abdel Kader Diagne, Elizabeth A. Hinkelman, Walter Kasper, Bernd Kiefer, Hans-Ulrich Krieger, Klaus Netter, Günter Neumann, Stephan Oepen and Stephen P. Spackman. DISCO — An HPSG-based NLP System and its Application for Appointment Scheduling. In *Proceedings COLING 1994*, Kyoto, 1994.

[van der Linden, 1993] Jan-Erik van der Linden. *A Categorial, Computational Theory of Idioms.* OTS-publications, Dissertation Series, 1993.

[Wall *et al.*, 1996] Larry Wall, Tom Christiansen and Randal L. Schwartz. *Programming Perl.* O'Reilly, 1996.

[Yuan, 1986] Jie Yuan. Funktionsverbgefüge im heutigen Deutsch. Eine Analyse und Kontrastierung mit ihren chinesischen Entsprechungen. Sammlung Groos 28, 1986.