LEXICAL ASSOCIATION MEASURES
Collocation Extraction

Pavel Pecina

ÚFAL
ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

STUDIES IN COMPUTATIONAL AND THEORETICAL LINGUISTICS

Pavel Pecina

LEXICAL ASSOCIATION MEASURES
Collocation Extraction

Published by Institute of Formal and Applied Linguistics
as the 4th publication in the series
Studies in Computational and Theoretical Linguistics.

Reviewers: Timothy Baldwin
Jiří Semecký

ISBN 978-80-904175-5-7

to my family

Contents

Motto:

"You shall know a word by the company it keeps!"
— John Rupert Firth 1890–1960

Acknowledgements

This work would not have succeeded without the support of many exceptional people who deserve my special thanks: my advisor Jan Hajič, for his guidance and support during my study at the *Institute of Formal and Applied Linguistics*; Bill Byrne for hosting me at the C*enter for Language and Speech Processing* and other colleagues and friends from the *Johns Hopkins University*, namely Jason Eisner, Erin Fitzgerald, Arnab Goshal, Frederick Jelinek, Sanjeev Khudanpur, Shankar Kumar, Veera Venkatramani, Paola Virga, Peng Xu, and David Yarowsky; my mentor Chris Quirk at *Microsoft Research* and other members of the Natural Language Processing group for the opportunity to work with them, namely Bill Dolan, Arul Menezes, and Lucy Vanderwende; my colleagues from the *University of Maryland* and *University of West Bohemia* who participated in the Malach project, namely Pavel Ircing, Craig Murray, Douglas Oard, Josef Psutka, Dagobert Soergel, Jianqiang Wang, and Ryen White; my colleagues and friends from the *Institute of Formal and Applied Linguistics*, especially those who contributed to my research: Silvie Cinková, Pavel Češka, Petra Hoffmannová, Martin Holub, Petr Homola, Vladislav Kuboň, Michal Marek, Petr Podveský, Pavel Schlesinger, Miroslav Spousta, Drahomíra Spoustová, Jana Straková, and Pavel Straňák; my loving wife Eliška, my dear parents Pavel and Hana, and the whole of my family for their immense patience, support, and love.

The work was also supported by the Ministry of Education of the Czech Republic, project MSM 0021620838.

Pavel Pecina

1
Introduction

Word association is a popular word game based on exchanging words that are in some way associated together. The game is initialized by a randomly or arbitrarily chosen word. A player then finds another word associated with the initial one, usually the first word that comes to his or her mind, and writes it down. A next player does the same with this word and the game continues in turns until a time or word limit is met. The amusement of the game comes from the analysis of the resulting chain of words – how far one can get from the initial word and what the logic behind the individual associations is. An example of a possible run of the game might be this word sequence: *dog*, *cat*, *meow*, *woof*, *bark*, *tree*, *plant*, *green*, *grass*, *weed*, *smoke*, *cigarette*, *lighter*, *fluid*.[1]

Similar concepts are commonly used in *psychology* to study a subconscious mind based on subject's word associations and disassociations, and in *psycholinguistics* to study the way knowledge is structured in the human mind, e.g. by *word association norms* measured as subject's responses to words when preceded by associated words (Palermo and Jenkins, 1964). "Generally speaking, subjects respond quicker than normal to the word *nurse* if it follows a highly associated word such as *doctor*" (Church and Hanks, 1990).

1.1 Lexical association

Our interest in word association is *linguistic* and hence, we use the term **lexical association** to refer to *association between words*. In general, we distinguish between three types of association between words: **collocational association** restricting combination of words into phrases (e.g. *crystal clear*, *cosmetic surgery*, *weapons of mass destruction*), **semantic association** reflecting semantic relationship between words (e.g. *sick – ill*, *baby – infant*, *dog – cat*), and **cross-language association** corresponding to potential translations of words between different languages (e.g. *maison (FR) – house (EN)*, *baum (GER) – tree (EN)*, *květina (CZ) – flower (EN)*).

In the word association game and the fields mentioned above, it is a human mind what directly provides evidence for exploring word associations. In this work, our source of such evidence is a **corpus** – a collection of texts containing examples of word usages. Based on such data and its statistical interpretation, we attempt to estimate lexical associations automatically by means of **lexical association measures** determin-

[1] examples from http://www.wordassociation.org/

ing the strength of association between two or more words based on their occurrences and cooccurrences in a corpus. Although our study is focused on the association on the collocational level only, most of these measures can be easily used to explore also other types of lexical association.

1.1.1 Collocational association

The process of combining words into phrases and sentences of natural language is governed by a complex system of rules and constraints. In general, basic rules are given by *syntax*, however there are also other restrictions (semantic and pragmatic) that must be adhered to in order to produce correct, meaningful, and fluent utterances. These constrains form important linguistic and lexicographic phenomena generally denoted by the term **collocation**. Collocations range from lexically restricted expressions (*strong tea*, *broad daylight*), phrasal verbs (*switch off*, *look after*), technical terms (*car oil*, *stock owl*), and proper names (*New York*, *Old Town*) to idioms (*kick the bucket*, *hear through the grapevine*), etc. As opposed to free word combinations, collocations are not entirely predictable only on the basis of syntactic rules. They should be listed in a **lexicon** and learned the same way as single words are.

Components of collocations are involved in a syntactic relation and usually tend to cooccur (in this relation) more often than would be expected in other cases. This empirical aspect typically distinguishes collocations from free word combinations. Collocations are often characterized by semantic **non-compositionality** – when the exact meaning of a collocation cannot be (fully) inferred from the meaning of its components (*kick the bucket*), syntactic **non-modifiability** – when their syntactic structure cannot be freely modified, e.g. by changing the word order, inserting another word, or changing morphological categories (*poor as a church mouse* vs. **poor as a big church mouse*), and lexical **non-substitutability** – when collocation components cannot be substituted by synonyms or other related words (*stiff breeze* vs. **stiff wind*) (Manning and Schütze, 1999, Chapter 5). Another property of some collocations is their **translatability** into other languages: a translation of a collocation cannot generally be performed blindly, word by word (e.g. the two-word collocation *ice cream* in English should be translated into Czech as one word *zmrzlina*, or perhaps as *zmrzlinový krém* (rarely) but not as *ledový krém* which would be a straightforward word-by-word translation).

1.1.2 Semantic association

Semantic association requires no grammatical boundedness between words. This type of association is concerned with words that are used in similar contexts and domains – word pairs whose meanings are in some kind of semantic relation. Compiled information of such type is usually presented in the form of a **thesaurus** and includes the following types of relationships: **synonyms** with exactly or nearly equivalent meaning (*car – automobile, glasses – spectacles*), **antonyms** with the opposite meaning (*high – low, love – hate*), **meronyms** with the part-whole relationship (*door – house, page –book*), **hyperonyms** based on superordination (*building – house, tree – oak*), **hyponyms** based on subordination (*lily – flower, car – machine*), and perhaps other word combinations with even looser relations (*table – chair, lecture – teach*).

Semantic association is closest to the process involved in the word game mentioned in the beginning of this chapter. Although presented as a relation between words themselves, the actual association exists between their meanings (concepts). Before a word association emerges in the human mind, the initial word is semantically disambiguated and only one selected sense of the word participates in the association, e.g. the word *bark* has different meaning in association with *woof* and *tree*. For the same reason, semantic association exists not only between single words but also between multiword expressions constituting indivisible semantic units (i.e. collocations).

Similarly to collocational association, semantically associated words cooccur in the same context more often than others, but in this case the context is understood as a much wider span of words and, as we have already mentioned, no direct syntactic relation between the words is necessary.

1.1.3 Cross-language association

Cross-language association corresponds to possible translations of words in one language to another. This information is usually presented in a form of a bilingual **dictionary**, where each word (with all its senses) is provided with all its equivalents in the other language. Although every word (in one of its meanings) usually has one or two common and generally accepted translations sufficient to understand its meaning, it can be potentially expressed by a larger number of (more or less equivalent but in a certain context entirely adequate) options. For example, the Czech adjective *důležitý* is in most dictionaries translated into English as *important* or *significant*, but in a text it can be translated also as: *considerable, material, momentous, high, heavy, relevant, solid, live, substantial, serious, notable, pompous, responsible, consequential, gutty, great, grand, big, major, solemn, guttily, fateful, grave, weighty, vital, fundamental,*[2] and possibly also as other options depending on context. Not even a highly competent speaker of both languages could not be expected to enumerate them exhaustively. Similarly to the case of semantic association, dictionary items are not only single words but also multiword expressions which cannot be translated in a word-by-word manner (i.e. collocations).

Cross-language association can be acquired not only from the human mind, it can also be extracted from examples of already realized translations, e.g. in the form of **parallel texts** – where texts (sentences) are placed alongside their translations. Also in such data, associated word pairs (translation equivalents) cooccur more often that would be expected in the case of non-associated (random) pairs.

[2] translations from http://slovnik.seznam.cz/

1.2 Motivation and applications

A monolingual **lexicon** enriched by collocations, a **thesaurus** comprised of semantically related words, and a bilingual **dictionary** containing translation equivalents – all of these are important (and mutually interlinked) resources not only for *language teaching* but in a machine-readable form also for many tasks of *computational linguistics* and *natural language processing*.

The traditional **manual approaches** to building these resources are in many ways insufficient (especially for computational use). The major problem is their lack of exhaustiveness and completeness. They are only "snapshots of a language".[3] Although modern lexicons, dictionaries, and thesauri are developed with the help of language corpora, utilization of these corpora is usually quite shallow and reduced to analysis of the most frequent and typical (multi)word usages. Natural language is a live system and no such resource can perhaps ever be expected to be complete and fully reflect the actual language use. All these resources must also deal with the problem of domain specificity. Either, they are general, domain-independent and thus in special domains usable only to a certain extent, or they are specialized, domain-specific and exist only for certain areas. Considerable limitations lie in the fact that the manually built resources are discrete in character, while lexical association, as presented in this work, should be perceived as a continuous phenomenon. Manually built language resources are usually reliable and contain only a small number of errors and mistakes. However, their development is an expensive and time-consuming process.

Automatic approaches extract association information on the basis of statistical interpretation of corpus evidence (by means of lexical association measures). They should eliminate (to a certain extent) all the mentioned disadvantages (lack of exhaustiveness and completeness, domain-specificity, continuousness). However, they heavily rely on the quality and extent of the source corpora the associations are extracted from. Compared to manually built resources, the automatically built ones will contain certain errors and this fact must be taken into account when these resources are applied. In the following passages, we present some of the tasks that make use of such automatically built resources.

Applications of lexical association measures

Generally, **collocation extraction** is the most popular application of lexical association measures and quite a lot of significant studies have been published on this topic, (e.g. Dunning, 1993; Smadja, 1993; Pedersen, 1996; Krenn, 2000; Weeber et al., 2000; Schone and Jurafsky, 2001; Pearce, 2002; Bartsch, 2004; Evert, 2004). In **computational lexicography**, automatic identification of collocations is employed to help human lexicographers in compiling lexicographic information (identification of possible word senses, lexical preferences, usage examples, etc.) for traditional lexicons (Church and

[3] A quote by Yorick Wilks, LREC 2008, Marrakech, Morocco.

Hanks, 1990) or for special lexicons of idioms or collocations (Klégr et al., 2005; Čermák et al., 2004), used e.g. in translation studies (Fontenelle, 1994a), bilingual dictionaries, or for language teaching (Smadja et al., 1996; Haruno et al., 1996; Tiedemann, 1997; Kita and Ogata, 1997; Baddorf and Evens, 1998). Collocations play an important role in systems of **natural language generation** where lexicons of collocations and frequent phrases are used during the process of word selection in order to enhance fluency of the automatically generated text (Smadja and McKeown, 1990; Smadja, 1993; Stone and Doran, 1996; Edmonds, 1997; Inkpen and Hirst, 2002).

In the area of **word sense disambiguation**, two applicable principles have been described: First, a word with a certain meaning tends to cooccur with different words than when it is used in another sense, e.g. *bank* as a financial institution occurs in context with words like *money*, *loan*, *interest*, etc., while *bank* as land along the side of a river or lake occurs with words like *river*, *lake*, *water*, etc. (Justeson and Katz, 1995; Resnik, 1997; Pedersen, 2001; Rapp, 2004). Second, according to Yarowsky's "one sense per collocation" hypothesis, all occurrences of a word in the same collocation have the same meaning (Yarowsky, 1995), e.g. the sense of the word *river* in the collocation *river bank* is the same across all its occurrences. There has also been some research on unsupervised discovery of word senses from text (Pantel and Lin, 2002; Tamir and Rapp, 2003). Association measures are used also for **detecting semantic similarity** between words, either on a general level (Biemann et al., 2004) or with a focus to specific relationships, such as synonymy (Terra and Clarke, 2003) or antonymy (Justeson and Katz, 1991).

An important application of collocations is in the field of **machine translation**. Collocations often cannot be translated in a word-by-word fashion. In translation, they should be treated rather as lexical units distinct from syntactically and semantically regular expressions. In this environment, association measures are employed in the **identification of translation equivalents** from sentence-aligned parallel corpora (Church and Gale, 1991; Smadja et al., 1996; Melamed, 2000) and also from non--parallel corpora (Rapp, 1999; Tanaka and Matsuo, 1999). In **statistical machine translation**, association measures are used over sentence aligned, parallel corpora to perform **bilingual word alignment** to identify translation pairs of words and phrases (or more complex structures) stored in the form of translation tables and used for constructing possible translation hypotheses (Mihalcea and Pedersen, 2003; Taskar et al., 2005; Moore et al., 2006).

Application of collocations in **information retrieval** has been studied as a natural extension of indexing single word terms to multiword units (phrases). Early studies were focused on small domain-specific collections (Lesk, 1969; Fagan, 1987, 1989) and yielded inconsistent and minor performance improvement. Later, similar techniques were applied over larger, more diverse collections within the Text Retrieval Conference (TREC) but still with only minor success (Evans and Zhai, 1996; Mittendorf et al., 2000; Khoo et al., 2001). Other studies were only motivated by information retrieval with no actual application presented (Dias et al., 2000). Recently, some researchers have attempted to incorporate cooccurrence information in probabilistic models (Vechtomova, 2001) but no consistent improvement in performance has been demonstrated (Alvarez et al., 2004; Jiang et al., 2004). Despite these results, using collocations in information retrieval is still of relatively high interest (e.g. Arazy and Woo, 2007). Collocational phrases have also been employed also in **cross-lingual information retrieval** (Ballesteros and Croft, 1996; Hull and Grefenstette, 1996). A significant amount of work has been done in the area of **identification of technical terminology** (Ananiadou, 1994; Justeson and Katz, 1995; Fung et al., 1996; Maynard and Ananiadou, 1999) and its translation (Dagan and Church, 1994; Fung and McKeown, 1997).

Lexical association measures have been applied to various other tasks from which we select the following examples: named entity recognition (Lin, 1998), syntactic constituent boundary detection (Magerman and Marcus, 1990), syntactic parsing (Church et al., 1991; Alshawi and Carter, 1994), syntactic disambiguation (Basili et al., 1993), discourse categorization (Wiebe and McKeever, 1998), adapted language modeling (Beefermam et al., 1997), extraction of Japanese-English morpheme pairs from bilingual terminological corpora (Tsuji and Kageura, 2001), sentence boundary detection (Kiss and Strunk, 2002b), identification of abbreviations (Kiss and Strunk, 2002a), computation of word associations norms (Rapp, 2002), topic segmentation and link detection (Ferret, 2002), discovering morphologically related words based on semantic similarity (Baroni et al., 2002), and possibly others.

1.3 Goals and objectives

This work is devoted to lexical association measures and their application to collocation extraction. The importance of this research was demonstrated in the previous section by the large range of applications in natural language processing and computational linguistics where the role of lexical association measures in general, or collocation extraction in particular, is essential. This significance was emphasized already in 1964 at the *Symposium on Statistical Association Methods For Mechanized Documentation* (Stevens et al., 1965), where Giuliano advocated better understanding of the measures and their empirical evaluation (as cited by Evert, 2004, p. 19):

> [First,] it soon becomes evident [to the reader] that at least a dozen somewhat different procedures and formulae for association are suggested [in the book]. One suspects that each has its own possible merits and disadvantages, but the line between the profound and the trivial often appears blurred. One thing which is badly needed is a better understanding of the boundary conditions under which the various techniques are applicable and the expected gains to be achieved through using one or the other of them. This advance would primarily be one in theory, not in abstract statistical theory but in a problem-oriented branch of statistical theory. (Giuliano, 1965, p. 259)

> [Secondly,] it is clear that carefully controlled experiments to evaluate the efficacy and usefulness of the statistical association techniques have not yet been undertaken except in a few isolated instances …Nonetheless, it is my feeling that the time is now ripe to conduct carefully controlled experiments of an evaluative nature, …(Giuliano, 1965, p. 259).

Since that time, the issue of lexical association has attracted many researchers and a number of works have been published in this field. Among those related to collocation extraction, we point out especially: Chapter 5 in Manning and Schütze (1999), Chapter 15 by McKeown and Radev in Dale et al. (2000), theses of Krenn (2000), Vechtomova (2001), Bartsch (2004), Evert (2004), and Moirón (2005). This work enriches the current state of the art in this field by achieving the following specific goals:

1) Compilation of a comprehensive inventory of lexical association measures

The range of various association measures proposed to estimate lexical association based on corpus evidence is enormous. They originate mostly in mathematical statistics, but also in other (both theoretical and applied) fields. Most of them were targeted mainly for collocation extraction, (e.g. Church and Hanks, 1990; Dunning, 1993; Smadja, 1993; Pedersen, 1996). The early publications were devoted to individual association measures, their formal and practical properties, and to the analysis of their application to a corpus. The first overview text appeared in Manning and Schütze (1999, Chapter 5) and described the three most popular association measures (and also other techniques for collocation extraction). Later, other authors (e.g. Weeber et al., 2000; Schone and Jurafsky, 2001; Pearce, 2002) attempted to describe (and compare) multiple measures. However, none of the authors, at the time our research started, had aspired to compile a comprehensive inventory of such measures.

A significant contribution in this direction was made by Stephan Evert, who set up a web page to "provide a repository for the large number of association measures that have been suggested in the literature, together with a short discussion of their mathematical background and key references"[4]. His effort, however, has focused only on measures applied to 2-by-2 contingency tables representing cooccurrence frequencies of word pairs, see details in Evert (2004). Our goal in this work is to provide a more comprehensive list of measures without this restriction. Such measures should be applicable to determine various types of lexical association but our key application and main research interest are in collocation extraction. The theoretical background to the concept of collocation and principles of collocation extraction from text corpora are covered in Chapter 2, and the inventory of lexical association measures is presented in Chapter 3.

[4] http://www.collocations.de/

2) Acquisition of reference data for collocation extraction

Before this work began, no widely acceptable evaluation resources for collocation extraction were available. In order to evaluate our own experiments, we were compelled to develop appropriate *gold-standard* reference data sets on our own. This comprised several important steps: to specify the task precisely, select a suitable source corpus, decide how to extract collocation candidates, define annotation guidelines, perform annotation by multiple subjects, and combine their judgments. The entire process and details of the acquired reference data sets are discussed in Chapter 4.

3) Empirical evaluation of association measures for collocation extraction

A strong request for empirical evaluation of association measures in specific tasks was made already by Giuliano in 1965. Later, other authors also emphasized the importance of such evaluation in order to determine "efficacy and usefulness" of different measures in different tasks and suggested various evaluation schemes for comparative evaluation of collocation extraction methods, e.g. Kita et al. (1994) or Evert and Krenn (2001). Empirical evaluation studies were published e.g. by Pearce (2002) and Thanopoulos et al. (2002). A comprehensive study of statistical aspects of word cooccurrences can be found in Krenn (2000) or Evert (2004).

Our evaluation scheme should be based on *ranking*, not classification (identification), and it should reflect the ability of association measure to rank potential collocations according to their chance to form true collocations (judged by human annotators). Special attention should be paid to statistical significance tests of the evaluation results. Our experiments, their results, and comparison are described in Chapter 5.

4) Combination of association measures for collocation extraction

The main focus of this work lies in the investigation of the possibility for combining association measures into more complex models in order to improve performance in collocation extraction. Our approach is based both on the application of supervised machine learning techniques and the fact that different measures discover different collocations. This novel insight into the application of association measures for collocation extraction is explored in Chapter 6.

Notes

In this work, no special attention is paid to semantic and cross-language association as they were discussed earlier in this chapter. We focus entirely on collocational association and the study of methods for automatic collocation extraction from text corpora. However, the inventory of association measures presented in this work, the evaluation scheme, as well as the principle of combining association measures can be easily adapted and used for other types of lexical association. As can be judged from the volume of published works in this field, collocation extraction has really been the most popular application of lexical association measures. The high interest in this field is also expressed in the activities of the ACL Special Interest Group on the Lexicon (SIGLEX) and the long tradition of workshops focused on problems related to this field.[5]

Our attention is restricted exclusively to two-word (*bigram*) collocations – primarily for the limited scalability of some methods to higher-order n-grams and also for the reason that experiments with longer expressions would require processing of a substantially larger corpus to obtain enough evidence of the observed events. For example, the Prague Dependency Treebank (see Chapter 4) contains approximately 623 000 different dependency bigrams – only about 27 000 of them occur with frequency greater than five, which can be considered sufficient evidence for our purposes. The same data contains more than twice as many trigrams (1 715 000), but only half the number (14 000) occurring more than five times.

The methods proposed in our work are language independent, although some language-specific tools are required for linguistic preprocessing of source corpora (e.g. part-of-speech taggers, lemmatizers, and syntactic parsers). However, the evaluation results are certainly language dependent and cannot be easily generalized for other languages. Mainly due to source constraints, we perform our experiments only on a limited selection of languages: Czech, Swedish, and German.

Some preliminary results of this research have already been published (see Pecina, 2005; Pecina and Schlesinger, 2006; Cinková et al., 2006; Pecina, 2008a,b).

[5] ACL 2001 Workshop on Collocations, Toulouse, France; 2002 Workshop on Computational Approaches to Collocations, Vienna, Austria; ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan; ACL 2004 Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain; COLING/ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, Australia; EACL 2006 Workshop on Multiword expressions in a multilingual context, Trento, Italy; 2006 Workshop on Collocations and idioms: linguistic, computational, and psycholinguistic perspectives, Berlin, Germany; ACL 2007 Workshop on a Broader Perspective on Multiword Expressions, Prague, Czech Republic; LREC 2008 Workshop, Towards a Shared Task for Multiword Expressions, Marrakech, Morocco.

2
Theory and Principles

This chapter is devoted to the theoretical background to collocations and principles of collocation extraction from text corpora. First, we present the notion of collocation based on the work of F. Čermák who introduced this concept into Czech linguistics (Čermák and Holub, 1982). This part is followed by an overview of various other approaches to this phenomenon presented from the perspective of theoretical and also applied linguistics. In the second half of the chapter, we describe details of the process of collocation extraction employed in the experimental part of this work.

2.1 Notion of collocation

The term **collocation** is derived from the Latin *collorale* (to place side by side, to co--locate). In linguistics, it is usually related to co-location of **words** and the fact that they can not be combined freely and randomly only by rules of grammar. Collocations are a borderline phenomenon ranging between lexicon and grammar and as such it is quite difficult to define and treat systematically. This section is intended to illustrate the diverse notions of collocation advocated by various authors in the last 70 years.

2.1.1 Lexical combinatorics

Although in traditional linguistics, *lexis* (vocabulary) and *grammar* (morphology and syntax) were perceived as separate and distinct components of a natural language, they are nowadays considered inseparable and completely interdependent. **Syntactic rules** are not the only restrictions imposed on arranging words into meaningful expressions and sentences. Čermák (2006) emphasizes that **semantic rules** are those which primarily govern the combination of words. These rules determine semantic **compatibility**, i.e. whether a lexical combination is meaningful or not (or to what extent), which combinations are (proto)typical and most frequent, which are common and ordinary, marginal and abnormal, or which are impossible. Syntax then plays only a subordinate role in the process of lexical selection. Omitting the semantic rules generally leads to grammatically correct but meaningless expressions and sentences. As a well-taken example, Čermák (2006) gives the famous sentence composed by Chomsky (1957): *Colorless green ideas sleep furiously*. Each word combination in this sentence (and thus the sentence itself) is grammatically correct but nonsensical.[1]

[1] Although the expression *green ideas* can nowadays have a figurative meaning and be interpreted as ideas that are "*environmentally friendly*."

In general, the ability of a word to combine with other words in text (or speech) is called **collocability**. It is governed by both semantic and grammatical (and pragmatic) rules and expressed in terms of paradigms – sets of words substitutable (functionally equivalent) in a specific context (as a combination with a given word). It can be specified either *intensionally* – by a description of the same syntactic and semantic properties, which forms **valency**, or *extensionally* – by enumeration, where no summary specification can be applied. On this basis, Čermák and Holub (1982, p. 10) defined **collocation** (quite broadly) as a realization of collocability in text, and later (2001) as a "meaningful combination of words [...] respecting their mutual collocability and also compatibility".

Naturally, different words have a different degree of collocability (examples from Čermák and Holub, 1982): On one hand, words like *be*, *good*, and *thing* can be combined with a wide range of other words and only general (syntactic) rules are required for producing correct expressions with such words. On the other hand, the collocability of words like *bark*, *cubic*, and *hypertension* is more restricted and knowledge of these (semantic) constraints is quite useful (together with the general rules) to produce a more cohesive text. Furthermore, there are words that can be combined with only one or a select few others; their knowledge (lexical and pragmatic) is absolutely essential for their correct usage in language, and they cannot be used otherwise (no general rules apply).

The scale of collocability ranges from **free word combinations** whose component words can be substituted by another word (i.e. synonym) without significant change in the overall meaning and if omitted, they can not be easily predicted from the remaining components, to **idioms** whose semantics can not be inferred from the meanings of the components. Čermák's notion of collocation based on mutual collocability and compatibility spans a wide range of this scale. The research in natural language processing is usually focused on the narrower concept: word combinations with extensionally restricted collocability – in literature described as significant (Sinclair, 1966), habitual, fixed, anomalous and holistic (Moon, 1998), unpredictable, mutually expected (Palmer, 1968), mutually selective (Cruse, 1986), or idiosyncratic (Sag et al., 2002).

2.1.2 Historical perspective

The idea of collocation was first introduced into linguistics by Palmer (1938), an English linguist and teacher. As a concept, however, collocations were studied by Greek Stoic philosophers as early as in the third century B.C. They believed that "word meanings do not exist in isolation, and may differ according to the collocation in which they are used" (Robins, 1967). Palmer (1938) defined collocations as "successions of two or more words the meaning of which can hardly be deduced from a knowledge of their component words" and pointed out that such concepts "must each be learnt as one learns single words", e.g. *at least*, *give up*, *let alone*, *as a matter of fact*, *how do you do* (see also Palmer and Hornby, 1937). Collocations as a linguistic phenomenon were studied mostly in British linguistics (Firth, Halliday, Sinclair) and rather neglected in structural linguistics (Saussure, Chomsky).

An important contribution to the theoretical research of collocations was made by John R. Firth who used the concept of collocation in his study of lexis to define a meaning of a single word (Firth, 1951, 1957). He introduced the term **meaning by collocation** as a new *mode of meaning* of words and distinguished it from both the "conceptual or idea approach to the meaning of words" and "contextual meaning". Uniquely, he attempted to explain it at the syntagmatic, not the traditional paradigmatic, level (by semantic relations such as synonymy or antonymy)[2]. With the example *dark night*, he claimed that one of the meanings of *night* is its collocability with *dark*, and one of the meanings of *dark* is its collocability with *night*. Thus, a complete analysis of the meaning of a word would have to include all its collocations. In Firth (1957, p. 181), "collocations of a given word" were defined as "statements of the habitual or customary places of that word." Later (see Palmer, 1968), he used a more famous definition and described collocation as "the company a word keeps".

Firth's students and disciples, known as Neo-Firthians, further developed his theory. They regarded lexis as complementary to grammar and used collocations as the basis for a lexical analysis of language alternative to (and independent from) the grammatical analysis. They argued that grammatical description does not account for all the patterns in a language, and promoted the study of lexis on the basis of corpus--based observations. Halliday (1966) defined collocation as "a linear co-occurrence relationship among lexical items which co-occur together" and introduced the term *set* as "the grouping of members with like privilege of occurrence in collocation". For example, *bright*, *hot*, *shine*, *light*, and *come out* belong to the same lexical set, since they all collocate with the word *sun* (Halliday, 1966, p. 158).

Sinclair (1966) also regarded grammar and lexicon as "two different interpenetrating aspects". He dealt with quite general "tendencies" of lexical items to collocate with one another which "ought to tell us facts about language that cannot be got by grammatical analysis". He introduced the following terminology for the structure of collocations: a *node* as the item whose collocations are studied, a *span* as the number of lexical items on each side of a node that are considered relevant to that node, and *collocates* as the items occurring within the span. He even argued that "there are virtually no impossible collocations, but some are much more likely than others" (Sinclair, 1966, p. 411) but later distinguished between *casual collocations* and *significant collocations* that "occur more frequently than would be expected on the basis of the individual items". In Sinclair (1991, p. 170), collocation were defined directly as "occurrence of two or more words within a short space of each other in a text", where "short space"

[2] The paradigmatic relationship of lexical items consists of sets of words belonging to the same class that can be substituted for one another in a certain grammatical and semantic context. The syntagmatic relationship of lexical items refers to the ability of a word to combine with other words (collocability).

was suggested as a maximum of four words intervening together. Sinclair also added that "Collocations can be dramatic and interesting because unexpected, or they can be important in the lexical structure of the language because of being frequently repeated."

Halliday and Hasan (1967, p. 287) described collocation as "a cover term for the cohesion that results from the cooccurrence of lexical items that are in some way or other typically associated with one another, because they tend to occur in similar environments" and gave examples such as: *sky – sunshine – cloud – rain* or *poetry – literature – reader – writer – style*, etc.

Mitchell (1971) considered lexis and grammar as interdependent, not separate and discrete, but forming a continuum. He argued for the "oneness of grammar, lexis and meaning" (p. 43) and suggested collocations "to be studied within grammatical matrices [which] in turn depend for their recognition on the observation of collocational similarities" (p. 65). By the grammatical matrices he understood patterns such as *adjective – noun*, *verb – adverb*, or *verb – gerund*. Fontenelle (1994b, p. 43), on the other hand, perceived the concept of collocation as "independent of grammatical categories: the relationship which holds between the verb *argue* and the adverb *strongly* is the same as that holding between the noun *argument* and the adjective *strong*".

2.1.3 Diversity of definitions

The disagreement on the notion of collocation among different linguists is quite remarkable not only in historical context but also in current research. None of the existing definitions of collocation is commonly accepted either in formal or computational linguistics. In general, the definitions are based on five fundamental aspects, which we address in the following passages (cf. Moon (1998) and Bartsch (2004)):

1) grammatical boundedness,
2) lexical selection,
3) semantic cohesion,
4) language institutionalization,
5) frequency and recurrence.

1) Grammatical boundedness

By grammatical boundedness, we mean a (direct) syntactic relationship between components of collocation. This criterion was omitted in early studies on collocations. Sinclair's concept of collocation presented in the previous section (Sinclair, 1966) suggests that all occurrences (including those not grammatically bounded) of two or more words can be considered collocations (they are co-located). More notably, Halliday's and Hasan's (1967) definition describing words which "tend to occur in similar environments“ directly implies that collocations do not necessarily appear as grammatical units with a specific word order, e.g. *hair, comb, curl, wave* or *candle, flame, flicker* (see also above). Halliday and Hasan (1967, p. 287) even emphasized that they are ”largely independent of the grammatical structure“. For such classes, of words that are “likely to be used in the same context” (semantically related but not syntactically dependent) Manning and Schütze (1999, p. 185) suggested to use the terms *association* or *co-occurrence*, e.g. *doctor, nurse, hospital*. In his later work, Hasan (1984) rejected his previous definition of collocation as too broad and used the term *lexical chain* for this concept.

The grammatical aspect became important in the notion of collocation based on lexical collocability by Čermák (see below). Also Kjellmer (1994, p. xiv) explicitly defined collocations as “recurring sequences that are grammatically well formed”. Similarly, Choueka (1988) used the expression “a syntactic and semantic unit” in his definition of collocation. Although most of the current definitions are not explicit about grammatical boundedness, they usually assume that collocations form grammatical expressions implicitly.

2) Lexical selection

The process of lexical selection in natural language production (generation) is closely related to collocability (expressing the ability of words to be combined with other words, see Section 2.1.1). Collocations (as opposed to free word combinations) are often characterized by restricted (or preferred) lexical selection, i.e. not-easily-explainable patterns of word usage (Manning and Schütze, 1999, p. 141). For example, *Meals will be served outside on the terrace, weather permitting* vs. **Meals will be served outside on the terrace, weather allowing*. Although *to allow* and *to permit* have very similar meanings, in this combination, only *permitting* is correct. For the same reason, *stiff breeze* is correct but **stiff wind* is not, *strong tea* is correct and **powerful tea* not, although *powerful drugs* and *strong cigarette* are correct too (examples from Manning and Schütze, 1999, Chapter 5).

Constrained lexical selection (morpho-syntactic preference) is what distinguishes free word combinations from collocations, which Bahns (1993, p. 253) depicted as “springing to mind in such a way as to be said to be psychologically salient”. Kjellmer (1991, p. 112) stated that “the occurrence of one of the words in such combination can be said to predict the occurrence of the other(s)”. Similarly Bartsch (2004, p. 11) claimed that “the choice of one of the constituents appears to automatically trigger the selection of one or more other constituents in their immediate context” and “block the selection of other lexical items that, according to their meaning and morpho-syntactic properties, appear to be eligible choices in the same expression”. Bartsch (2004, p. 60) also discussed directionality of the process of co-selection, but for the notion of collocation it seems not important.

3) Semantic cohesion

The criterion of semantic cohesion reflects the semantic transparency or opacity (compositionality or non-compositionality) of word combinations. Many researchers use cohesion to distinguish between idioms and collocations as different lexical phenomena. Benson (1985, p. 62) clearly stated that "the collocations [...] are not idioms: their meanings are more or less inferrable from the meanings of their parts". Idioms do not reflect the meanings of their component parts at all, whereas the meaning of collocations does reflect the meanings of the parts (Benson et al., 1986, p. 253).

Cruse (1986, p. 37–41) also distinguished between collocations and idioms. He perceived idioms as "lexically complex" units, forming a "single minimal semantic constituent", "whose meaning cannot be inferred from the meaning of its parts". He used the term collocation to "refer to sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent" and gave examples such as *fine weather*, *torrential rain*, *light drizzle*, and *high winds*. He also added that collocations are "easy to distinguish from idioms; nonetheless they do have a kind of semantic cohesion – the constituent elements are, to varying degrees, mutually selective". The cohesion is especially evident when "the meaning carried by one (or more) of the constituent elements is highly restricted contextually, and different from its meaning in more neutral contexts". He also introduces "bound collocations" as expressions "whose constituents do not like to be separated" and "transitional area bordering on idiom" (e.g. *foot the bill* and *curry flavour*).

Fontenelle (1994b) stated that collocations are both "non-idiomatic expressions" as well as "non-free combinations". He characterized idiomatic expressions by "the fact that they constitute a single semantic entity and that their meaning is not tantamount to the sum of the meanings of the words they are made up of" (e.g. *to lick somebody's boots* which is neither about licking, nor about boots). To illustrate the difference between collocations and free-combinations he gave an example of adjectives *sour*, *bad*, *addled*, *rotten*, and *rancid* that all can be combined with nouns denoting food, but they are not freely interchangeable. Only *sour milk*, *bad/addled/rotten egg*, and *rancid butter* are correct collocations in English. Other combinations such as **rancid egg*, **sour butter*, and **addled milk* are unacceptable.

Some researchers, however, do not explicitly exclude idioms from collocations – Wallace (1979) even perceived collocations (and also proverbs) as subcategories of idioms. Carter (1987, p. 58) considered idioms and fixed expressions as subclasses of collocations. He described idioms as "restricted collocations which cannot normally be understood from the literal meaning of the words which make them up" such as *have cold feet* and *to let the cat out of the bag*. He argued that among collocations, there are also other fixed expressions, such as *as far as I know*, *as a matter of fact*, and *if I were you* that are not idioms but are also "semantically and structurally restricted".

Similarly, Kjellmer (1994, p. xxxiii) used *collocation* as an inclusive term and presented *idiom* as a "subcategory of the class of collocations" defined as "a collocation whose meaning cannot be deduced from the combined meanings of its constituents". Choueka (1988) also included idioms in his definition of collocation: "[A collocation expression] has a characteristics of a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components." Manning and Schütze (1999, p. 151) claimed that "collocations are often characterized by limited compositionality" and that "idioms are the most extreme examples of non-compositionality. Also Čermák (2001) explicitly conceived idioms as a subtype of collocations (see Section 2.1.4).

4) Language institutionalization

In general, language institutionalization is a process by which a phrase becomes "recognized and accepted as a lexical item of the language" (Bauer, 1983). Institutionalized phrases, originally fully compositional and free word combinations, become significant and idiosyncratic by their frequent and consistent usage (especially in comparison with other alternative lexicalizations of the same concept). Baldwin and Villavicencio (2002) illustrated this phenomenon on the example of *machine translation*: "There is no particular reason why one could not say *computer translation* [...] but people do not." Bauer (1983) gave examples such as *telephone booth* (correct in American English) vs. *telephone box* (correct in British English), *salt and pepper*, etc. Institutionalized phrases are domain-dependent – they are adopted only within a certain domain and not elsewhere (e.g. *carriage return* in computer science, *white water* in sports).

5) Frequency of occurrence

Frequency of occurrence plays an important role in many attempts to describe and define collocations. Benson et al. (1986, p. 253) characterized collocation as being "used frequently", Bartsch (2004) defined collocations as "frequently recurrent, relatively fixed syntagmatic combinations of two or more words". Frequency is closely related to institutionalization but it is difficult to be quantified. Kjellmer's (1987, p. 133) restriction on sequences "of words that occur more than once in identical form and is grammatically well-structured" is apparently insufficient. The key issue here is corpus representativeness – which is, in general, insufficient and therefore no absolute constraint can be imposed on a phrase as a frequency limit to become recognized as a collocation. Sinclair (1991) defined a collocation as the "occurrence of two or more words within a short space of each other in a text" that makes potentially any cooccurrence of two or more words a collocation – which is also questionable.

Some more statistically motivated definitions are not based on the absolute frequency of occurrence but rather on its statistical significance, where frequency of component words is also taken into account: Church and Hanks (1990) defined a collocation as "a word pair that occurs together more often than expected", McKeown and

Radev (2000) as "a group of words that occur together more often than by chance", and Kilgarriff (1992, p. 29) as words co-occuring "significantly more often than one would predict, given the frequency of occurrence of each word taken individualy". Sinclair (1966, p. 411) defined *significant collocations* as combinations occuring "more frequently than would be expected on the basis of the individual items". This approach is fundamental for methods of automatic collocation extraction but it suffers from the problem of a limited corpus representativeness and data sparsity in general.

2.1.4 Typology and classification

Several attempts have been made to design a topology or classification of collocations and related concepts. All of them are closely tied to the definition of the studied concept and the criteria used for its classification. In this section, we present four representative approaches to illustrate the diversity of the notion of collocation among theoretical and also applied linguists.

Word combinations by van der Wouden (1997)

Van der Wouden (1997, 8–9) used the following categorization of word combinations based on semantic cohesion (cf. also Benson et al., 1986). Here, collocations occupy a relatively narrow part of the scale but among the other types they are denoted as *fixed expressions* as opposed to *free word combinations*.

1) free combinations
the components combine most freely with other lexical items
a murder + verbs, such as *to analyze* and *to describe*

2) collocations
loosely fixed combinations between idioms and free combinations
to commit a murder

3) transitional combinations
appear between idioms and collocations, more frozen than ordinary collocations and, unlike idioms, these combinations seem to have a meaning close to that suggested by their component parts
to catch one's breath

4) idioms
relatively frozen, meanings do not reflect the meaning of the components
to kick the bucket

5) proverbs/sayings
usually more frozen than idioms but form complete sentences
a friend in need is a friend indeed

6) compounds
totally frozen with no possible variations
definite article

Fixed expressions and idioms by Moon (1998)

Moon (1998) dealt with the term *fixed expressions and idioms* (FEIs). In her work, she stated that ″no clear classifications [of FEIs] are possible″ and suggested that ″it should be stressed that FEIs are non-compositional (to some extent); *collocations* and *idioms* represent two large and amorphous subgroups of FEIs on continuum; transformational deficiencies are a feature of FEIs but not criterial; and discoursally or situationally constrained units should be considered FEIs″ 1998, p. 19–21 Her topology was based on the identification of the primary reasons why each potential FEI might be ″regarded lexicographically as a holistic unit: that is, whether the string is problematic and anomalous on grounds of lexicogrammar, pragmatics, or semantics″. This typology has three macrocategories *anomalous collocations*, *formulae*, and *metaphors*, each divided into finer grained subcategories.

A) anomalous collocations (problems of lexicogrammar)

1. **ill-formed collocations** – syntagmatically or paradigmatically aberrant
   at all, by and large
2. **cranberry collocations** – idiosyncratic lexical component
   in retrospect, kith and kin
3. **defective collocations** – idiosyncratic meaning component
   in effect, foot the bill
4. **phraseological collocations** – occurring in paradigms
   in/into/out of action, on show/display

B) formulae (problems of pragmatics)

1. **simple formulae** – routine compositional strings with a special discourse function; *alive and well, you know*
2. **sayings** – quotations catch-phrases, truism
   an eye for an eye; a horse, a horse, my kingdom for a horse
3. **proverbs (literal/metaphorical)** – traditional maxims with deontic functions
   you can't have your cake and eat it, enough is enough
4. **similes** – institutionalized comparisons
   as good as gold, live like a king

C) metaphors (problems of semantics)

1. **transparent metaphors** – expected to be decoded by real-world knowledge
   behind someone's back, pack one's bags
2. **semi-transparent metaphors** – special knowledge required for decoding
   on an even keel, pecking order
3. **opaque metaphors** – absolutely-compositional
   bite the bullet, kick the bucket

Lexical combinations by Čermák (2001)

Čermák (2001) attempted to classify lexical combinations by two basic linguistic distinctions: *stableness* (stable – unstable, langue – parole, system – text) and *regularity* (regular – irregular) into the types shown below (see also Čermák and Šulc, 2006). This classification, compared to others, is quite systematic but not quite consistent with Čermák's definition based on collocability and compatibility (see Section 2.1.1). Apparently, not all word combinations are considered to be collocations, but the collocations do subsume idioms in this case. Čermák also emphasized that the main types A and B are not absolutely distinct and introduced the C type into his classification as the boundary case between the types A1a and B3a.

A) Langue	1. *regular*	a)	**terminological collocations** (multiword technical terms) *cestovní kancelář (travel agency), kyselina sírová (sulphuric acid)*
		b)	**proprial collocations** (multiword proper names) *Kanárské ostrovy (Canary Islands), Velká Británie (Great Britain)*
	2. *irregular*		**idiomatic collocations** (idioms and phrasemes) *ležet ladem (lie fallow), jen aby (just to)*
B) Parole	3. *regular*	a)	**common collocations** (gram-semantic combinations) *letní dovolená (summer vacation), snadná odpověď (easy answer)*
		b)	**analytical form combinations** (analytical forms) *šel by (would go), byl zapsán (was subscribed)*
	4. *irregular*	a)	**individual metaphoric collocations** (authors' metaphors) *třeskutě vtipný (bitingly funny), virové hrátky (viral games)*
		b)	**random adjacent combinations** (adjacent occurrences) *uvnitř bytu (inside [an] apartment), že v (that in)*
		c)	**other combinations** (babble)
C) Langue/Parole	5.		**common usage collocations** (boundary type A1a–B3a) *umýt si ruce (wash hands), nastoupit do vlaku (board [the] train)*

Multiword expressions by Sag et al. (2002)

Sag et al. (2002, p. 2) defined *multiword expressions* (MWE) "roughly as idiosyncratic interpretations that cross word boundaries (or spaces)" and stated that the "problem of multiword expressions is underappreciated in the field at large" and later "MWEs appear in all text genres and pose significant problems for every kind of NLP." As the main problems, Sag et al. mentioned "overgeneration", when no attention is paid to collocational preferences in language generation (e.g. **telephone cabinet* instead of *telephone box* in British or *telephone booth* in American), and "idiomaticity" leading to missinterpretation of idiomatic and metaphoric expressions (e.g. *kick the bucket*). The terminology used in the proposed classification is adopted from (Bauer, 1983). The term collocation is not used at any level of the classification, it is used to refer to "any statistically significant cooccurrence, including all forms of MWE as described above and compositional phrases which are predictably frequent (because of real world events or other nonlinguistic factors)." For example, *sell* and *house* appear more often than one can predict from the frequency of the two words, but "there is no reason to think that this is due to anything other than real world facts."

A) lexicalized phrases
have at least partially idiosyncratic syntax or semantics, or contain words which do not occur in isolation:

1. fixed expressions
immutable expressions that defy conventions of grammar and compositional interpretation, e.g. *by and large*, *in short*, *kingdom come*, *every which way*; they are fully lexicalized and undergo neither morphosyntactic variation (cf. **in shorter*) nor internal modification (cf. **in very short*)

2. semi-fixed expressions
adhere to strict constraints on word order and composition, but undergo some degree of lexical variation, e.g. in the form of inflection, variation in reflexive form, and determiner selection

a) non-decomposable idioms
kick the bucket, *trip the light*

b) compound nominals
car park, *attorney general*, *part of speech*

c) proper names
San Francisco, *Oakland Riders*

3. syntactically-flexible expressions
exhibit a much wider range of syntactic variability

a) verb-particle constructions
write up, *look up*, *brush up on*

b) decomposable idioms
Idioms such as *spill the beans*, for example, can be analyzed as being made up of spill in a *reveal* sense and the beans in a *secret(s)* sense, resulting in the overall compositional reading of *reveal the secret(s)*
let the cat out of the bag, *sweep under the rug*

c) light verbs
make a mistake, *give a demo*

B) institutionalized phrases
syntactically and semantically compositional but statistically idiosyncratic, they occur with remarkably high frequency (in a given context), e.g. *traffic light*.

2.1.5 Conclusion

There is no commonly accepted definition of collocation and we do not aim to create one in this work. Based on Čermák's notion of compatibility and collocability (Section 2.1.1), we understand **collocation** as a meaningful and grammatical word combination constrained by extensionally specified restrictions and preferences. This approach has two important aspects: First, it restricts collocations only to meaningful grammatical expressions and therefore combinations of incompatible words (e.g. *green idea*) and combinations of words without direct syntactic relationship (e.g. *doctor – nurse*) cannot form collocations. Second, combination of words in a collocation must be governed not only by syntactic and semantic rules but also by some other restrictions that cannot be based on the description of syntactic and semantic properties of the components – they must be specified explicitly by enumeration, i.e. extensionally.

This approach is quite similar to that presented by Evert (2004). His notion of collocation is based on the definition by Choueka (1988) saying that "[A collocation expression] has a characteristics of a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components." Evert added to his notion only an explicit criterion that should help to distinguish between collocational and non-collocational expressions: "Does it deserve a special entry in a dictionary or lexical database of the language?" and defined **collocation** as "a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon" (Evert, 2004, p. 9), which only emphasizes the extensional character of collocations – to be enumerated, listed in a lexicon.

Also, in a similar manner to Evert (2004), we use **collocation** as "a generic term whose specific meaning can be narrowed down according to the requirements of a particular research question or application" (Evert, 2004, p. 9). However, each experiment presented in this work is performed on a specific data set and bounded with a particular definition of the studied concept (or its subtype) and thus it is always clear what phenomenon we deal with.

The presented notion of collocation is possibly interchangeable with the concept of **multiword expression** (MWE) that has became commonly preferred and accepted by many authors and researchers. Baldwin (2006) defined it as an expression that is "1) decomposable into multiple simplex words and 2) lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic". However, mainly for historical and traditional reasons, we keep using the term **collocation** to refer to this concept in our work.

2.2 Collocation extraction

Collocation extraction is a traditional task of *corpus linguistics*. The goal is to extract a list of collocations from a text corpus. Generally, it is not required to identify particular occurrences (instances, tokens) of collocations, but rather to produce a list of all collocations (types) appearing anywhere in the corpus – a **collocation lexicon**. This task is often restricted to a particular subtype or subset of collocations (defined e.g. by grammatical constraints), but we deal with it in a general sense. The first research attempts in this area are dated back to the era of "mechanized documentation" (Stevens et al., 1965). The first work focused particularly on collocation extraction was published by Berry-Rogghe (1973), and later followed by studies by Choueka et al. (1983), Church and Hanks (1990), Smadja (1993), Kita et al. (1994), Shimohata et al. (1997), and many others, especially in the last ten years (Krenn, 2000; Evert, 2004; Bartsch, 2004)

In the following sections, we first briefly discuss the basic principles of collocation extraction and then, in more detail, we describe individual steps of the entire extraction process. The reference corpus we use in our examples in this section is the Prague Dependency Treebank 2.0 (PDT), described in detail later in Section 4.2.

2.2.1 Extraction principles

Methods for collocation extraction are based on several different **extraction principles**. These principles exploit characteristic properties of collocations and are formulated as hypotheses (assumptions) about word occurrence and cooccurrence statistics extracted from a text corpus. Mathematically, they are expressed as formulas that determine the degree of collocational association between words. These formulas are commonly called **lexical association measures**. In this work, we focus our attention on measures based on the following three extraction principles:

1) Collocation components occur together more often than by chance

The simplest approach to discover collocations in a text corpus is counting their occurrences – if two words occur together a lot, then that might be the evidence that they have a special function that is not simply explained as a result of their combination (Manning and Schütze, 1999, p. 153). The assumption that collocations occur more frequently than arbitrary word combinations is reflected in many definitions of collocation (see Section 2.1.3) but in practice it presents certain difficulties:

First, natural language contains some highly frequent word combinations that do not form collocations, such as various combinations of function words (words with little lexical meaning, expressing only grammatical relationship with other words). For example, the most frequent word combination (with a direct syntactic relation between components) in PDT is *by měl* (*would have*) with frequency 2 124, while the most frequent combination that can be considered a collocation is *Česká republika* (*Czech Republic*) occurring only 527 times. Such "uninteresting" combinations should be identified and eliminated during the extraction process.

Second, high frequency of certain word combinations can be purely accidental – very frequent words are expected to occur together a lot just by chance, even if they do not form a collocation. For example, the expression *nový zákon* (*new law*, not in the sense of *new testament*) is among the 35 most frequent adjective-noun combinations although it is not a collocation (not surprisingly, the words *nový* (*new*) and *zákon* (*law*) are indeed very frequent; in PDT, the word *nový* (as masculine inanimate) occurs 777 times and the word *zákon* occurs 1575 times – both are among the most frequent adjectives and nouns).

The basic principle of collocation extraction is based on distinguishing between random (free) word combinations that occur together just by chance, and those that are not accidental and possibly form collocations. Herein, not only the frequency of word cooccurrences but also the frequencies of words occurring independently are taken into account. The corpus is observed as a sequence of randomly and independently generated word bigrams (a random sample), and their joint and marginal occurrence frequencies are then employed in various association measures to estimate how much the word cooccurrence is accidental.

One class of association measures using this principle is based on statistical hypothesis testing: The *null hypothesis* is formulated such that there is no association between the words beyond chance occurrences. The association measures are, in fact, the test statistics used in these hypothesis tests. Other classes of measures using this principle are *likelihood ratios* (expressing how much more likely one hypothesis is against the other), and other (mostly heuristic) measures of statistical association or measures adopted from other fields, such as information theory (Church et al., 1991) and others.

2) Collocations occur as units in information-theoretically noisy environment

While the previous principle deals with the relationship of words *inside* collocations, in this approach, we analyse the *outside* relationships of collocations, i.e. words which immediately precede or follow a collocation in the text stream (*immediate contexts*).

By determining the entropy of these contexts, we can discover points in the word stream with either low or high uncertainty (disorder) what the next (or previous) word will be. "Points with high uncertainty are likely to be phrase boundaries, which in turn are candidates for points where a collocation may start or end, whereas points with low uncertainty are likely to be located within a collocation." (Manning and Schütze, 1999, p. 181). In other words, entropy inside collocations is expected to be lower (low uncertainty, high association) and outside collocations to be higher (high uncertainty, low association). Methods based on this assumption has been employed e.g. by Evans and Zhai (1996), Shimohata et al. (1997), and Pearce (2002).

In this principle, the corpus is again interpreted as a sequence of randomly (and independently) generated words. For each collocation candidate, we estimate probability distribution of words occurring in its immediate contexts (left and right) and determine its lexical association based on measuring entropy of these contexts.

3) Collocations occur in different contexts to their components

Limited compositionality is a typical property of collocations – the meaning of a collocation cannot be fully inferred from the meanings of its components. In other words, meaning of a collocation must (to some extent) differ from the "union" of the meaning of its components (see Section 2.1.3). Traditional examples of this property are idiomatic expressions (e.g. *kick the bucket* – there is no *bucket* nor *kicking* in the meaning of this idiom).

A typical way of modeling senses in natural language processing is by *empirical contexts*, i.e. by a bag of words occurring within a specified context window of a word or an expression. The more different the contexts are, the higher the chance is that the expression is a collocation (Zhai, 1997). Lexical association measures based on this principle are adopted from mathematics (vector distance), information theory (cross entropy, divergence), and from the field of information retrieval (vector similarity).

A major weakness of most lexical association measures lies in their unreliability when applied to low frequency data. They either assume word occurrence probabilities to be approximately normally distributed (e.g. *t-test*), which is not true in general (Church and Mercer, 1993) and unensurable to assume when dealing with frequencies around five or less. Or, they are just sensitive to estimates that are inaccurate due to data sparsity (e.g. *Pointwise mutual information*), see Manning and Schütze (1999, p. 181).

Other extraction principles

Various other extraction principles have been proposed for collocation extraction but they are not of our interest in this work – they either require additional linguistic resources or they are not based on measuring lexical association. For example, Manning and Schütze (1999, Chapter 5) described a technique based on analysis of the mean and variance of distance between the components of word combinations. Pearce (2002) exploited another characteristic property of collocation – non-substitutability and measured whether collocation components can be replaced by their synonyms, where *Wordnet* (Fellbaum, 1998) was used as a source of such (lexical) synonyms. Several researchers have also attempted to extract collocations (and their translations) from bilingual parallel corpora and its word alignment, e.g. Ohmori and Higashida (1999) or Wu and Zhou (2003).

2.2.2 Extraction pipeline

Automatic collocation extraction is usually performed as a process consisting of several steps, called the **extraction pipeline** (Krenn, 2000; Evert and Kermes, 2003):

First, the corpus as a collection of machine-readable texts in one language is linguistically pre-processed – morphologically and syntactically analyzed and disambiguated. Second, all **collocation candidates** (i.e. potential collocations) are identified and their occurrence statistics extracted from the corpus. Third, the candidates are filtered to improve precision (based on grammatical patterns and/or occurrence frequency). Fourth, a lexical association measure is chosen and applied to the occurrence statistics obtained from the corpus. And finally, the collocation candidates are classified according to their association scores and a certain threshold – candidates above this threshold are classified as collocations and candidates below the threshold as non-collocations.

However, there is no principled way of finding the optimal classification threshold (Inkpen and Hirst, 2002) – its value depends primarily on the intended application (whether high precision or broad coverage is preferred) and is usually set empirically. To avoid this step, the task of collocation extraction is usually reformulated as **ranking collocation candidates** – the goal is not to extract a discreet set of collocations from a given corpus, but instead to **rank** all potential collocations according to their degree of association so that the most associated ones are concentrated at the top of the list. This approach to collocation extraction will be applied in the rest of our work. The extraction pipeline for bigram collocation extraction will be described in detail in the following sections, and lexical association measures will be presented separately in the next chapter.

2.2.3 Linguistic preprocessing

By linguistic preprocessing, we mean the analysis and disambiguation at the level of **morphology** and **surface syntax**. Higher levels (more abstract) of linguistic processing (e.g. deep syntax) are not useful since we are interested only in the association at the lexical level. In this step, information about word *base forms*, *morphological categories*, and sentence *syntax* is obtained in order to identify collocation candidates and all their occurrences – regardless of inflectional variance and sentence position.

Formally, a **source corpus** W is expected in the form of a linearly ordered set of n **word tokens** w_i identified as contiguous, non-overlapping strings v_i over an **alphabet** Σ distinguished by their position $i = 1, \dots, n$ in the corpus, so the i-th word token w_i is a pair $\langle i, v_i \rangle$. The ordering of W is defined by the natural ordering of the positions. The items v_i are called **word forms** and the set of all possible word forms is called the **vocabulary** V.

$$W = \{w_1, \dots, w_n\}, \quad w_i := \langle i, v_i \rangle, \quad v_i \in V \subset \Sigma^*, \quad i = 1, \dots, n.$$

During morphological analysis and disambiguation, each word token w_i from W is assigned (by mapping ϕ) a (basic) **word type** u (from a set of all such word types U). The word types define equivalence classes of word tokens based on *inflection*, so all inflectional variants are assigned the same value u. We denote u_i as the word type assigned to the word token w_i.

$$\phi\colon W \to U, \quad u_i := \phi(w_i), \quad i = 1, \dots, n.$$

Technically, each $u \in U$ is usually a pair $\langle l, t \rangle$ where l is a **lemma** – a word base form as it appears in the **lexicon** L, and t is a **tag** from the **tag set** T specifying detailed morphological characteristics (e.g. *derivational*) shared by all the inflectional variants.

$$u := \langle l, t \rangle, \quad l \in L, \quad t \in T.$$

The word types are defined to conflate all word tokens not only with the same word base form but also with the same *lexical meaning* – which may not be fully reflected in the word base form. Details strongly depend on the system employed for encoding the morphological information in the corpus. For example, in the Czech system used in PDT, the information about the morphological categories *negation* or *grade* (degree of comparison) which are considered derivational and which discriminate word meanings, is encoded in the *tag*, not in the *lemma*. For this reason, e.g. the word types of *nebezpečný* (*insecure*) and *nejvyšší* (*highest*) must be encoded as ⟨*bezpečný, 1N*⟩ (*secure*, 1^{st}*grade+negative*) and ⟨*vysoký, 3A*⟩ (*high*, 3^{rd}*grade+affirmative*), respectively (for details, see also Section 4.2.1).

During syntactic analysis and disambiguation, each word token w_i from the corpus W is assigned (by a function δ applied to its index i) an index j of its **head word** w_j (in terms of dependency syntax, w_j governs w_i) and (by a mapping α) the **analytical function** a (from the set A of all possible analytical functions enriched by a special value *HEAD*, see details bellow) specifying the type of syntactic relation between the word token and its head word. The head word of a word token w_i is either another word token $w_j, i \neq j$ from the same sentence, or the value *NULL* if w_i is the root of the sentence ($j = 0$). We denote a_i as the analytical function assigned to the word token w_i.

$$\delta\colon \{1, \dots, n\} \to \{0, \dots, n\}, \quad \delta(i) \neq i,$$
$$\alpha\colon W \to A, \quad a_i := \alpha(w_i), \quad i = 1, \dots, n.$$

In order to identify word tokens that are not only inflectional variants but also have the same syntactic function, each word token w_i can be assigned (by a mapping φ) an **extended word type** $\langle u_i, a_i \rangle$, which consists of its word type u_i and its analytical function a_i.

$$\varphi\colon W \to U \times A, \quad \varphi(w_i) := \langle u_i, a_i \rangle, \quad u_i = \phi(w_i), \quad a_i = \alpha(w_i), \quad i = 1, \dots, n.$$

For technical reasons, we also define a special extended word type that can be assigned (by a mapping φ') to any word token w_i and consists of its word type u_i and the special value of analytical function $a_i = \textit{HEAD}$. This extended word type will be used to label head words appearing in a dependency relation with other words.

$$\varphi' \colon W \rightarrow U \times A, \quad \varphi(w_i) := \langle u_i, \textit{HEAD} \rangle, \quad u_i = \phi(w_i), \quad i = 1, \dots, n.$$

Generally, linguistic preprocessing is not necessarily required for collocation extraction, especially when working with languages with simple morphology (such as English) and if we focus e.g. only on fixed adjacent and non-modifiable collocations. However, if we have to deal with complex morphology (e.g. in Czech) and if we want to extract syntactically bounded word combinations with free word order, this information is quite useful.

Linguistic information can also be used in the subsequent steps of the extraction pipeline for filtering collocation candidates (see Section 2.2.6) and to construct additional features in methods combining statistical and linguistic evidence in more complex classification and ranking models (see Chapter 6).

2.2.4 Collocation candidates

Collocation candidates represent the set of all potential collocations appearing in the corpus, i.e. the word combinations that satisfy some basic requirements imposed on collocations (e.g. components to be in a direct syntactic relation or to occur within a given distance in the text). Collocation candidates are examined with respect to the degree their components are associated, and ranked according to their strength of association, as specified in the task description. The goal in this step of the extraction pipeline is to identify all collocation candidates and their instances (occurrences) in the corpus. In the beginning, we describe this step on a general level, then with details of specific approaches.

First, the corpus W is by some means transformed to a set B consisting of **bigram tokens** $b_k = \langle w_i, w_j \rangle$, i.e. pairs of word tokens from the corpus satisfying some given conditions. Elements of B are indexed by $k \in \{1, \dots, N\}$, where $N = |B|$, although the actual ordering of this set is not important.

$$B = \{b_1, \dots, b_N\}, \quad B \subset W \times W, \quad b_k := \langle w_i, w_j \rangle, \quad k = 1, \dots, N.$$

Second, each bigram token b_k from the set B is assigned (by a mapping Φ) a **bigram type** c (from a set C^* of all possible bigram types) defining equivalence classes of bigram tokens based on inflection – all bigram tokens that differ only in inflection are assigned the same bigram type c. Bigram types identified by Φ in B are called **collocation candidates** and a set of all such bigram types is denoted by C. Each bigram token is thus an instance of a collocation candidate. We denote c_k as the bigram type of the bigram token b_k.

$$\Phi \colon B \rightarrow C^*, \quad c_k := \Phi(b_k), \quad k = 1, \dots, N, \quad C := \Phi(B), \quad C \subset C^*.$$

Third, a *multiset* (allowing repeated elements, also called a *bag*) D, referred to as the **candidate occurrence data** (or candidate data), is acquired as a result of Φ applied on all the elements from B, i.e. bigram types assigned to all bigram tokens. This data serves as a basis for the extraction of occurrence statistics described in the following section.

$$\mathrm{D} = \{c_1, \ldots, c_N\}, \quad c_k := \Phi(b_k), \quad b_k \in \mathrm{B}, \quad k = 1, \ldots, N.$$

The collocation candidate data can be obtained in several alternative ways, depending on the level of linguistic preprocessing of the corpus. These ways differ in how the set of bigram tokens B is constructed and how the mapping Φ is defined to produce the elements of D. In the following paragraphs, we describe three approaches employed in our experiments.

Dependency bigrams

The generic notion of collocation presented in Section 2.1.5 requires collocations to be syntactic units. In dependency syntax, as it is applied in PDT, this constraint can be interpreted as the presence of a *direct dependency relation* between the collocation components. Collocation candidates can then be identified as **dependency bigrams**. The set $\mathrm{B}_{\mathrm{dep}}$ then consists of dependency bigram tokens defined as pairs $\langle w_i, w_j \rangle$ of word tokens from the corpus W in a direct dependency relation of a certain type and in a certain word order.

$$\mathrm{B}_{\mathrm{dep}} = \{\langle w_i, w_j \rangle \in W \times W \colon\ i < j \ \wedge\ (j = \delta(i) \vee i = \delta(j))\}.$$

In general, word order can discriminate between the collocation candidates, and it should be distinguished between bigrams with the first component as the head word and the second one as the modifier and vice versa. For illustration, see the following example: dependency bigrams *velký výr* and *výr velký* differ only in word order; the component *výr* is in both the cases the head word and *velký* is its attribute but the meanings of these expressions are different – the first refers to a *big owl* and the latter denotes *stock owl* as a biological species. On the other hand, in some collocations, word order is not that important: For example, *naklepat maso* (*to tenderize meat*) can occur in this and also in the reverse word order: *Petr naklepal maso* and *Maso jsem naklepal včera* are both correct sentences containing the collocation *naklepat maso*. Since it is not clear how to determine when word order is important and when it is not, we decided to preserve word order in all collocation candidates. This is done by the condition $i < j$ (the first component must always precede the second one in the corpus). For this reason, dependency relations are possible in both directions, either $j = \delta(i)$ or $i = \delta(j)$.

The mapping Φ_{dep} that assigns to each bigram token from $\mathrm{B}_{\mathrm{dep}}$ its bigram type is for dependency bigrams defined by *extended word types* in the following way:

$$\Phi_{\mathrm{dep}}\left(\langle w_i, w_j \rangle\right) = \begin{cases} \langle \varphi(w_i), \varphi'(w_j) \rangle & \text{for} \quad j = \delta(i), \\ \langle \varphi'(w_i), \varphi(w_j) \rangle & \text{for} \quad i = \delta(j). \end{cases}$$

One component of a dependency bigram appearing in a sentence always acts as the head and the other one as the modifier. The head word, however, also participates in another relation outside the bigram as a modifier. This relation is ignored in the dependency bigram and the analytical function of the bigram head word is set to the value *HEAD* (by the mapping φ').

Surface bigrams

Extracting the collocation candidates as dependency bigrams seems quite a reasonable approach. It is guaranteed that each potential collocation is a syntactic unit. However, the source corpus is, in this case, expected to be syntactically analyzed and disambiguated in order to identify such bigrams. If this is not the case, we can detect collocation candidates heuristically, based just on the surface word order. We can assume that most collocations occur as adjacent word expressions that cannot be modified by the insertion of another word, and identify bigram collocation candidates as **surface bigrams** – pairs of adjacent words. The set B_{surf} of surface bigram tokens is formally defined as follows:

$$B_{\mathrm{surf}} = \{\langle w_i, w_j\rangle \in W \times W\colon\ j = i + 1\}.$$

The mapping Φ_{surf} that assigns a surface bigram type to each surface bigram token from B_{surf} is defined by word types of both components in the following way:

$$\Phi_{\mathrm{surf}}\left(\langle w_i, w_j\rangle\right) := \langle \phi(w_i), \phi(w_j)\rangle\,.$$

Distance bigrams

The constraint that collocation candidates are only adjacent word pairs might be too restrictive. Obviously, it is not valid for certain types of collocations, such as *support--verb constructions* or *verb–noun combinations* in general. Collocations of these (and perhaps other) types can often be modified by the insertion of another word and their components can occur at various distances, as in the example *naklepat maso* (*to tenderize meat*) mentioned earlier. In Czech, it can occur not only with free word order but also with various distances between the components. Of course, these cases can be captured by dependency bigrams, but if the syntactic information is not available in the source corpus, we can identify collocation candidates as **distance bigrams** – word pairs occurring within a given distance specified by a distance function d_b and a threshold t_b. The set B_{dist} is then defined by this formula:

$$B_{\mathrm{dist}} = \{\langle w_i, w_j\rangle \in W \times W\colon\ i < j\ \wedge\ d_b(i,j) \leq t_b\}.$$

The mapping Φ_{dist} that assigns a bigram type to each distance bigram token from B_{dist} is then defined in the same way as for surface bigrams:

$$\Phi_{\mathrm{dist}}\left(\langle w_i, w_j\rangle\right) := \Phi_{\mathrm{surf}}\left(\langle w_i, w_j\rangle\right) = \langle \phi(w_i), \phi(w_j)\rangle\,.$$

By one of the mentioned approaches, the candidate data D is constructed as follows:

$$\langle B, \Phi \rangle \in \{\langle B_{\mathrm{dep}}, \Phi_{\mathrm{dep}} \rangle, \langle B_{\mathrm{surf}}, \Phi_{\mathrm{surf}} \rangle, \langle B_{\mathrm{dist}}, \Phi_{\mathrm{dist}} \rangle\},$$

$$D = \{\Phi(b_1), \ldots, \Phi(b_N)\}, \quad b_k \in B, \quad k = 1, \ldots, N, \quad N = |B|.$$

The candidate data of *dependency* and *surface* bigrams are of approximately the same size as the corpus (the number of bigram tokens roughly corresponds to the number of word tokens in the corpus), but the candidate data of *distance* bigrams is larger, depending on the distance function and the threshold (usually set to 3–5 intervening words).

2.2.5 Occurrence statistics

In this step of the extraction pipeline, the occurrence statistics of bigrams and their components are obtained from the candidate occurrence data D and the corpus W. We assume that D is a multiset of generic bigram types (either *dependency*, *surface*, or *distance*) whose components are generic word types (either *basic* or *extended*), elements of U^*. For simplicity of notation, we further denote the elements of D as pairs $\langle x_k, y_k \rangle$:

$$D = \{\langle x_k, y_k \rangle : \; k \in \{1, \ldots, N\}\}, \quad x_k, y_k \in U^*$$

The statistics extracted for each collocation candidate (bigram type) $\langle x, y \rangle \in C$ (for simpler notation further denoted as xy) and its components (word types) x, y from the candidate data, range from simple *frequency counts* and *contingency tables* to more complex models such as *immediate* or *empirical contexts*.

Frequency counts

The basic occurrence model consists of the **frequency counts** of the bigram xy, its components x, y, and the size of the candidate data $N = |D|$.

$$\begin{aligned} f(xy) &:= |\{k : x_k = x \wedge y_k = y\}| \\ f(x*) &:= |\{k : x_k = x\}| \\ f(*y) &:= |\{k : y_k = y\}| \end{aligned}$$

The **bigram frequency** $f(xy)$ (also called **joint frequency**) denotes the number of pairs $\langle x_k, y_k \rangle = \langle x, y \rangle$ in the candidate data D. The **component frequencies** $f(x*)$ and $f(*y)$ (also called **marginal frequencies**) denote the number of pairs where the first component is x and pairs where the second component is y, respectively. N denotes the number of all pairs in D. Evert (2004, p. 28) refers to the quadruple $\langle f(xy)$, $f(x*)$, $f(*y)$, $N\rangle$ as the **frequency signature** of the bigram xy.

$a := f(xy) =: f_{11}$	$b := f(x\bar{y}) =: f_{12}$	$f(x*) =: f_1$
$c := f(\bar{x}y) =: f_{21}$	$d := f(\bar{x}\bar{y}) =: f_{22}$	$f(\bar{x}*)$
$f(*y) =: f_2$	$f(*\bar{y})$	N

Table 2.1: Observed contingency table frequencies (f_{11}, f_{12}, f_{21}, and f_{22}) of a bigram xy, including marginal frequencies (f_1, f_2) summing over the first row and the first column, respectively.

Contingency tables

A more detailed model of bigram occurrences has the form of an (observed) **contingency table**. In addition, it also counts frequencies of pairs of the bigram components x, y with words other than y and x, respectively. The contingency table contains four cells with the following counts:

$$
\begin{aligned}
f(xy) &:= |\{k : x_k = x \land y_k = y\}| \\
f(x\bar{y}) &:= |\{k : x_k = x \land y_k \neq y\}| \\
f(\bar{x}y) &:= |\{k : x_k \neq x \land y_k = y\}| \\
f(\bar{x}\bar{y}) &:= |\{k : x_k \neq x \land y_k \neq y\}|
\end{aligned}
$$

These counts are organized in the table as depicted in Table 2.1: for a given bigram xy, the counts are often also denoted also by the letters a, b, c, and d or by the letter f indexed by $i,j \in \{1, 2\}$. An example of a contingency table is shown in Table 2.2. It also illustrates how the contingency table is constructed and what types of bigrams are counted in which table cells.

Immediate contexts

Another possible approach to describe bigram occurrences is modeling occurrences of words that appear in an immediate context of the bigram, i.e. words that immediately precede or follow the bigram in the corpus. According to the second extraction principle (page 24), composition of these contexts should also, in a sense, reflect the degree of association between the bigram components.

For this purpose, we formally define the **left immediate context** C^l_{xy} and the **right immediate context** C^r_{xy} of a bigram xy as *multisets* (also called *bags of words*) whose elements are word types $\phi(w_m)$ of word tokens $w_m \in W$ that appear at a particular position before (the left context) or after (the right context) an occurrence of the bigram xy:

$$
\begin{aligned}
C^l_{xy} &= \{u_m = \phi(w_m) \colon w_m \in W \land \exists\, i,j\, (\Phi(\langle w_i, w_j \rangle) = \langle x, y \rangle \ \land \ m = i - 1)\}, \\
C^r_{xy} &= \{u_m = \phi(w_m) \colon w_m \in W \land \exists\, i,j\, (\Phi(\langle w_i, w_j \rangle) = \langle x, y \rangle \ \land \ m = i + 1)\}.
\end{aligned}
$$

	$X =$ *black*	$X \neq$ *black*	$X = *$
$Y =$ *market*	***black market***	***new market***	$*$ ***market***
$Y \neq$ *market*	***black horse***	***new horse***	$*$ ***horse***
$Y = *$	***black*** $*$	***new*** $*$	$*$ $*$

	$X =$ *black*	$X \neq$ *black*	$X = *$
$Y =$ *market*	15	38	53
$Y \neq$ *market*	654	1 330 171	1 330 825
$Y = *$	669	1 330 209	1 330 878

Table 2.2: An example of an observed contingency table for the bigram *černý trh* (*black market*). X and Y denote the first and the second bigram components, respectively. The actual frequencies contained in the bottom table refer to the occurrences of dependency bigrams in PDT.

Empirical contexts

Occurrences of bigrams (and words) can also be described by a broader **empirical context** which captures occurrences of words appearing not only in the immediate contexts but also within a longer distance from a given bigram (or a word). This approach is mainly used by lexical association measures based on the third extraction principle (page 25).

Formally, for a given word type $z \in U^*$, we define a *multiset* C_x of word types $\phi(w_m)$ of word tokens w_m from the corpus W that appear within a predefined distance (determined by a distance function d_c and a threshold t_c) from a particular occurrence of the word type z in the corpus; analogically, we define C_{xy} for a bigram type $xy \in C^*$.

$$\begin{aligned} C_x &= \{u_m = \phi(w_m) : w_m \in W \wedge \exists i\, (\phi(w_i) = x \ \wedge\ d_c(i,m) < t_c)\}, \\ C_{xy} &= \{u_m = \phi(w_m) : \\ &\qquad w_m \in W \wedge \exists i,j\, (\Phi(\langle w_i, w_j\rangle) = \langle x, y\rangle \wedge (d_c(i,m) \le t_c \vee d_c(j,m) \le t_c))\}. \end{aligned}$$

Examples of these contexts (immediate and empirical) are shown in Figures 2.1 and 2.2 on the next page. In the examples, the words are displayed as word tokens, but in fact, the contexts contain their word types.

2.2.6 Filtering candidate data

Filtering is often used to improve the precision of the extraction process by eliminating such data that does not help discover true collocations or can bias their extraction. It can be performed either *before* obtaining the occurrence statistics or *after*. Evert (2004, p. 32–33) described these two approaches as token filtering and type filtering.

. . . součástí trhu, vznikl **obratem**	**černý trh**	s plyšovými medvídky a .
zabránit přísunu drog na **domácí**	**černý trh**	v hodnotě 32 milionu . . . .
. stejnými jednotlivci i **kompletní**	**černý trh**	. Jinými slovy, byla by . . . .
. . . pomáhali pašování cigaret **na**	**černý trh**	do východního Německa.
. . . . . nájemních práv **nezaručený**	**černý trh**	. Libor Dellin, člen . . . .
. . . . . . pašovaného zboží a **kypící**	**černý trh**	jsou toho výmluvným . . .
. Také například tím, že **vznikne**	**černý trh**	, který je ke spotřebitelům
. . . . nabídku a pak **nastupuje**	**černý trh**	. Za možnost přestupu na
. . . . . Řídí gangy, které **kontrolují**	**černý trh**	a okrádají cizince. Oba . . .
. . . . nájemného " bylo a je **omezit**	**černý trh**	s byty, nestane se nic. . . . .

. . . . nějak negativně tento	**černý trh**	**naše** hospodářství? Je to . . . .
. . . . . inzeráty. Rozmohl se	**černý trh**	**bytů** a skutečné náklady na . . . . .
. . . . . jak se říká na Arbatu,	**černý trh**	**něco** do sebe. Je - li hlad . . . .
. . . . . Naplno se již rozjíždí	**černý trh**	**se** vstupenkami. Na závod . . . . . .
. . starožitnostmi měl řídit,	**černý trh**	**podporuje** na straně jedné . . . . . .
. . . Našim lidem pro samý	**černý trh**	**nezbýval** čas na sex, a tak . . . .
. . .unie však ukazují, že	**černý trh**	**překonal** stagnaci a piráti . . . .
. . . . ceny, funguje čilý	**černý trh**	**dosud**. Země bez chudých . . . . . .
. . . . novými zbraněmi. Na	**černý trh**	**odhalený** specialisty z útvaru . . .
. . . . se vlastně jedná o	**černý trh**	**s** byty. Připustil ovšem, že . . . . . .

Figure 2.1: Examples of left (at the top) and right (at the bottom) immediate contexts (not underlined words in bold) of the expression *černý trh* (*black market*) as they appear in PDT.

. . . . . oparu. Muž byl velmi malý, menší než žena. **Měl**	**černý**	**kabát se sametovým límcem.** Nevšímali si ho. Sedni
. . . . . rozsadili se kolem stolu. **Kordič si sundal sako a**	**černý**	**vlčák mu ulehl oddaně k nohám**. Po předchozím . . .
. . . zn. **Hořčák přibyl ještě tuzemský rum a činže, zel**	**černý**	**plstěný klobouk brzo prázdnotou.** Tehdy začal pan .
. . našla správnou odpověď. **Táhla za bílého a vzápětí**	**černý**	**svým posledním tahem.** V pořadí sto šedesátým . . . .
. . . Ani se o to nepokoušel. **Náhle se před ním vynořil**	**černý**	**kůň**. Na koni klidně seděla mladá policistka, světlé . .
. . . jsou bílé. **Zobák je u obou pohlaví v prostém šatě**	**černý**	**, u samice v době hnízdění žlutý.** Domovem tohoto .
. . . . v kapli. **Ruce ve volném rukávu, umělá květina a**	**černý**	**klobouk na bílém stolku.** Starý kněz vypíná pomalu
Poslanecké sněmovny. **Na budově je zároveň vyvěšen**	**černý**	**prapor.** Rozpočet armády v příštích letech vzroste . . .
. . . zdravou reakci. A pak je tu ještě smích. **Humor tak**	**černý**	**, že se můžete jen smát.** Smích je poslední výspa . . . .
. . . . ženy. **Chodily zahalené od hlavy až k patě, jejich**	**černý**	**hábit měl jen dva otvory pro oči.** Nesměly tehdy . . .

. miliónů dolarů. **Ovlivňuje nějak negativně tento**	**černý trh**	**naše hospodářství?** Je to pouze ztráta na daních . . . .
. . **Maltské liry lze nakoupit pouze ve směnárnách,**	**černý trh**	**s valutami neexistuje.** Na Maltě je v porovnání s . . .
operoval i ženu. **A přece má, jak se říká na Arbatu,**	**černý trh**	**něco do sebe.** Je - li hlad nejlepší kuchař, je . . . .
. . přestal. **V patách za krizí vstoupil do Bělehradu**	**černý trh**	**, pašování a zvýšená kriminalita.** Překupníci . . . .
. . . . . . z toho obviněni. **Řídí gangy, které kontrolují**	**černý trh**	**a okrádají cizince.** Oba byli zbaveni funkcí a byl . . . .
. drogové hysterii. **Následkem toho neexistoval ani**	**černý trh**	**, protože nebylo na čem vydělávat**. V roce 1957 bylo
. . . . . k rychlému zpracování. **Naplno se již rozjíždí**	**černý trh**	**se vstupenkami.** Na závod na 5000 m v . . . .
. . . na čelném místě obchodu se zbraněmi. **Zatímco**	**černý trh**	**se zbraněmi se pro celý svět stává čím dál tím větší.**
. . . . . v parlamentu. **Věřím, že brzy bude regulovat**	**černý trh**	**s ohroženými druhy zvířat**, míní. Promoravské . . . . .
. . . 100 tisíc korun. **Podle Piňose se vlastně jedná o**	**černý trh**	**s byty.** Připustil ovšem, že právě v případě bytového

Figure 2.2: Example of empirical contexts (not underlined words in bold) of the word *černý* (*black*) and the expression *černý trh* (*black market*) from the Prague Dependency Treebank.

f	*POS*	*bigram*	*f*	*POS*	*bigram*
2124	V:V	být mít	527	A:N	Česká republika
1815	V:R	být v	488	N:N	milión korun
1362	P:J	ten že	242	A:N	příští rok
1344	J:V	že být	221	A:N	loňský rok
1287	R:V	v být	220	A:N	životní prostředí
1196	V:P	být ten	210	A:N	letošní rok
1165	V:J	být a	190	A:N	současná doba
1010	P:V	ten být	182	N:N	ministr zahraničí
985	V:R	jít o	179	N:N	miliarda korun
973	V:J	být a	169	A:N	Spojené státy
904	J:V	a být	164	A:N	minulý týden
883	R:N	v roce	162	A:N	Evropský unie
841	V:V	být moci	156	N:N	Václav Klaus
826	V:J	být že	156	A:N	druhá strana
798	P:V	který být	156	A:N	akciová společnost
771	J:J	že a	155	N:N	návrh zákona
712	R:N	v době	155	A:N	New York
700	P:V	se stát	152	N:N	milión dolarů
675	J:R	a v	150	A:N	cenný papír
661	R:N	v případě	148	N:N	konec roku
627	V:R	být na	145	A:N	státní rozpočet
627	R:J	mezi a	142	A:N	politická strana
620	D:J	hodně než	142	A:N	akciová společnost
618	V:V	být být	141	A:N	trestný čin
618	P:V	který mít	130	A:N	hlavní město
573	J:V	že být	129	A:N	generální ředitel
560	R:P	o ten	128	A:N	poslední rok
543	V:R	mít v	126	A:N	poslední doba
542	R:J	v a	121	A:N	Komerční banka
527	A:N	Česká republika	120	N:N	Václav Havel

Table 2.3: Part-of-speech filtering: the top collocation candidates from PDT sorted by raw bigram frequency (*f*) before (left) and after (right) the filtering was applied.

Token filtering is applied before the extraction of occurrence statistics and can be understood as a set of additional constraints on the identification of bigram tokens in the set B. Token filtering affects the candidate occurrence data D and the statistics obtained from it. This step must be theoretically substantiated and must not bias the occurrence models. Appropriately designed token filtering can even improve the validity of assumptions required by certain extraction principles (e.g. the independence of randomly generated word pairs). According to Evert (2004, p. 33), it is quite adequate to restrict the bigram tokens e.g. only to *adjective-noun* combinations, if we focus only on collocations of this type, however, we cannot remove bigrams with certain general adjectives that ”usually produce uninteresting results“. Such a step would decrease marginal frequencies of nouns appearing in the affected bigrams which could unjustly prioritize other combinations of these nouns in ranking. It is quite reasonable, on the other hand, to restrict the bigram tokens only to combinations without punctuation marks.

PMI	*f*	*POS*	*bigram*
20.34	1	N:N	Čchien Čchi
20.34	1	N:N	Čaněk Gridoux
20.34	1	N:N	ČLS JEP
20.34	1	N:N	Áron Mónus
20.34	1	N:N	škodlivost narkomanie
20.34	1	N:N	šiška konifery
20.34	1	N:N	šestka Davenportová
20.34	1	N:N	Šan Čching
20.34	1	N:N	Šalom Achšav
20.34	1	N:N	Ľuba Lauffová
20.34	1	N:N	zúžení hrdla
20.34	1	N:N	zvýraznění koloritu
20.34	1	N:N	zplozenec Paynea
20.34	1	N:N	zopakování seskoku
20.34	1	N:N	znechucení naladění
20.34	1	N:N	zjevení démantu
20.34	1	N:N	zbožnost císaře
20.34	1	N:N	zavření tavírny
20.34	1	N:N	zastánce vhodu
20.34	1	N:N	zaměřování zlomek
20.34	1	N:N	zadeček Chera
20.34	1	N:N	výškař Ruffíni
20.34	1	N:N	výstřednost slavíka
20.34	1	N:N	výsev jařiny

PMI	*f*	*POS*	*bigram*
17.53	7	N:N	TTI Therm
17.53	6	N:N	Guido Reni
17.34	8	N:N	Buenos Aires
17.34	7	N:N	Monte Carlo
17.34	7	A:N	laskavé svolení
17.34	7	A:N	AG Flek
17.34	6	A:N	Tchaj wan
17.31	6	N:N	AIK Stockholm
17.17	9	N:N	Twin Peaks
17.17	9	N:N	Kazimír Jánoška
17.17	7	A:N	Geigerův čítač
17.17	6	N:N	Karol Štěpánová
17.17	6	A:N	Saudská Arábie
17.12	6	N:N	cash flow
17.02	7	A:N	Beastie Boy
16.98	7	A:N	čtvrtletní slosování
16.95	6	N:N	Kaučuk Kralupy
16.92	6	A:N	Třinecké železárny
16.88	9	N:N	tie break
16.88	9	N:N	Four Seasons
16.88	7	A:N	kochleární implantát
16.88	6	N:N	Saccheriho čtyřúhelník
16.88	6	N:N	José Carreras
16.88	6	N:N	Baruch Goldstein
16.85	8	A:N	clearingové zúčtování

Table 2.4: Frequency filtering: the top collocation candidates from PDT sorted by *Pointwise mutual information)* (*PMI*) before (left) and after (right) the filtering was applied.

Type filtering is applied after the extraction of occurrence statistics and has no effect on the candidate occurrence data D and the extracted statistics. It divides the collocation candidates into subsets which are then handled separately. A typical case of type filtering is the commonly used **part-of-speech filtering** based on the morphological information obtained during linguistic preprocessing (see e.g. Justeson and Katz, 1995; Manning and Schütze, 1999; Evert, 2004). With the knowledge of morphological characteristics of collocation candidates and their components, we can identify those that are not very likely to form collocations, and exclude them from further analysis. They can be explicitly classified as *non-collocations* or, in the case of ranking, placed at the end of the list or discarded entirely.

The effect of type filtering is illustrated in Table 2.3. It shows the top 20 collocation candidates from PDT, ranked by bigram frequency obtained before part-of-speech filtering (on the left), and the top 20 candidates from the same set, obtained after the filter was applied, where only *adjective-noun* and *noun-noun* combinations were kept. The first table contains only one true collocation *Česká republika*, which appears at the very bottom of the list (*Czech Republic*). After the application of the filter, almost all the top candidates, as they appear in the other table, can be considered collocations.

Another case of type filtering is **frequency filtering**. It is based on setting a limit on the minimal frequency of collocation candidates before association measures are applied. It is a well-known fact that many association measures are unreliable when applied to low-frequency data and that certain minimal frequency is required in order to expect meaningful results. This issue was thoroughly studied by Evert in his thesis (Evert, 2004) where he demonstrated that "it is impossible *in principle* to compute meaningful association scores for the lowest-frequency data" (p. 22, 95–108).

The effect of frequency filtering is illustrated in Table 2.4. The top positions in the list of collocation candidates from PDT, ranked according to scores of *Pointwise mutual information*, are occupied by bigrams whose components appear in PDT just once, that is, in this bigram. There is no way to distinguish between collocations and non--collocations in this list – from the perspective of statistics, they have the same properties (occurrence frequency) and cannot be differentiated. The top candidates obtained after applying the frequency filter that discarded candidates occurring 5 times or less is shown on the right – almost all of them can be considered collocations.

Context filtering is a special case of filtering that can be employed during the construction of empirical contexts. These structures are intended for modeling the semantics of collocation candidates and their components (see the third extraction principle in Section 2.2.1). The way they are defined in Section 2.2.5 implies that they contain types of all word tokens occurring within specified context windows which also includes words with a little or no semantic content that do not determine meaning of a given bigram or word. In empirical contexts, such word tokens can be ignored. This idea, however, cannot be applied to immediate contexts that model an immediate word environment from an information-theoretical point of view, and therefore the occurrence of all word tokens should be taken into account.

3
Association Measures

The last step of the extraction pipeline involves applying a chosen lexical association measure to the occurrence and context statistics extracted from the corpus for all collocation candidates and obtaining their association scores. A list of the candidates ranked according to their association scores is the desired result of the entire process.

In this chapter, we introduce an inventory of 82 such lexical association measures. These measures are based on the extraction principles described in Section 2.2.1 which correspond to the three basic approaches to determine collocational association: by measuring the *statistical association* between the components of the collocation candidates, by measuring the *quality of context* of the collocation candidates, and by measuring the *dissimilarity of contexts* of the collocation candidates and their components.

For each of these approaches, we first present its mathematical foundations and then a list of the relevant measures including their formulas and key references. We do not discuss each of the measures in detail. An exhaustive description of many of these measures (applied to collocation extraction) was published in the dissertation of Evert (2004). A general description (not applied to collocation extraction) of other measures can be found in the thesis of Warrens (2008) or in the provided references.

3.1 Statistical association

In order to measure the statistical association, the candidate occurrence data D extracted from the corpus is interpreted as a **random sample** obtained by sampling (with replacement) from the (unknown) population of all possible bigram types $xy \in C^*$. The random sample consists of N realizations (observed values) of a pair of discrete random variables $\langle X, Y \rangle$ representing the component types $x, y \in U^*$. The population is characterized by the **occurrence probability** (also called **joint probability**) of the bigram types:

$$P(xy) := P(X = x \wedge Y = y).$$

The probabilities $P(X = x)$ and $P(Y = y)$ of the components types x and y are called the **marginal probabilities** and can be computed from the joint probabilities as:

$$P(x*) := P(X = x) = \sum_{y'} P(X = x \wedge Y = y'),$$
$$P(*y) := P(Y = y) = \sum_{x'} P(X = x' \wedge Y = y).$$

$P(xy) =: P_{11}$	$P(x\bar{y}) =: P_{12}$	$P(x*) =: P_1$
$P(\bar{x}y) =: P_{21}$	$P(\bar{x}\bar{y}) =: P_{22}$	$P(\bar{x}*)$
$P(*y) =: P_2$	$P(*\bar{y})$	N

Table 3.1: A contingency table of the probabilities associated with a bigram xy.

Similarly to the occurrence frequencies, the population can also be described by the following probabilities organized into a contingency table (Table 3.1):

$$
\begin{aligned}
P(xy) &:= P(X{=}\,x \wedge Y{=}\,y)\\
P(x\bar{y}) &:= P(X{=}\,x \wedge Y{\neq}\,y) = \sum_{y'\neq y} P(X{=}\,x \wedge Y{=}\,y'),\\
P(\bar{x}y) &:= P(X{\neq}\,x \wedge Y{=}\,y) = \sum_{x'\neq x} P(X{=}\,x' \wedge Y{=}\,y),\\
P(\bar{x}\bar{y}) &:= P(X{\neq}\,x \wedge Y{\neq}\,y) = \sum_{x'\neq x, y'\neq y} P(X{=}\,x' \wedge Y{=}\,y')
\end{aligned}
$$

These probabilities are considered *unknown* parameters of the population. Any inferences concerning these parameters can be made only on the basis of the observed frequencies obtained from the random sample D.

In order to estimate values of these probabilities for each bigram separately, we introduce random variables F_{ij}, $i,j \in \{1,2\}$ that correspond to the values in the observed contingency table of a given bigram xy as depicted in Table 3.2. These random variables are defined as the number of successes in a sequence of N independent experiments (Bernoulli trials) that determine whether a particular bigram type (xy, $x\bar{y}$, $\bar{x}y$, or $\bar{x}\bar{y}$) occurs or not, and where each experiment yields success with probability P_{ij}. The observed values of a contingency table $\langle f_{11}, f_{12}, f_{21}, f_{22}\rangle$ can be interpreted as the realization of the random variables $\langle F_{11}, F_{12}, F_{21}, F_{22}\rangle$ denoted by $\boldsymbol{F}$. Their joint distribution is a **multinomial distribution** with parameters $N, P_{11}, P_{12}, P_{21}$, and P_{22}:

$$\boldsymbol{F} \sim \mathrm{Multi}(N, P_{11}, P_{12}, P_{21}, P_{22}).$$

The probability of an observation of the values $f_{11}, f_{12}, f_{21}, f_{22}$, where $\sum f_{ij} = N$, is:

$$P(F_{11}{=}f_{11} \wedge F_{12}{=}f_{12} \wedge F_{21}{=}f_{21} \wedge F_{22}{=}f_{22}) = \frac{N!}{f_{11}!f_{12}!f_{21}!f_{22}!} \cdot P_{11}^{f_{11}} \cdot P_{12}^{f_{12}} \cdot P_{21}^{f_{21}} \cdot P_{22}^{f_{22}}.$$

Each random variable F_{ij} has then a **binomial distribution** with parameters (N, P_{ij}):

$$F_{ij} \sim \mathrm{Bi}(N, P_{ij}).$$

	$X = x$	$X \neq x$
$Y = y$	F_{11}	F_{12}
$Y \neq y$	F_{21}	F_{22}

Table 3.2: Random variables representing event frequencies in a contingency table.

The probability of observing the value f_{ij} is for these variables defined by the formula:

$$P(F_{ij}{=}f_{ij}) = \binom{N}{f_{ij}} P_{ij}^{f_{ij}} \, (1 - P_{ij})^{N-f_{ij}}.$$

The expected value and variance for binomially distributed variables are defined as:

$$E(F_{ij}) = NP_{ij}, \quad Var(F_{ij}) = NP_{ij}(1 - P_{ij}).$$

In the same manner, we can introduce random variables $F_i, i \in \{1, 2\}$ representing the marginal frequencies f_1, f_2 that have binomial distribution with the parameters N and P_1, P_2, respectively. Under the binomial distribution of F_{ij}, the **maximum-likelihood estimates** of the population parameters P_{ij} that maximize the probability of the data (the observed contingency table) are defined as:

$$p_{11} := \frac{f_{11}}{N} \approx P_{11}, \qquad p_{21} := \frac{f_{21}}{N} \approx P_{21},$$
$$p_{12} := \frac{f_{12}}{N} \approx P_{12}, \qquad p_{22} := \frac{f_{22}}{N} \approx P_{22}.$$

And analogically, the maximum-likelihood estimates of the marginal probabilities are:

$$p_1 := \frac{f_1}{N} \approx P_1 \qquad p_2 := \frac{f_2}{N} \approx P_2$$

The last step to measuring statistical association is to define this concept by the notion of **statistical independence**. We say that there is *no* statistical association between the components of a bigram type if the occurrence of one component has *no* influence on the occurrence of the other one, i.e. the occurrences of the components (as random events) are statistically independent.

In the terminology of statistical hypothesis testing, this can be formulated as the **null hypothesis of independence** H_0 where the probability of observing the components together (as a bigram) is just the product of their marginal probabilities:

$$H_0: \quad P = P_1 \cdot P_2$$

We are then interested in those bigram types (collocation candidates) for which this hypothesis can be (based on the evidence obtained from the random sample) **rejected**

$\widehat{f}(xy) =: \widehat{f}_{11}$	$\widehat{f}(x\bar{y}) =: \widehat{f}_{12}$	$\widehat{f}(x*) =: \widehat{f}_1$
$\widehat{f}(\bar{x}y) =: \widehat{f}_{21}$	$\widehat{f}(\bar{x}\bar{y}) =: \widehat{f}_{22}$	$\widehat{f}(\bar{x}*)$
$\widehat{f}(*y) =: \widehat{f}_2$	$\widehat{f}(*\bar{y})$	N

Table 3.3: Expected contingency table frequencies of a bigram xy (under the null hypothesis).

in favor of the **alternative hypothesis** H_1 stating the observed bigram occurrences have not resulted from random chance:

$$H_1: \quad P \neq P_1 \cdot P_2$$

With the maximum-likelihood estimates $p_1 \approx P_1$ and $p_2 \approx P_2$, we can determine the probabilities P_{ij} under the null hypothesis H_0 as:

$$\begin{aligned} H_0: \quad P_{11} &= p_1 \cdot p_2, \\ P_{12} &= p_1 \cdot (1-p_2), \\ P_{21} &= (1-p_1) \cdot p_2, \\ P_{21} &= (1-p_1) \cdot (1-p_2). \end{aligned}$$

Consequently, the expected values of the variables F_{ij} that form the **expected contingency table** under the null hypothesis H_0 (Table 3.3) are:

$$\begin{aligned} H_0: \quad E(F_{11}) &= \frac{f_1 \cdot f_2}{N} =: \widehat{f}_{11}, & E(F_{12}) &= \frac{f_1 \cdot (N-f_2)}{N} =: \widehat{f}_{12}, \\ E(F_{21}) &= \frac{(N-f_1) \cdot f_2}{N} =: \widehat{f}_{21}, & E(F_{22}) &= \frac{(N-f_1) \cdot (N-f_2)}{N} =: \widehat{f}_{22}. \end{aligned}$$

There are various approaches that can be employed for testing the null hypothesis of independence. **Test statistics** calculate the probability (p-value) that the observed values (frequencies) would occur if the null hypothesis were true. If the p-value is too low (beneath a significance level α, typically set to 0.05), the null hypothesis is rejected in favor of the alternative hypothesis (at the significance level α) and held as possible otherwise. In other words, the tests compare the observed values (frequencies) with those that are expected under the null hypothesis and if the difference is too large, the null hypothesis is rejected (again at the significance level α). However, the test statistics are more useful as methods for determining the strength of association (the level of significance is ignored) and their scores are directly used as the association scores for ranking. The statistical association measures base on statistical tests are *Pearson's χ^2 test (10)*, *Fisher's exact test (11)*, *t-test (12)*, *z score (13)*, and *Poisson significance (14)* (the numbers in parentheses refer to Table 3.4).

More interpretable are **likelihood ratios** that simply express how much more likely one hypothesis is than the other (H_0 vs. H_1). These ratios can also be employed to test the null hypothesis in order to attempt rejecting it (at the significance level α) or not, but it is more useful to use them directly to compute the association scores for ranking, e.g. *Log likelihood ratio (15).*

Various other measures have been proposed to determine the statistical association of two events (and its strength). Although they originate in all sorts of fields (e.g. information theory) and are based on various principles (often heuristic), they can be successfully used for measuring lexical association. All the statistical association measures are presented in Table 3.4.

3.2 Context analysis

The second and the third extraction principle, described in Section 2.2.1, deal with the concept of **context**. Generally, a context is defined as a multiset (bag) of word types occurring within a predefined distance (also called a **context window**) from any occurrence of a given bigram type or word type (their tokens, more precisely) in the corpus. The main idea of using this concept is to model the **average context** of an occurrence of the bigram/word type in the corpus, i.e. word types that *typically* occur in its neighborhood.

In this work, we employ two approaches representing the average context: by estimating the **probability distribution** of word types appearing in such a neighborhood and by the **vector space model** adopted from the field of information retrieval.

The four specific context types used in this work are formally defined on page 32. In the following sections, we use C_e to denote the context of an event e (occurrence of a bigram type xy or a word type z) of any of those types (left/right immediate context or empirical context). For simplicity of notation, elements of C_e are denoted by z_k:

$$C_e = \{z_k : z_k \in \{1, \dots, M\}\}, \quad M = |C_e|, \quad C_e \in \left\{C_{xy}^l, C_{xy}^r, C_x, C_{xy}\right\}.$$

Probability distribution estimation

In order to estimate the **probability distribution** $P(Z|C_e)$ of word types z appearing in the context C_e, this multiset is interpreted as a **random sample** obtained by sampling (with replacement) from the population of all possible (basic) word types $z \in U$. The random sample consists of M realizations of a (discrete) random variable Z representing the word type appearing in the context C_e. The population parameters are the **context occurrence probabilities** of the word types $z \in U$.

$$P(z|C_e) := P(Z = z|C_e).$$

These parameters can be estimated on the basis of the observed frequencies of word types $z \in U$ obtained from the random sample C_e by the following formula:

$$f(z|C_e) = |\{k : z_k \in C_e \ \wedge \ z_k = z\}|.$$

#	*name*	*formula*	*reference*
1.	**Joint probability**	$p(xy)$	(Giuliano, 1964)
2.	**Conditional probability**	$p(y\mid x)$	(Gregory et al., 1999)
3.	**Reverse cond. probability**	$p(x\mid y)$	(Gregory et al., 1999)
4.	**Pointwise mutual inf. (*MI*)**	$\log \frac{p(xy)}{p(x*)p(*y)}$	(Church and Hanks, 1990)
5.	**Mutual dependency (*MD*)**	$\log \frac{p(xy)^2}{p(x*)p(*y)}$	(Thanopoulos et al., 2002)
6.	**Log frequency biased *MD***	$\log \frac{p(xy)^2}{p(x*)p(*y)} + \log p(xy)$	(Thanopoulos et al., 2002)
7.	**Normalized expectation**	$\frac{2f(xy)}{f(x*)+f(*y)}$	(Smadja and McKeown, 1990)
8.	**Mutual expectation**	$\frac{2f(xy)}{f(x*)+f(*y)} \cdot p(xy)$	(Dias et al., 2000)
9.	**Salience**	$\log \frac{p(xy)^2}{p(x*)p(*y)} \cdot \log f(xy)$	(Kilgarriff and Tugwell, 2001)
10.	**Pearson's χ^2 test**	$\sum_{i,j} \frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$	(Manning and Schütze, 1999)
11.	**Fisher's exact test**	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$	(Pedersen, 1996)
12.	**t test**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$	(Church and Hanks, 1990)
13.	**z score**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$	(Berry-Rogghe, 1973)
14.	**Poisson significance**	$\frac{\hat{f}(xy)-f(xy)\log \hat{f}(xy)+\log f(xy)!}{\log N}$	(Quasthoff and Wolff, 2002)
15.	**Log likelihood ratio**	$-2\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$	(Dunning, 1993)
16.	**Squared log likelihood ratio**	$-2\sum_{i,j} \frac{\log f_{ij}^2}{\hat{f}_{ij}}$	(Inkpen and Hirst, 2002)
17.	**Russel-Rao**	$\frac{a}{a+b+c+d}$	(Russel and Rao, 1940)
18.	**Sokal-Michiner**	$\frac{a+d}{a+b+c+d}$	(Sokal and Michener, 1958)
19.	**Rogers-Tanimoto**	$\frac{a+d}{a+2b+2c+d}$	(Rogers and Tanimoto, 1960)
20.	**Hamann**	$\frac{(a+d)-(b+c)}{a+b+c+d}$	(Hamann, 1961)
21.	**Third Sokal-Sneath**	$\frac{b+c}{a+d}$	(Sokal and Sneath, 1963)
22.	**Jaccard**	$\frac{a}{a+b+c}$	(Jaccard, 1912)
23.	**First Kulczynsky**	$\frac{a}{b+c}$	(Kulczynski, 1927)
24.	**Second Sokal-Sneath**	$\frac{a}{a+2(b+c)}$	(Sokal and Sneath, 1963)
25.	**Second Kulczynski**	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$	(Kulczynski, 1927)
26.	**Fourth Sokal-Sneath**	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$	(Kulczynski, 1927)
27.	**Odds ratio**	$\frac{ad}{bc}$	(Tan et al., 2002)
28.	**Yulle's ω**	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	(Tan et al., 2002)
29.	**Yulle's Q**	$\frac{ad-bc}{ad+bc}$	(Tan et al., 2002)
30.	**Driver-Kroeber**	$\frac{a}{\sqrt{(a+b)(a+c)}}$	(Driver and Kroeber, 1932)

# name	*formula*	*reference*
31. **Fifth Sokal-Sneath**	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	(Sokal and Sneath, 1963)
32. **Pearson**	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	(Pearson,1950)
33. **Baroni-Urbani**	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	(Baroni-Urbani and Buser, 1976)
34. **Braun-Blanquet**	$\frac{a}{\max(a+b,a+c)}$	(Braun-Blanquet, 1932)
35. **Simpson**	$\frac{a}{\min(a+b,a+c)}$	(Simpson, 1943)
36. **Michael**	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	(Michael, 1920)
37. **Mountford**	$\frac{2a}{2bc+ab+ac}$	(Kaufman and Rousseeuw, 1990)
38. **Fager**	$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2}\max(b,c)$	(Kaufman and Rousseeuw, 1990)
39. **Unigram subtuples**	$\log\frac{ad}{bc} - 3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$	(Blaheta and Johnson, 2001)
40. ***U* cost**	$\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$	(Tulloss, 1997)
41. ***S* cost**	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$	(Tulloss, 1997)
42. ***R* cost**	$\log(1+\frac{a}{a+b})\cdot\log(1+\frac{a}{a+c})$	(Tulloss, 1997)
43. ***T* combined cost**	$\sqrt{U\times S\times R}$	(Tulloss, 1997)
44. **Phi**	$\frac{p(xy)-p(x*)p(*y)}{\sqrt{p(x*)p(*y)(1-p(x*))(1-p(*y))}}$	(Tan et al., 2002)
45. **Kappa**	$\frac{p(xy)+p(\bar{x}\bar{y})-p(x*)p(*y)-p(\bar{x}*)p(*\bar{y})}{1-p(x*)p(*y)-p(\bar{x}*)p(*\bar{y})}$	(Tan et al., 2002)
46. ***J* measure**	$\max[p(xy)\log\frac{p(y\|x)}{p(*y)}+p(x\bar{y})\log\frac{p(\bar{y}\|x)}{p(*\bar{y})},$ $p(xy)\log\frac{p(x\|y)}{p(x*)}+p(\bar{x}y)\log\frac{p(\bar{x}\|y)}{p(\bar{x}*)}]$	(Tan et al., 2002)
47. **Gini index**	$\max[p(x*)(p(y\|x)^2+p(\bar{y}\|x)^2)-p(*y)^2$ $+p(\overline{x*})(p(y\|\bar{x})^2+p(\bar{y}\|\bar{x})^2)-p(*\bar{y})^2,$ $p(*y)(p(x\|y)^2+p(\bar{x}\|y)^2)-p(x*)^2$ $+p(*\bar{y})(p(x\|\bar{y})^2+p(\bar{x}\|\bar{y})^2)-p(\bar{x}*)^2]$	(Tan et al., 2002)
48. **Confidence**	$\max[p(y\|x),p(x\|y)]$	(Tan et al., 2002)
49. **Laplace**	$\max[\frac{Np(xy)+1}{Np(x*)+2},\frac{Np(xy)+1}{Np(*y)+2}]$	(Tan et al., 2002)
50. **Conviction**	$\max[\frac{p(x*)p(*y)}{p(x\bar{y})},\frac{p(\bar{x}*)p(*y)}{p(\bar{x}y)}]$	(Tan et al., 2002)
51. **Piatersky-Shapiro**	$p(xy)-p(x*)p(*y)$	(Tan et al., 2002)
52. **Certainity factor**	$\max[\frac{p(y\|x)-p(*y)}{1-p(*y)},\frac{p(x\|y)-p(x*)}{1-p(x*)}]$	(Tan et al., 2002)
53. **Added value (*AV*)**	$\max[p(y\|x)-p(*y),p(x\|y)-p(x*)]$	(Tan et al., 2002)
54. **Collective strength**	$\frac{p(xy)+p(\bar{x}\bar{y})}{p(x*)p(y)+p(\bar{x}*)p(*y)}\cdot\frac{1-p(x*)p(*y)-p(\bar{x}*)p(*y)}{1-p(xy)-p(\bar{x}\bar{y})}$	(Tan et al., 2002)
55. **Klosgen**	$\sqrt{p(xy)}\cdot AV$	(Tan et al., 2002)

Table 3.4: Statistical association measures.

We introduce a random variable F that represents the observed frequencies of word types in the context C_e which has a **binomial distribution** with parameters M and P. The probability of observing the value f for the binomial distribution with these parameters is defined as:

$$P(F{=}f) = \binom{M}{f} P^f \, (1-P)^{M-f}, \qquad F \sim Bi(M,P).$$

Under the binomial distribution of F, the **maximum-likelihood estimates** of the population parameters P that maximize the probability of the observed frequencies are:

$$p(z|C_e) := \frac{f(z|C_e)}{M} \;\approx\; P(z|C_e)$$

Having estimated the probabilities of word types occurring within the context of collocation candidates and their components, we can compute the association scores of measures based on the second and third extraction principles, such as entropy, cross entropy, divergence, and distance of these contexts, such as measures *56–62* and *63–76* in Table 3.5.

Vector space model

The **vector space model model** (Salton et al., 1975; van Rijsbergen, 1979; Baeza-Yates and Ribeiro-Neto, 1999) is a mathematical model used in information retrieval and related areas for representing text documents as vectors of *terms*. Each dimension of the vector corresponds to a separate term. The value of the term in the vector corresponds to its weight in the document: if the term appears in the document, its weight is greater than zero. In our case, the document is a context and the terms are the word types from the set of all possible word types U.

Formally, for a context C_e, we define its vector model $\mathbf{c}_e$ as the vector of **term weights** ω_{l,C_e}, where $l = 1,\ldots,|U|$. The value of ω_{l,C_e} then represents the weight of the word type u_l in the context C_e.

$$\mathbf{c}_e = \left\langle \omega_{1,C_e}, \ldots, \omega_{|U|,C_e} \right\rangle .$$

Several different techniques for computing term weights have been proposed. In this work, we employ three of the most common ones:

In the **boolean model**, the weights have boolean values {0, 1} and simply indicate if a term appears in the context or not. If the term occurs in the context at least once, its weight is 1 and 0 otherwise.

$$\omega_{l,C_e} := I(u_l, C_e), \qquad\qquad I(u_l, C_e) := \begin{cases} 1 & \text{if} \quad f(u_l|C_e) > 0, \\ 0 & \text{if} \quad f(u_l|C_e) = 0. \end{cases}$$

The **term frequency model** (TF) is equivalent to the context probability distribution and the term weights are computed as normalized occurrence frequencies. This approach should reflect how important the term is for the context – its importance increases proportionally to the number of times the term appears in the context.

$$\omega_{l,C_e} := \mathrm{TF}(u_l, C_e), \qquad \mathrm{TF}(u_l, C_e) := \frac{f(u_l|C_e)}{M}$$

The **term frequency-document frequency model** (TF-IDF) weights terms not only by their importance in the actual context but also by their importance in other contexts. The formula for computing term weights consists of two parts: term frequency is the same as in the previous case and document frequency counts all contexts where the term appears. C_e' denotes any context of the same type as C_e.

$$\omega_{l,C_e} := \mathrm{TF}(u_l, C_e) \cdot \mathrm{IDF}(u_l), \qquad \mathrm{IDF}(u_l) := \log \frac{|\{C_e'\}|}{|\{C_e' \colon u_l \in C_e'\}|}$$

The numerator in the IDF part of the formula is the total number of contexts of the same type as C_e. The denominator corresponds to the number of contexts of the same type as C_e containing u_l.

Any of the specified models can be used for quantifying similarity between two contexts by comparing their vector representations. Several techniques have been proposed, e.g. *Jaccard, Dice, Cosine* (Frakes and Baeza-Yates, 1992) but in our work, we employ two of the most popular ones:

The **cosine similarity** computes the cosine of the angle between the vectors. The numerator is the inner product of the vectors, and the denominator is the product of their lengths, thus normalizing the context vectors:

$$\cos(\mathbf{c}_x, \mathbf{c}_y) := \frac{\mathbf{c}_x \cdot \mathbf{c}_y}{||\mathbf{c}_x|| \cdot ||\mathbf{c}_y||} = \frac{\sum \omega_{l,x}\, \omega_{l,y}}{\sqrt{\sum {\omega_{l,x}}^2} \cdot \sqrt{\sum {\omega_{l,y}}^2}}.$$

The **dice similarity** computes a similarity score on the basis of the formula given bellow. It is also based on the inner product but the normalizing factor is the average quadratic length of the two vectors:

$$\mathrm{dice}(\mathbf{c}_x, \mathbf{c}_y) := \frac{2\, \mathbf{c}_x \cdot \mathbf{c}_y}{||\mathbf{c}_x||^2 + ||\mathbf{c}_y||^2} = \frac{2 \sum \omega_{l,x}\, \omega_{l,y}}{\sum {\omega_{l,x}}^2 + \sum {\omega_{l,y}}^2}$$

These techniques combined with the different vector models are the basis of association measures comparing empirical contexts of collocation candidates and their components, such as measures *63–82* in Table 3.5.

# *name*	*formula*	*reference*
56. **Context entropy**	$-\sum_z p(z \mid C_{xy}) \log p(z \mid C_{xy})$	(Krenn, 2000)
57. **Left context entropy**	$-\sum_z p(z \mid C^l_{xy}) \log p(z \mid C^l_{xy})$	(Shimohata et al., 1997)
58. **Right context entropy**	$-\sum_z p(z \mid C^r_{xy}) \log p(z \mid C^r_{xy})$	(Shimohata et al., 1997)
59. **Left context divergence**	$p(x*) \log p(x*) - \sum_z p(z \mid C^l_{xy}) \log p(z \mid C^l_{xy})$	
60. **Right context divergence**	$p(*y) \log p(*y) - \sum_z p(z \mid C^r_{xy}) \log p(z \mid C^r_{xy})$	
61. **Cross entropy**	$-\sum_z p(z \mid C_x) \log p(z \mid C_y)$	(Cover and Thomas, 1991)
62. **Reverse cross entropy**	$-\sum_z p(z \mid C_y) \log p(z \mid C_x)$	(Cover and Thomas, 1991)
63. **Intersection measure**	$\frac{2\lvert C_x \cap C_y \rvert}{\lvert C_x \rvert + \lvert C_y \rvert}$	(Lin, 1998)
64. **Euclidean norm**	$\sqrt{\sum_z (p(z \mid C_x) - p(z \mid C_y))^2}$	(Lee, 2001)
65. **Cosine norm**	$\frac{\sum_z p(z \mid C_x) p(z \mid C_y)}{\sum_z p(z \mid C_x)^2 \cdot \sum_z p(z \mid C_y)^2}$	(Lee, 2001)
66. ***L1* norm**	$\sum_z \lvert p(z \mid C_x) - p(z \mid C_y) \rvert$	(Dagan et al., 1999)
67. **Confusion probability**	$\sum_z \frac{p(x \mid C_z) p(y \mid C_z) p(z)}{p(x*)}$	(Dagan et al., 1999)
68. **Reverse confusion prob.**	$\sum_z \frac{p(y \mid C_z) p(x \mid C_z) p(z)}{p(*y)}$	
69. **Jensen-Shannon divergence**	$\frac{1}{2}[D(p(Z \mid C_x) \Vert \frac{1}{2}(p(Z \mid C_x) + p(Z \mid C_y)))$	(Dagan et al., 1999)
	$+D(p(Z \mid C_y) \Vert \frac{1}{2}(p(Z \mid C_x) + p(Z \mid C_y)))]$	
70. **Cosine of pointwise *MI***	$\frac{\sum_z MI(z,x) MI(z,y)}{\sqrt{\sum_z MI(z,x)^2} \cdot \sqrt{\sum_z MI(z,y)^2}}$	
71. **KL divergence**	$\sum_z p(z \mid C_x) \log \frac{p(z \mid C_x)}{p(z \mid C_y)}$	(Dagan et al., 1999)
72. **Reverse KL divergence**	$\sum_z p(z \mid C_y) \log \frac{p(z \mid C_y)}{p(z \mid C_x)}$	
73. **Skew divergence**	$D(p(Z \mid C_x) \Vert \alpha\, p(Z \mid C_y) + (1 - \alpha)\, p(Z \mid C_x))$	(Lee, 2001)
74. **Reverse skew divergence**	$D(p(Z \mid C_y) \Vert \alpha\, p(Z \mid C_x) + (1 - \alpha)\, p(Z \mid C_y))$	
75. **Phrase word coocurrence**	$\frac{1}{2}(\frac{f(x \mid C_{xy})}{f(xy)} + \frac{f(y \mid C_{xy})}{f(xy)})$	(Zhai, 1997)
76. **Word association**	$\frac{1}{2}(\frac{f(x \mid C_y) - f(xy)}{f(xy)} + \frac{f(y \mid C_x) - f(xy)}{f(xy)})$	(Zhai, 1997)
Cosine context similarity:	$\frac{1}{2}(\cos(\mathbf{c}_x, \mathbf{c}_{xy}) + \cos(\mathbf{c}_y, \mathbf{c}_{xy}))$	(Frakes, Baeza-Yates,1992)
77. **in boolean vector space**	$\omega_{l,C_e} = I(u_l, C_e)$	
78. **in** TF **vector space**	$\omega_{l,C_e} = \mathrm{TF}(u_l, C_e)$	
79. **in** TF·IDF **vector space**	$\omega_{l,C_e} = \mathrm{TF}(u_l, C_e) \cdot \mathrm{IDF}(u_l)$	
Dice context similarity:	$\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x, \mathbf{c}_{xy}) + \mathrm{dice}(\mathbf{c}_y, \mathbf{c}_{xy}))$	(Frakes, Baeza-Yates,1992)
80. **in boolean vector space**	$\omega_{l,C_e} = I(u_l, C_e)$	
81. **in** TF **vector space**	$\omega_{l,C_e} = \mathrm{TF}(u_l, C_e)$	
82. **in** TF·IDF **vector space**	$\omega_{l,C_e} = \mathrm{TF}(u_l, C_e) \cdot \mathrm{IDF}(u_l)$	

Table 3.5: Context-based association measures.

4
Reference Data

Gold-standard reference data is absolutely essential for empirical evaluation. For many tasks of computational linguistics and natural language processing (such as machine translation or word sense disambiguation), standard and well designed reference data sets are widely available for evaluation and development purposes, often developed for various shared task evaluation campaigns. Since this has not been the case for the task of collocation extraction (at the time of starting of this work) we decided to develop a complete *testbed* of our own. In the following sections, we describe requirements imposed on the reference data, the source corpora the data was extracted from, and the actual reference data sets we created and used in our experiments.

The main set of our experiments was conducted on the Czech *Prague Dependency Treebank*, a medium-sized corpus featuring manual morphological and syntactic annotation. In additional experiments, we used the *Czech National Corpus*, a much larger data automatically processed by a part-of-speech tagger. In order to compare the results with experiments on a different language, we also carried out some experiments on the Swedish *PAROLE* corpus provided with automatic part-of-speech tagging.

4.1 Requirements

With respect to the nature of the task (defined as ranking collocation candidates; see Chapter 2), and the evaluation scheme (based on precision and recall; see Chapter 5) the reference data should be composed of a set of collocation candidates indicated (annotated) as *true collocations* and *false collocations* (non-collocations). The design and development of the reference data is thus influenced by two main factors: 1) how and from where to extract the candidate data and 2) how to perform the annotation.

4.1.1 Candidate data extraction

When choosing the source corpus and preparing the candidate data for annotation, we considered the following requirements (or recommendations):

1. Czech, similarly to many other languages, has very complex morphology. Appropriate morphological normalization is required to conflate all morphological variants of individual collocation candidates so all occurrences of a collocation candidate in the source corpus are correctly recognized regardless of their actual surface forms.

2. According to our notion of collocation (see Section 2.1.5), collocations are grammatically bounded. Syntactic information is required to identify collocation candidates solely as syntactic units (and not as other non-syntactic word combinations). Also, each occurrence of a collocation candidate must be correctly recognized regardless of the actual word order of its components.
3. In order to minimize the bias caused by underlying linguistic data preprocessing (such as part-of-speech tagging, lemmatization, and parsing) the source corpus should be provided with manual morphological and syntactic annotation.
4. Most of the extraction methods assume normal distribution of observations or become unreliable when dealing with rare events for other reasons (see Chapter 3). The source corpus must be large enough to provide enough occurrence evidence for sufficient numbers of collocation candidates.
5. Ideally, the annotation should be performed on a full candidate data extracted from the corpus (e.g. all occurring n-grams) to avoid sampling (taking only a subset of the full data) and potential problems with estimating performance over the full data based on the sample estimation.
6. The amount of collocation candidates must be small enough that the annotation process is feasible for a human annotator(s), and at the same time large enough to provide good and reliable estimation of the performance scores.

4.1.2 Annotation process

The annotation process should result in a set of collocation candidates, each reliably judged either as a *true collocation* or as a *false collocation*. The entire procedure must follow a-priori established guidelines covering the following points:

1. Clear and exact definition of annotated phenomena must be provided. All the participating annotators must share the same notion of these phenomena and be able to achieve maximum agreement.
2. Subjectivity and other factors play an important role in the notion of collocation and have a negative influence on the annotation quality. The annotation should be performed independently by multiple annotators in parallel in order to estimate the output quality and to minimize the subjectivity of the work by combining annotators' judgments.
3. There are several possible scenarios how to combine multiple annotators' outcomes: at least one positive judgment required, taking a majority vote, full agreement required etc. The most appropriate approach should be considered with respect to the nature of the annotated phenomena.
4. During annotation, annotators can assess each occurrence of a collocation candidate as a *token* with complete knowledge of its current context, or judge each candidates as a *type* independently on its occurrences and without contextual information assuming that all occurrences would share the same annotation.

4.2 Prague Dependency Treebank

To accomplish all requirements imposed in the previous section, we chose the Prague Dependency Treebank 2.0 (PDT) as the source corpus of our candidate data. It is a moderate sized corpus provided with manual morphological and syntactic annotation and by focusing only on two-word collocations, PDT provides sufficient evidence of observations for a sound evaluation. By default, the data is divided into training, development, and evaluation sets (e.g. for the purposes of part-of-speech tagging, parsing, etc.). We ignored this split and used all data annotated on the morphological and analytical layer – a total of 1 504 847 tokens in 87 980 sentences and 5 338 documents.

4.2.1 Treebank details

The Prague Dependency Treebank has been developed by the Institute of Formal and Applied Linguistics and the Center for Computational Linguistics, Charles University, Prague[1] and it is available from Linguistic Data Consortium[2] (catalog number LDC2006T01). It contains a large amount of Czech texts (comprising samples from daily newspapers, a weekly business magazine, and a popular scientific magazine) with complex and interlinked annotation on morphological, analytical (surface syntax), and tectogrammatical (deep syntax) layer. The annotation is based on the long--standing Praguian linguistic tradition, adapted for the current computational linguistics research needs.[3]

Morphological layer

On the morphological layer, each word form (token) is assigned a **lemma** and a **morphological tag**. Combination of the lemma and the tag uniquely identifies the word form. Two different word forms differ either in their lemmas or in morphological tags.

A lemma has two parts. The first part, the **lemma proper**, is a unique identifier of the lexical item. Usually it is the **base form** of the word (e.g. first case singular for nouns, infinitive for verbs, etc.), possibly followed by a number distinguishing different lemmas with the same base forms (different word senses). The second part is optional. It contains additional information about the lemma (e.g. semantic or derivational information). A morphological tag is a string of 15 characters where every position encodes one morphological category using one character. Description of the categories and range of their possible values are summarized in Table 4.1. Detailed information of the morphological annotation can be found in Zeman et al. (2005).

[1] http://ufal.mff.cuni.cz/

[2] http://www.ldc.upenn.edu/

[3] http://ufal.mff.cuni.cz/pdt2.0/

position	*name*	*description*	*# values*
1	**POS**	**Part of speech**	**12**
2	SubPOS	Detailed part of speech	60
3	**Gender**	**Gender**	**9**
4	Number	Number	5
5	Case	Case	8
6	PossGender	Possessor's gender	4
7	PossNumber	Possessor's number	3
8	Person	Person	4
9	Tense	Tense	5
10	**Grade**	**Degree of comparison**	3
11	**Negation**	**Negation**	2
12	Voice	Voice	2
13	Reserve1	Reserve	-
14	Reserve2	Reserve	-
15	Var	Variant, style	10

Table 4.1: Morphological categories encoded in Czech positional tags (Zeman et al., 2005).

afun	*description*
Pred	Predicate, a node not depending on another node
Sb	Subject
Obj	Object
Adv	Adverbial
Atr	Attribute
AtrAtr	An attribute of any of several preceding (syntactic) nouns
AtrAdv	Structural ambiguity between adverbial and adnominal dependency
AdvAtr	Dtto with reverse preference
AtrObj	Structural ambiguity between object and adnominal dependency
ObjAtr	Dtto with reverse preference
Atv	Complement (determining), hung on a non-verb. element
AtvV	Complement (determining), hung on a verb, no 2^{nd} gov. node
Pnom	Nominal predicate, or nom. part of predicate with copula *be*
Coord	Coordinated node
Apos	Apposition (main node)
ExD	Main element of a sentence without predicate, or deleted item
AuxV	Auxiliary verb *be*
AuxT	Reflexive tantum
AuxR	Reflexive passive
AuxP	Primary preposition, parts of a secondary preposition
AuxC	Conjunction (subordinate)
AuxO	Redundant or emotional item, 'coreferential' pronoun
AuxZ	Emphasizing word
AuxX	Comma (not serving as a coordinating conjunciton)
AuxG	Other graphic symbols, not terminal
AuxY	Adverbs, particles not classed elsewhere
AuxK	Terminal punctuation of a sentence

Table 4.2: Analytical functions and their description (Hajič et al., 1997).

ID	*form*	*lemma*	*tag*	*parentID*	*afun*
1	Zbraně	zbraň	`NNFP1-----A----`	0	ExD
2	hromadného	hromadný	`AANS2----1A----`	3	Atr
3	ničení	ničení_^(*3it)	`NNNS2-----A----`	1	Atr

Table 4.3: Example of a text (*zbraně hromadného ničení – weapons of mass destruction*) annotated on morphological (*lemma* + *tag*) and analytical (*parentID* + *afun*) layers.

Analytical layer

Analytical layer of PDT serves to encode sentence dependency structures. Each word in a sentence is linked to its **head word** and assigned its **analytical function** (dependency type). If we think of a sentence as a graph with words as nodes and dependency relations as edges, the resulting structure is a tree – a directed acyclic graph having one root, in the theory of dependency syntax called *dependency tree*. Possible values of analytical functions are listed in Table 4.2. Complete details of analytical annotation can be found in Hajič et al. (1997) and a small example of an annotated text in Table 4.3. Each token (either a word or a punctuation mark) is represented by: its position in the sentence (*ID*), word form as it appears in the original text, lemma, morphological tag, position of the governing word (*parentID*) (or 0 if the token is the root), and analytical function (*afun*).

4.2.2 Candidate data sets

Two collocation candidate data sets were obtained from the Prague Dependency Treebank. Both were extracted from morphologically normalized texts and filtered by a part-of-speech filter and a frequency filter. Details of these steps are described in the following parts:

Morphological normalization

The usual role of morphological normalization is to canonize morphological variants of words so that each word (lexical item) can be identified regardless of its actual morphological form. This technique has been found to be very beneficial in information retrieval, for example, especially when dealing with morphologically rich languages such as Czech (Pecina et al., 2008). Two basic approaches to this problem are: a) **stemming**, where a word is transformed (usually heuristically) into its *stem* which often does not represent a meaningful word, and b) **lemmatization**, where a word is properly transformed into its base form (lemma) by means of morphological analysis and disambiguation. For details, see e.g. Frakes and Baeza-Yates (1992) or Manning et al. (2008).

The latter approach seems more reasonable in our case (manually assigned lemmas are available in PDT) but it is not completely adequate. By transforming words

form	*lemma*	*full tag*	*lemma proper*	*reduced tag*
Zbraně	zbraň	**N**N**F**P1----**-A**----	zbraň	**NF-A**
hromadného	hromadný	**A**A**N**S2----**1A**----	hromadný	**AN1A**
ničení	ničení_^(*3it)	**N**N**N**S2----**-A**----	ničení	**NN-A**

Table 4.4: Morphological normalization of surface word forms. A normalized form consists of a lemma proper (lemma without technical suffixes) and a reduced morphological tag.

only into lemmas, we would lose important information about their lexical senses that we need to distinguish between the occurrences of different collocation candidates. For example, *negation* and *grade* (degree of comparison) significantly change word meanings and differentiate between collocation candidates (e.g. *secure area* vs. *insecure area*, *big mountain* vs. *(the) highest mountain*). Indication of such morphological categories in the PDT system is not encoded in the lemma but rather in the tag. With respect to our task, we normalized word forms by transforming them into a combination of a **lemma** (lemma proper, in fact; the technical suffixes in PDT lemmas are omitted) and a **reduced tag** that comprises the following morphological categories: *part-of-speech*, *gender*, *grade*, and *negation* (highlighted in Table 4.1). An example of morphological normalization is shown in Table 4.4.

For similar reasons and in order to decrease the granularity of collocation candidates, we also simplified the system of Czech analytical functions by merging some of them into a single value. Details are depicted in Table 4.2, where only the highlighted part of analytical function values is kept.

Part-of-speech filtering

A part-of-speech filter is a simple heuristic that improves the results of collocation extraction methods (Justeson and Katz, 1995): the collocation candidates are passed through a filter which only lets through those patterns that are likely to be "phrases" (potential collocations) and not random word combinations. Similar approaches were used also by Ross and Tukey (1975) and Kupiec et al. (1995). Our motivation for part--of-speech filtering is similar but not quite identical. Justeson and Katz (1995) filtered the data in order to keep those that are more likely to be collocations than others; for bigram collocation extraction they suggest to use only patterns A:N (*adjective–noun*) and N:N (*noun–noun*). On the other hand, we deal with a broader notion of collocation in this work and this constraint would be too constraining. We filtered out candidates with part-of-speech patterns that *never* form a collocation (at least in our data), in other words, we allow all part-of-speech patterns that can *possibly* form a collocation. This step does not affect the evaluation because it can be done prior to all extraction methods (token filtering). A list of the employed patterns is presented in Table 4.6. It was proposed congruently by the annotators before the annotation process, described in Section 4.2.3, started.

ID	*lemma proper*	*reduced tag*	*parentID*	*afun*
1	zbraň	`NF-A`	0	Head
2	hromadný	`AN1A`	3	Atr
3	ničení	`NN-A`	1	Atr

Table 4.5: Example of a normalized collocation candidate.

Frequency filtering

To make sure that the evaluation is not biased by low-frequency data, we limit ourselves only to collocation candidates occurring in PDT more than five times. The less frequent candidates do not meet the requirement for sufficient evidence of observations needed by some methods used in this work (they are unreliable when dealing with rare events) and thus were not included in our evaluation. While Moore (2004) clearly stated that these cases comprise the majority of all the data (the well-known Zipfian phenomenon (Zipf, 1949)) and should not be excluded from real-world applications, Evert (2004, p. 22) argues that "it is impossible *in principle* to compute meaningful association scores for the lowest-frequency data".

PDT-Dep

Dependency trees from the treebank were broken down into **dependency bigrams** (Section 2.2.4). From all sentences in PDT, we obtained a total of 635 952 different dependency bigram types (494 499 of them were singletons). Only 26 450 of them occur in the data more than five times. After applying the frequency and part-of-speech pattern filter, we obtained a list of 12 232 collocation candidates (consisting of a normalized head word and its modifier, plus their dependency type), further referred to as *PDT-Dep*.

PDT-Surf

Although collocations form syntactic units by the definition we use (Section 2.1.5), it is also possible to extract collocations as **surface bigrams**, i.e. pairs of adjacent words (Section 2.2.4) without the guarantee that they form such units but under the assumption that a majority of bigram collocations cannot be modified by the insertion of another word and in text they occur as surface bigrams (Manning and Schütze, 1999, Chapter 5). In real-world applications this approach would not require the source corpus to be parsed, which is usually a time-consuming process, accurate only to a certain extent. A total of 638 030 surface bigram types was extracted from PDT, 29 035 of which occurred more than five times. After applying the part-of-speech filter, we obtained a list of 10 021 collocation candidates (consisting of normalized component words), further referred to as *PDT-Surf*. 974 of these bigrams do not appear in the *PDT-Dep* test set (ignoring syntactic information).

POS pattern	*example*	*translation*
A:N	trestný čin	*criminal act*
N:N	doba splatnosti	*term of expiration*
V:N	kroutit hlavou	*shake head*
R:N	bez problémů	*no problem*
C:N	první republika	*First Republic*
N:V	zranění podlehnout	*succumb*
N:C	Charta 77	*Charta 77*
D:A	volně směnitelný	*freely convertible*
N:A	metr čtvereční	*square meter*
D:V	těžce zranit	*badly hurt*
N:T	play off	*play-off*
N:D	MF Dnes	*MF Dnes*
D:D	jak jinak	*how else*

Table 4.6: Patterns for part-of-speech filtering with their examples. The capital letters denote: adjectives (A), nouns (N), numerals (C), verbs (V), adverbs (D), prepositions (R), particles (T).

4.2.3 Manual annotation

Three educated linguists, familiar with the phenomenon of collocation, were hired to annotate the reference data sets extracted from PDT. These annotators agreed on a definition of collocation adopted from Choueka (1988): "[A collocation expression] has the characteristics of a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components." This notion requires collocations to be grammatical units (subtrees of sentence dependency trees in case of dependency syntax employed in PDT) that are not entirely predictable (semantically and syntactically). This definition is relatively wide and covers a broad range of lexical phenomena such as idioms, phrasal verbs, light verb constructions, technical expressions, proper names, stock phrases, and also weaker lexical preferences. Basically, the annotators had to judge each candidate whether it could be considered a free word combination (with intensionally restricted collocability) or not (and hence, should be placed in a lexicon as collocation).

The dependency bigrams in *PDT-Dep* were assessed first. The annotation was performed independently, in parallel, and without any knowledge of context. In order to minimize the cost of the process, each collocation candidate was presented to each annotator only once although it could appear in various different contexts. The annotators were instructed to judge any bigram which could *eventually* appear in a context where it has a character of collocation as a *true collocation*. For example, idiomatic expressions were judged as collocations although they can also occur in contexts where they have a literal meaning; similarly for other types of collocations. As a result, the annotators were relatively liberal in their judgments, but their full agreement was

	0	*1*	*2*	*3*	*4*	*5*
0	7 066	644	135	78	208	3
1	590	265	125	0	96	0
2	13	8	621	0	46	1
3	74	0	1	185	0	0
4	409	442	87	0	1075	7
5	25	3	2	2	15	6

	0	*1–5*
0	7 066	1 068
1–5	1 111	2 987

Table 4.7: Confusion matrix of two annotators measured on the fine-grained (0-5) categories (on the left) and after the collocation categories (1–5) were merged together (on the right).

required to mark a candidate as a true collocation in the reference data set. Problems could have arisen in cases where the annotators had poor knowledge of some (e.g. technical) domain and could have misjudged certain less-known technical terms from this domain. The Prague Dependency Treebank, fortunately, does not contain such texts (see Section 4.2.1) and this sort of problems was not observed (according to the annotators).

During the assessment, the annotators also attempted to distinguish between subtypes of collocations and classified each collocation into one of the categories below. This classification, however, was not intended as a result of the process (our primary goal was binary classification) but rather as a way to clarify and simplify the annotation. Any bigram that can be assigned to any of the categories was considered a collocation.

1. stock phrases, frequent unpredictable usages
   zásadní problém (major problem), konec roku (end of the year)
2. proper names
   Pražský hrad (Prague Castle), Červený kříž (Red Cross)
3. support-verb constructions
   mít pravdu (to be right), činit rozhodnutí (make decision)
4. technical terms
   předseda vlády (prime minister), očitý svědek (eye witness)
5. idiomatic expressions
   *studená válka (cold war), visí otazník (*lit. *hanging question mark ~ open question)*

The surface bigrams from *PDT-Surf* were annotated in the same fashion but only those collocation candidates that do not appear in *PDT-Dep* were actually judged. Technically, we removed the syntactic information from *PDT-Dep* data and transfered the annotations to *PDT-Surf*. If a surface bigram from *PDT-Surf* appears also in *PDT--Dep*, it is assigned the same annotation from all three annotators.

annotations	*fine grained*		*binary*	
	accuracy	*Fleiss' κ*	*accuracy*	*Fleiss' κ*
A1–A2	72.1	0.49	79.5	0.55
A2–A3	71.1	0.47	78.6	0.53
A1–A3	75.4	0.53	82.2	0.60
A1–A2–A3	61.7	0.49	70.1	0.56

Table 4.8: Inter-annotator agreement for annotators A1, A2, A3 on *PDT-Dep* measured by accuracy and Fleiss' κ on all 6 categories (fine-grained) and after merging categories 1–5 (binary).

Inter-annotator agreement

The inter-annotator agreement on all the categories of collocations (plus a 0 category for non-collocations) was relatively low: the simple percent agreement (accuracy) between two annotators on *PDT-Dep* ranged from 71.1% to 75.4% and Cohen's κ[4] ranged from 0.47 to 0.53. The exact Fleiss' κ[5] among all the annotators was 0.49.

This demonstrates that the notion of collocation is very subjective, domain-specific, and also somewhat vague. In our experiments, we did not distinguish between different collocation categories – ignoring them (considering only two categories: *true collocations* and *false collocations*) increased Fleiss' κ among all the annotators to 0.56 (see details in Tables 4.7 and 4.8). Multiple annotation was performed in order to get a more precise and objective idea about what can be considered a collocation by combining independent outcomes of the annotators. Only those candidates that *all* three annotators recognized as collocations (of any type) were considered *true collocations* (full agreement required). The *PDT-Dep* reference data set contains 2 557 such bigrams (21.02% of all the candidates) and *PDT-Surf* data set 2 293 (22.88% of all the candidates). For comparison of these reference data sets, see Figure 4.1.

4.3 Czech National Corpus

In an era of multi-billion word corpora, a corpus of the size of the PDT is certainly not sufficient for real-world applications and thus we attempted to extract collocations also from a larger data – a set of a total of 242 million tokens from the Czech National Corpus. This data, however, lacks any manual annotation, and hence we settled for automatic part-of-speech tagging (Hajič, 2004) and extracted collocation candidates as surface bigrams similarly to the case of *PDT-Surf*.

[4] An agreement measure for two annotators (Cohen, 1960): $\kappa = \frac{P_o - P_e}{1 - P_e}$, where P_o is the relative *observed* agreement between annotators and P_e is the theoretical probability of *chance* agreement (each annotator randomly choosing each category). The factor $1 - P_e$ then corresponds to the level of agreement achievable above chance and $P_o - P_e$ is the level of agreement actually achieved above chance. We used this commonly accepted and robust measure although Krenn et al. (2004) argued against using it for linguistic annotations.

[5] A generalization of Cohen's κ for any numbers of annotators (Fleiss, 1971).

units	*all tokens*	*relevant tokens*
tags	95.78	94.77
lemmas	97.21	96.30
lemmas + tags	94.14	92.52
reduced tags	98.15	97.83
lemmas + reduced tags	96.34	95.37

Table 4.9: Accuracy of a Czech state-of-the-art morphological tagger measured on various different units. By default, accuracy is measured on tags of all tokens. *Relevant tokens* refer to words with part-of-speech used in the part-of-speech pattern filter described in Section 4.2.2.

4.3.1 Corpus details

The *Czech National Corpus* (CNC) is a log-term academic project with the aim of building up a large computer-based corpus, containing mainly written Czech.[6] This project consists of two main parts – synchronous and diachronic – and already produced a number of various valuable corpora that are available for academic purposes. The data we used in our evaluation experiments comprises two synchronous (containing contemporary written language) corpora SYN2000 (ICNC, 2000) and SYN2005 (ICNC, 2005), each containing about 100 million running words (excluding punctuation).

SYN2000 (released to the public in 2000) contains complete texts selected to cover the widest range of literary genres. It contains contemporary written Czech mostly from the period 1990-1999. SYN2005 (released in 2005) is again a synchronous but also a representative collection of texts (mostly from 1990-2004) reflecting the current distribution of text genres.

4.3.2 Automatic preprocessing

SYN2000 and SYN2005 are not manually annotated, neither on the morphological nor the analytical layer. Manual annotation of such an amount of data would be unfeasible. These corpora, however, were processed by a part-of-speech tagger (Spoustová et al., 2007) and provided at least with automatically assigned morphological tags. On the one hand, we do not want our evaluation to be biased by automatic linguistic preprocessing (hence we chose the manually annotated PDT as the source corpus for our main experiments), but on the other hand, we are interested in estimating the performance of the methods in real-world applications where the availability of a large-scale manually annotated data cannot be expected.

In order to better understand the possible bias caused by the automatic preprocessing tools, let us now study their actual performance. The part-of-speech tagging of our CNC data was performed by a hybrid tagger described in Spoustová et al. (2007). It is a complicated system based on a combination of statistical and rule-based methods.

[6] http://ucnk.ff.cuni.cz/

window span	1	2	3	4	5	6	7	8	9	Inf
accuracy (%)	90.89	89.45	88.12	87.16	86.47	85.99	85.56	85.27	85.04	84.76

Table 4.10: Accuracy of a current Czech state-of-the-art dependency parser with respect to the maximum span of a word and its head, ranging between 1–9 and more (Inf) words.

Its expected accuracy (ratio of correctly assigned tags) measured on the PDT evaluation test set is 95.68% (Spoustová et al., 2007). One of the statistical components used in this system is a state-of-the-art tagger based on discriminative training of Hidden Markov Models by the Averaged Perceptron algorithm. This approach was first introduced by Collins (2002) and for Czech morphology implemented by Votrubec (2006). Its current (unpublished) accuracy measured on full morphological tags (described in Section 4.2.1) is 95.78%. For measuring the accuracy of taggers, lemmas are typically ignored. If we count both the correctly assigned tags and lemmas, the accuracy will drop to 94.14%. The accuracy evaluated on lemmas and reduced tags which were used in our experiments (Section 4.2.2) is relatively high, a 96.34% (Table 4.9).

Based on this observation, we can assume that in an automatically tagged text approximately one out of 28 randomly selected tokens is assigned a wrong tag and/or lemma. Such a token, however, usually appears in more than one bigram. For surface bigrams, only the first and the last token of a sentence affect one bigram: all other tokens affect two different bigrams. In the case of dependency bigrams, only the root and leaf tokens appear in one bigram, other tokens can appear in two or more bigrams depending on the sentence tree structure. For both surface and dependency bigrams, the average number of bigrams affected by one token depends on the sentence length and is equal to $2(n-1)/n$, where n is the sentence length. For an average sentence from the PDT data, which has 17.1 tokens, the number of bigrams affected by one token equals 1.88. This implies that if one out of 28 tokens is not assigned a correct tag and/or lemma (accuracy of 96.34%), then approximately one out of 15 selected bigrams occurring in an automatically normalized text is misleading and contains an error (at least in one of its components). Furthermore, we should estimate the performance only on words that pass through our part-of-speech filter (Section 4.2.2). Accuracy on such data measured on lemmas and reduced tags is equal to 95.37%. Thus, we can assume that approximately every 12^{th} bigram occurrence contains an error. All the accuracy figures mentioned above are summarized in Table 4.9.

Both SYN2000 and SYN2005 are provided with automatic part-of-speech tagging but no syntactic analysis. Although automatic dependency parsers for Czech do exist, they were not used to obtain automatic sentence dependency structures of the data from CNC – mainly for reasons of time complexity. The state-of-the-art dependency parser is based on McDonald's maximum spanning tree approach (McDonald et al., 2005) and enhanced by Novák and Žabokrtský (2007). Its accuracy (ratio of correctly assigned head words and corresponding values of analytical function) measured on the evaluation test set from the PDT is 84.76%. This performance is much higher if

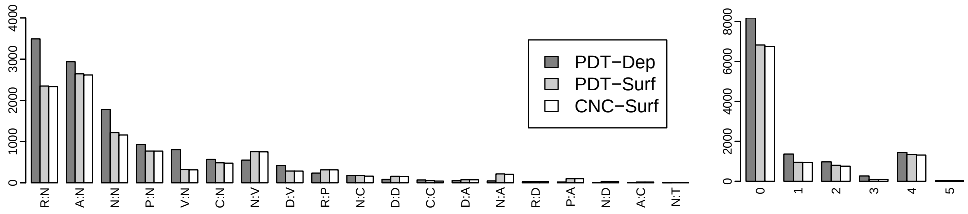

Figure 4.1: Distribution of part-of-speech patterns (left) and annotation categories (right) assigned by one of the three annotators in all the three Czech reference data sets.

we analyze words only in a limited surface distance. If we focus only on adjacent dependency bigrams, which are more likely to form collocations, the tagger's accuracy is almost 91%. As we allow for more distant dependencies (less likely to form collocations) the accuracy constantly decreases. See Table 4.10 for details.

4.3.3 Candidate data set

CNC-Surf

From the total of 242 million tokens from SYN2000 and SYN2005, we extracted more than 30 million **surface bigrams** (types) (Section 2.2.4). We followed the same procedure as for the PDT reference data. After applying the part-of-speech and frequency filters, the list of collocation candidates contained 1 503 072 surface bigrams. Manual annotation of such an amount of data was infeasible. To minimize the cost, we selected only a small sample of it – the already annotated bigrams from the *PDT-Surf* reference data set, a total of 9 868 surface bigrams, further called *CNC-Surf*. All these bigrams appear also in *PDT-Surf*, but 153 do not occur in it more than five times. *CNC-Surf* contains 2 263 (22.66%) *true collocations* – candidates that all three annotators recognized as collocations (of any type). For comparison with the reference data sets extracted from the PDT, see Figure 4.1.

4.4 Swedish PAROLE corpus

So far, all the reference data sets presented in this work have been extracted from Czech texts. In this section, we describe our last reference data set – *support-verb construction* candidates obtained from the Swedish *PAROLE* corpus, containing about 20 million words. This data differs not only in the language and the type of collocations used, but also in the extraction procedure. Our original motivation was to evaluate methods for semi-automatic building of a Swedish lexicon of support-verb constructions. Preliminary results of this work were described in Cinková et al. (2006).

category	*w=1*	*w=2*	*w=3*	*w=1–3*
0. non-collocations	7 320	7 080	2 119	15 735
1. phrasemes	63	24	8	79
2. quasimodals	24	14	8	31
3. support-verb constructions	557	559	232	1 182
all	7 964	7 677	2 367	17 027

Table 4.11: Distribution of the annotation categories in the Swedish reference data with respect to the surface distance between the candidate components (*w*) in the range 1–3.

4.4.1 Corpus details

The PAROLE corpus is a collection of modern Swedish texts comprising 20 million running words. It belongs to Språkbanken, the set of corpora at Språkdata, University in Gothenburg, Sweden.[7] The corpus was built within the EU project PAROLE (finished in 1997), which aimed at creating a European network of language resources (corpora and lexicons). It has automatic morphological annotation but lacks lemmatization. In order to deal with morphological normalization, an automatic lemmatizer developed by Cinková and Pomikálek (2006) was employed to transform all word forms into their lemmas.

4.4.2 Support-verb constructions

Support-verb construction (SVC) is combination of a lexical verb and a noun or a nominal group containing a predication and denoting an event or a state, e.g. *to take/make a decision, to undergo a change*. From the semantic point of view, the noun seems to be part of a complex predicate rather than the object of the verb, whatever the surface syntax may suggest (Cinková et al., 2006). The meaning of a SVC is concentrated in the predicate noun, whereas the semantic content of the verb is reduced or generalized. The notion of SVC and related concepts has already been studied elsewhere, e.g. in Grefenstette and Teufel (1995), Tapanainen et al. (1998), Lin (1999), McCarthy et al. (2003), and Bannard et al. (2003).

Our interest in SVCs is mainly in the perspective of foreign language learners and building a lexicon (see Cinková et al., 2006). Although SVCs are easily understood by foreign language learners, they pose substantial problems for foreign language production (Heid, 1998) due to the unpredictability of the support verb. For example, the predicate noun *question* in an SVC meaning *to ask* takes different support verbs in Czech and in Swedish: Czech uses the verb *položit* (i.e. *to put horizontally*) while Swedish uses the verb *ställa* (i.e. *to put vertically*). The translation equivalent to the support verb is unpredictable, though the common semantic motivation can be traced

[7] http://spraakbanken.gu.se/PAROLE/

back. The unpredictability of the support verb places SVCs into the lexicon, while the semantic generality of support verbs and their productivity move them to the very borders of grammar (Cinková et al., 2006).

4.4.3 Manual extraction

The reference data was obtained by the following manual extraction procedure. It was inspired by other similar approaches (e.g. Heid, 1998) and comprises these steps:

1. semi-automatic extraction of word expressions whose morphosyntactic character suggests that they are potential support-verb constructions,
2. subsequent manual elimination of non-collocations,
3. manual sorting of collocations into three groups: *SVCs*, *quasimodals*, *phrasemes*.

Step 1 involved formulating several corpus queries and obtaining the results. The queries basically varied the distance between the verb and the noun (ranging from 1 to 3). Some queries introduced article, number, and adjective insertion restrictions. To ensure that the noun was the object of the verb, the verbs had to follow a modal or an auxiliary verb.

In step 2, the collocation candidates were ordered according to their frequency in the corpus. Each *collocation interval* (the distance between the noun and the verb) was processed separately. Equally frequent collocation candidates were sorted alphabetically according to their verbs. This facilitated manual processing, as some very frequent verbs could be instantly recognized as never forming support verbs, and ignored in blocks, i.e. *kåpa (to buy)* or *såga (to say)*.

Step 3 included a fine-grained semantic classification. Three groups were set at the beginning: *SVCs*, *quasimodals*, and *phrasemes*. The *SVCs* group included collocations with nouns denoting an event (also a state) or containing a predication, e.g. *få hjålp (to get help)* and *få betydelse* (lit. *to get significance – to become significant*). In the *SVCs* group, it is the event described by the predicate noun that actually "takes place". In *quasimodals*, on the other hand, the verb and the predicate noun form one semantic unit that resembles a modal verb (e.g. *to get the chance to V = to start to be able to V* etc.) (Cinková and Kolářová, 2004) and must be completed by the event in question (here marked as V). *Phrasemes* include frequent collocations in which the noun is not a predicate noun and the meaning of the entire unit is idiomatic (e.g. *ta hand om X*, lit. *to take hand about X – to take care of X*).

Naturally, this sorting was strongly based on intuition. Basically, the phraseme and quasimodal groups also allow for nouns which do not contain any predication (e.g. *hand*), while the "pure *SVCs*" are intended to be denoting events and states. With respect to this, we were not able to find a consistent solution for constructions like *begå en dummhet* (lit. *to commit a stupidity*), which underspecifies the given event.

reference data set	*PDT-Dep*	*PDT-Surf*	*CNC-Surf*	*PAR-Dist*
morphology	*manual*	*manual*	*auto*	*auto*
syntax	*manual*	*none*	*none*	*none*
bigram types	*dependency*	*surface*	*surface*	*distance*
sentences	87 980	87 980	15 934 590	2 639 283
tokens	1 504 847	1 504 847	242 272 798	22 883 361
words (no punctuation)	1 282 536	1 282 536	200 498 152	20 240 346
bigram types	635 952	638 030	30 608 916	13 370 375
after frequency filtering	26 450	29 035	2 941 414	13 370 375
after part-of-speech filtering	12 232	10 021	1 503 072	898 324
collocation candidates	12 232	10 021	9 868	17 027
data sample size (%)	100	100	0.66	1.90
true collocations	2 557	2 293	2 263	1 292
baseline precision (%)	21.02	22.88	22.66	7.59

Table 4.12: Summary statistics of all the four reference data sets and their source corpora.

PAR-Dist

The extraction procedure was designed and performed by Dr. Silvie Cinková and yielded 17 027 SVC candidates occurring at collocation intervals 1–3, out of which 15 735 were classified as negative examples, not collocations of our interest. 1 182 collocations were classified as *SVCs*, 21 were labeled as *quasimodal*, 79 were labeled as *phrasemes*. All of these cases are considered *true collocations* in our experiments. Details are shown in Table 4.11. This reference data set is further referred to as *PAR-Dist*and detailed comparison of the four reference data sets is shown in Table 4.12.

Crossvalidation split

For the purposes of significance testing (Section 5.1.3) and crossvalidation in our experiments, all the data sets were split into seven stratified subsets (folds), each containing the same ratio of true collocations (to ensure the prior probabilities of true collocations, i.e. the baseline precision scores, are equal in all the folds). This number was chosen as a compromise between two contradictory needs: 1) to have enough folds for a paired test of significance, and 2) to have enough instances in each fold for reliable estimates of evaluation scores. Six of the folds (called the **evaluation folds**) were used for 6-fold cross validation and estimation of average performance including significance testing (Chapter 5). The one remaining fold (called the **held-out fold**) was put aside and used as held-out data in additional experiments (Section 6.5).

5
Empirical Evaluation

In this chapter, we present a comparative performance evaluation of the 82 association measures discussed in Chapter 3. The evaluation experiments were performed on the four data sets described in Chapter 4: dependency bigrams from the Prague Dependency Treebank (*PDT-Dep*), surface bigrams from the same source (*PDT-Surf*), instances of surface bigrams from the Czech National Corpus (*CNC-Surf*), and distance verb-noun combinations from the Swedish PAROLE corpus (*PAR-Dist*).

In the first section, we introduce our evaluation scheme based on precision and recall. Then, we evaluate performance of the association measures separately on the individual data sets and attempt to compare the obtained results across the different data sets.

5.1 Evaluation methods

From the statistical point of view, collocation extraction can be viewed as a **classification problem**, where each collocation candidate from a given data set must be assigned to one of two categories: *collocation* or *non-collocation*. By setting a threshold, any association measure becomes a *binary classifier*: the candidates with higher association scores fall into one class (collocation), the rest into the other class (non--collocation). Effectiveness of such a classifier can be visualized in the form of a **confusion matrix** (Kohavi and Provost, 1998), also called a *table of confusion*, or a *matching matrix*. This matrix contains information about the actual and predicted classifications done by the classifier on a given data set. An example of a confusion matrix for a classifier of collocations is shown in Table 5.1.

The rows in the confusion matrix represent instances of the *true* (gold-standard) classes and the columns represent instances of the *predicted* classes. The cells then contain counts of the instances divided into four sets according to their true and predicted classification as depicted in Table 5.1: *true positives* (TP) are correctly classified true collocations, *false negatives* (FN) are misclassified true collocations, *false positives* (FP) are misclassified true non-collocations, and *true negatives* (TN) are correctly classified true non-collocations.

The performance of this classifier can be evaluated using the data in its confusion matrix, e.g. by a common evaluation measure **accuracy** – the fraction of correct predictions, i.e. the candidates correctly predicted either as collocations or non-collocations:

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}, \quad A \in \langle 0, 1 \rangle.$$

		predicted	
		collocation	*non-collocation*
true	*collocation*	TP	FN
	non-collocation	FP	TN

Table 5.1: A confusion matrix of prediction of collocations.

However, the prior probabilities of the two classes (the number of true collocations vs. non-collocations) are usually unbalanced and in that case, the accuracy is not a very representative evaluation measure of the classifier performance – the classifier can be biased towards non-collocations. Since we are more interested in correct prediction of collocations rather than non-collocations, several authors (e.g. Evert and Krenn, 2001) have suggested to use *precision* and *recall* as more appropriate evaluation measures: **precision** (P) is the fraction of positive predictions that are correct (correctly predicted true collocations), **recall** (R) is the fraction of positives that are correctly predicted (true collocations correctly predicted):

$$\mathsf{P} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}}, \quad \mathsf{P} \in \langle 0, 1 \rangle, \qquad \mathsf{R} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}}, \quad \mathsf{R} \in \langle 0, 1 \rangle.$$

These two evaluation measures are interdependent – by changing the classification threshold (also called discrimination threshold), we can tune the classifier and trade off between recall and precision, as illustrated in Figure 5.2

5.1.1 Precision-recall curves

Choosing the optimal classification threshold depends primarily on the intended application and there is no principled way of finding its optimal value (Inkpen and Hirst, 2002). Instead, we can measure the performance of association measures by pairs of precision-recall scores within the entire interval of possible threshold values. In this manner, individual association measures can be thoroughly compared by their two--dimensional *precision-recall curves* visualizing the quality of ranking collocation candidates without committing to a classification threshold. The closer the curve stays to the top and right, the better the ranking procedure is.

Formally, the **precision-recall curve** is a graphical plot of recall vs. precision for a classifier as its classification threshold is varied. The concept of the precision-recall curve is closely related to a *receiver operating characteristic* (ROC) curve which compares two *operating characteristics* computed also from the data of the confusion matrix – namely the fraction of true positives ($\mathsf{TPR} = \mathsf{TP}/(\mathsf{TP}+\mathsf{FP})$) vs. the fraction of false positives ($\mathsf{FPR} = \mathsf{FP}/(\mathsf{FP}+\mathsf{TN})$) as the criterion (threshold) changes (Fawcett, 2003). ROC analysis is a popular diagnostic tool used to select optimal classification models. Originally, it was used in signal detection theory (in 1960s) but recently, it was introduced

collocation candidate	*translation*	*PMI*	*precision*	*recall*
Červený kříž	*Red Cross*	15.66	**100.00**	**12.50**
železná opona	*iron curtain*	15.23	**100.00**	**25.00**
řádová čárka	*decimal point*	14.01	**100.00**	**37.50**
kupónová knížka	*coupon book*	13.83	**100.00**	**50.00**
autor knihy	*book author*	11.05	80.00	50.00
aritmetická operace	*arithmetical operation*	10.52	**83.33**	**62.50**
podavač papíru	*paper feeder*	10.17	**85.71**	**75.00**
nová kniha	*new book*	10.09	75.00	75.00
kulatý stůl	*round table*	7.03	**77.77**	**87.50**
nová vlna	*new wave*	6.59	70.00	87.50
čerpací stanice	*gas station*	6.04	**72.72**	**100.00**
systém typu	*system of a type*	3.54	66.66	100.00
centrum města	*city center*	1.54	61.53	100.00
na další	*on next*	0.54	57.14	100.00
program v	*program in*	0.35	53.33	100.00
úroveň je	*level is*	0.25	50.00	100.00

Table 5.2: Precision-recall trade-off illustrated on a ranked list of collocation candidates (true collocations are in bold) sampled from *PDT-Dep* and ranked by *Pointwise mutual information (4)*.

also into areas such as machine learning and data mining . The precision-recall (PR) curves are commonly used for the evaluation of methods in natural language processing and information retrieval when dealing with unbalanced data sets (which is also the case of collocation extraction) because they give a more informative picture of the classifier's performance. For a more detailed comparison of ROC and PR curves, see e.g. Davis and Goadrich (2006).

From the statistical perspective, the precision-recall curves must be viewed as estimates of their true (unknown) shapes from a (random) data sample (fold). As such they have a certain statistical variance and are sensitive to data. For illustration, see Figure 5.1 showing PR curves obtained on the six crossvalidation folds of *PDT-Dep* (each of the thin curves corresponds to one data fold). In order to obtain a good estimation of their true shape we must apply some kind of **curve averaging** where all crossvalidation folds with precision-recall scores are combined and a single curve is drawn. Such averaging can be done in three ways (Fawcett, 2003): *vertical* – averaging precision at the same fixed levels of recall, *horizontal* – averaging recall at the same fixed levels of precision, and *combined* – fixing threshold, averaging both precision and recall. The averaged results are then presented on a curve. Vertical averaging, as illustrated in Figure 5.1, worked reasonably well in our case and was used in our further experiments. The thin curves are produced by a single association measure on six separate data folds; the thick one is obtained by vertical averaging and better characterizes the true performance on the whole data set.

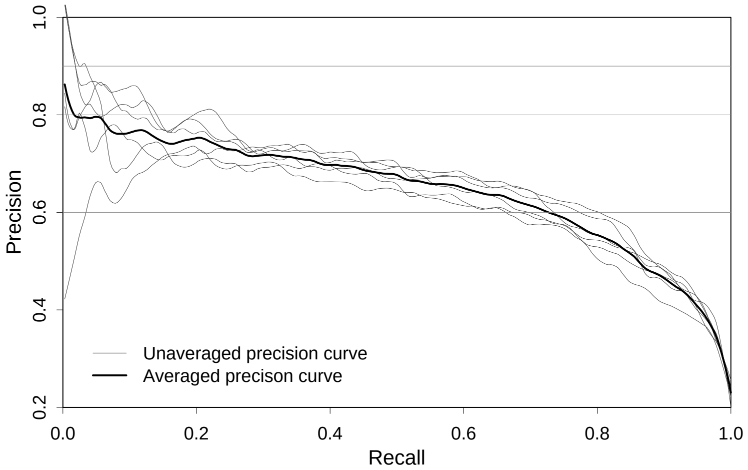

Figure 5.1: An example of vertical averaging of precision-recall curves. The thin curves represent individual non-averaged curves obtained by *Pointwise mutual information (4)* on each of the six evaluation data folds of *PDT-Dep*, the thick one is obtained by vertical averaging.

5.1.2 Mean average precision

The visual comparison of precision-recall curves is a powerful evaluation tool. However, it has a certain weak point: while a curve that predominates another one within the entire interval of recall is evidently better (although it might not be significantly better), when this is not the case, the judgment is not so obvious. Also the significance testing of the difference on the curves is non-trivial – it should be done *interval-wise* by comparing the curves globally on the whole interval of recall (Prchal, 2008), not only *point-wise* by comparing the points of precision at fixed levels of recall independently of each other (Evert, 2004). Instead of evaluating association measures directly by their PR curves, we propose the *average precision* (AP) as a more appropriate evaluation measure that can simply compare the evaluated methods by their overall performance. This measure is adopted from information retrieval, where it is widely used for comparing the performance between retrieval techniques or systems (Buckley and Voorhees, 2000).

Formally, for a ranked list of collocation candidates, we define the **average precision** as the mean of the precision scores obtained after each true collocation appears in the list:

$$AP = \frac{1}{r}\sum_{i=1}^{n} x_i p_i, \quad p_m = \frac{1}{m}\sum_{k=1}^{m} x_k, \quad x_k \in \{0,1\},$$

where r is the total number of true collocations in the fold, n is the total number of all candidates in the fold, p_m is the precision after m candidates in the ranked list, and

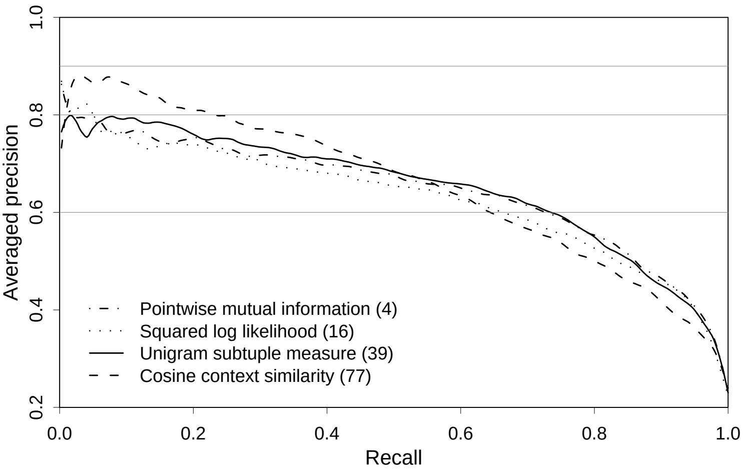

Figure 5.2: Examples of crossvalidated and averaged precision-recall curves of some well--performing association measures obtained on the *PDT-Dep* data set.

x_k indicates if the k^{th} candidate in the list is a true collocation (x_k=1) or not (x_k=0). The average precision can also be understood as the expected value of precision for all possible values of recall, assuming uniform distribution of recall (all possible values of recall are equally probable):

$$AP \approx E\left(P(R)\right), \quad R \sim U(0,1).$$

In the example in Table 5.2, the average precision would be computed from the precision scores highlighted in bold. Another interpretation of the average precision is the *area under the (PR) curve* (AUC). Nevertheless, our approach does not require the precision-recall values to be transformed into a (continuous) curve in order to estimate the area under it.

Based on the average precision scores AP_j computed for N data folds, we define the **mean average precision** (MAP) as the sample mean of these scores and use it as the main evaluation measures in our work:

$$MAP = \frac{1}{N} \sum_{j=1}^{N} AP_j$$

Note: In order to reduce the bias caused by the unreliable precision scores for low recall and their fast changes for high recall (see again Figure 5.1), we limit the estimation of AP to a narrower range of recall $\langle 0.1, 0.9 \rangle$ and use this estimation in all our experiments.

5.1.3 Significance testing

Statistical tests of the difference between the ranking methods are necessary to examine whether the observed differences in the evaluation scores (MAP) are measurable or whether they occur only by chance. Because MAP is averaged over a number of AP values computed on the separate (independent) data folds, we can employ tests based on estimating the error of this measure.

As we have mentioned earlier, the precision-recall curves are quite sensitive to the data and thus, we can expect differences in the AP values to be greater *between data folds* than *between methods*. Therefore, when comparing two ranking methods, we should analyze their AP difference for each matched pair of data folds (D_i) rather than the difference between AP values averaged over all the folds ($\bar{D}$). This problem is usually solved by the **paired Student's t-test** which compares the average difference of AP between two methods on the separate data folds to the variation of the difference across the folds. If the average difference is large enough compared to its standard error, then the methods are significantly different.

$$t = \frac{\bar{D}}{S_{\bar{D}}/\sqrt{N}}, \quad \bar{D} = \frac{1}{N}\sum_{i=1}^{N} D_i, \quad S_{\bar{D}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(D_i - \bar{D})^2},$$

where D_i is the AP difference on the i^{th} data fold, $\bar{D}$ is the average difference over all folds ($i = 1, \dots, N$), and $S_{\bar{D}}$ is the sample standard deviation.

Although the t-test requires the differences to be normally distributed, it works quite well even if this assumption is not completely valid. However, as a non-parametric alternative, we can apply the **paired Wilcoxon signed-ranked tests** which is commonly used in information retrieval. This test is more conservative and takes into account only the rank and sign of the difference and ignores the actual magnitude. The differences in AP on each data fold are replaced with the ranks of their absolute values and each rank is multiplied by the sign of the difference (R_i). The sum of the signed-ranks is compared to its expected value under the assumption that the two groups are equal. For details and description of other possible tests, see e.g. Hull (1993).

$$T = \frac{\sum_{i=1}^{N} R_i}{\sqrt{\sum_{i=1}^{N} R_i^2}}, \quad R_i = \mathrm{sign}(D_i) \cdot \mathrm{rank}|D_i|.$$

5.2 Experiments

In order to evaluate the performance of the individual association measures, we performed the following experiment on each of the four data sets introduced in Chapter 4. For all collocation candidates, we extracted their frequency information (the observed contingency tables) and context information (the immediate and empirical contexts)

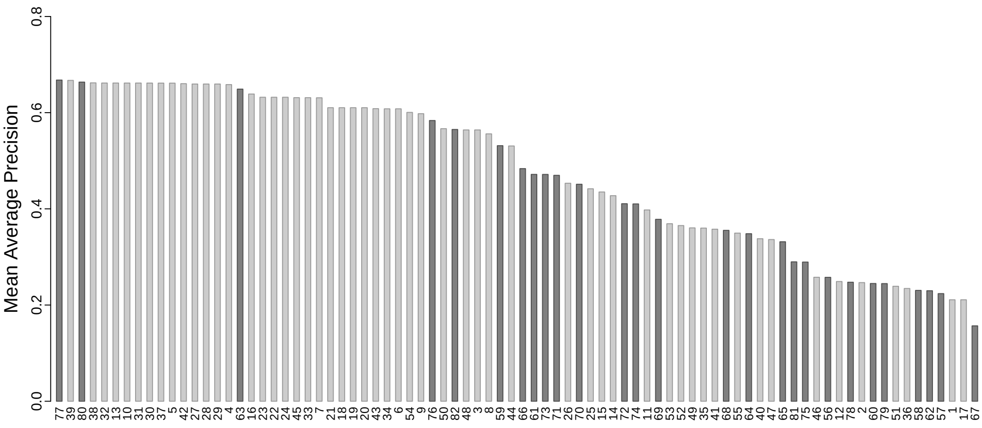

Figure 5.3: Sorted MAP scores of all association measures computed on *PDT-Dep*. The light bars correspond to the statistical measures and the dark bars to the context-based measures.

from their source corpora as described in Section 2.2.5. The empirical contexts were limited to a context window of 3 sentences (the actual one, the one preceding, and the one following) and filtered to include only open-class word types as described in Section 2.2.6. Based on this information, we computed the scores for all 82 association measures for all the candidates in each evaluation data fold. Then, for each association measure and each fold, we ranked the candidates according to their descending association scores, computed values of precision and recall after each true collocation appearing in the ranked list, plotted the averaged precision-recall curve, and computed the average precision on the recall interval $\langle 0.1, 0.9 \rangle$. The AP values obtained on the evaluation data folds were used to estimate the mean average precision as the main evaluation measure. Further, we ranked the association measures according to their MAP values in descending order and depicted the results in a graph. Finally, we applied the paired Student's and Wilcoxon tests to detect the measures with statistically indistinguishable performance. The actual results are presented in the following subsections.

5.2.1 Prague Dependency Treebank

First, we evaluated the association measures on the *PDT-Dep* data set of dependency bigrams extracted from the morphologically and syntactically annotated Prague Dependency Treebank, filtered by the part-of-speech and frequency filters as described in Section 4.2. A baseline system ranking the *PDT-Dep* candidates randomly would operate with the expected precision (AP and also MAP) of 21.02%, which is the prior probability of a collocation candidate to be a true collocation. Precision-recall curves of some well-performing methods are plotted in Figure 5.2. The best method evalu-

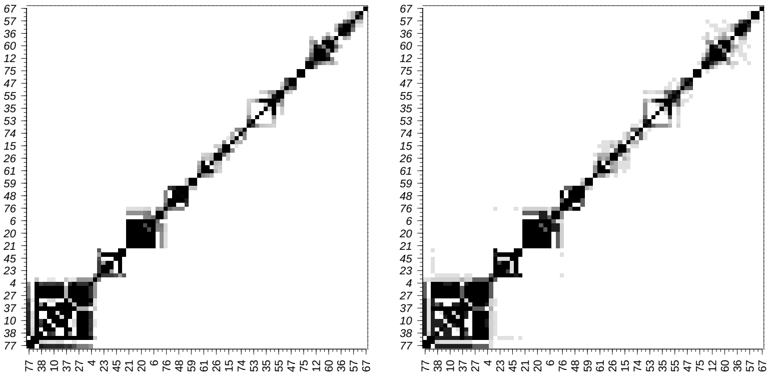

Figure 5.4: Visualization of p-values from the significance tests of difference (Student's t-test on the left and Wilcoxon signed-rank test on the right) between all methods on *PDT-Dep* sorted according to their MAP. The gray points correspond to p-values greater than α=0.05 and indicate pairs of methods with statistically indistinguishable performance.

ated by the mean average precision is *Cosine context similarity in boolean vector space (77)* with MAP=66.79%, followed by *Unigram subtuple measure (39)* with MAP=66.72% and other 14 association measures with nearly identical performance (in terms of MAP, see Figure 5.3). They include some popular methods known to perform reliably in this task, such as *Pointwise mutual information (4)*, *Mutual dependency (5)*, *Pearson's χ^2 test (10)*, *Z score (13)*, or *Odds ratio (27)*. Surprisingly, another commonly used method *T test (12)* only achieved MAP=24.89% and performed only slightly above the baseline. Although the best association measure uses the empirical context information, most of the other context-based methods are concentrated in the second half of the ranked list of the association measures (indicated by dark-gray bars) and did not preform well.

The significance tests were applied on all pairs of the association measures and their results are visualized in Figure 5.4 in the form of a matrix of p-values for both types of the test (the Student's t-test on the left and Wilcoxon signed-rank test on the right). The dark points indicate pairs of measures with statistically indistinguishable MAP ($p \geq 0.05$), the white space indicates pairs that are statistically different ($p < 0.05$). The big dark square in the bottom left corner corresponds to the 16 best measures mentioned earlier. Almost all of them are statistically indistinguishable from one another (with some exceptions). Further in the ranked list of association measures, we can observe also other "clusters" of measures with statistically equal performance determined by the dark squares on the diagonal. If we compare these two tests, we can conclude that the Wilcoxon test is indeed more conservative (more pairs of the measures are indistinguishable) but in general, the results are not very distinct.

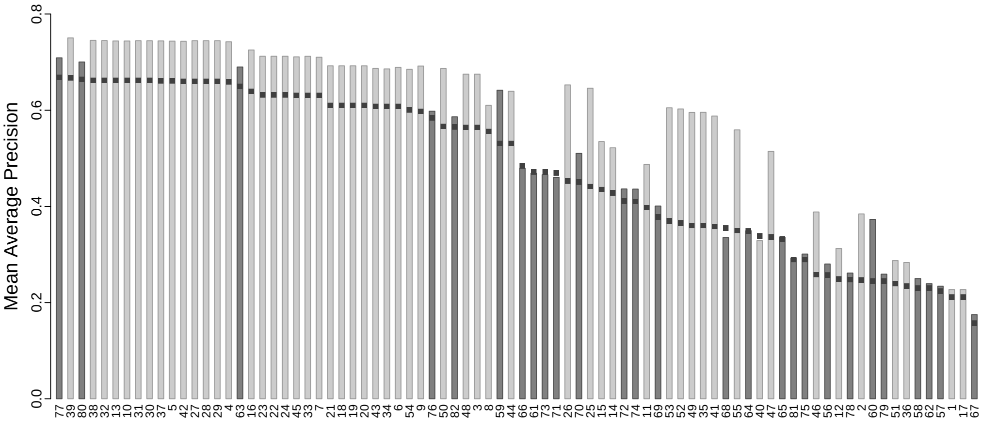

Figure 5.5: MAP scores of association measures computed on the *PDT-Surf* data set (bars) and sorted by the MAP scores obtained on the *PDT-Dep* data set (square points) in descending order.

As the second experiment, we performed the same procedure on the the *PDT-Surf* data set of surface bigrams extracted from the Prague Dependency Treebank (exploiting only the morphological information), and depicted the resulting MAP scores of all association measures in Figure 5.5. For a better comparison, the methods are sorted according to the results obtained on *PDT-Dep*. Extracting collocations as surface bigrams seems to be more reasonable than as dependency bigrams. The MAP scores of most association measures increased dramatically. The best performing method was *Unigram subtuple measure (39)* with MAP=75.03% compared to 66.71% achieved on the dependency bigrams (absolute improvement of 11.68%). This is probably due to the non-directly adjacent dependency bigrams not appearing in the *PDT-Surf* data set: in most cases, they do not form collocations. Interestingly, this improvement is not so significant for context-based association measures (see the dark-gray bars in Figure 5.5). The best context-based measure on the dependency bigrams *(77)* ended up as the 22nd on the surface data and its score increased only by absolute 4.1%.

5.2.2 Czech National Corpus

The third experiment was performed analogously on the instances of *PDT-Surf* in the Czech National Corpus – the *CNC-Surf* reference data set. The content of these two data sets is almost the same, *CNC-Surf* shares 98.46% of the collocation candidates with *PDT-Surf*. The main difference is in their source corpora. The data from the Czech National corpus are approximately 150 times larger (in terms of the number of tokens). The average frequency of candidates in *PDT-Surf* is 161 compared to 1 662 in *CNC-Surf*.

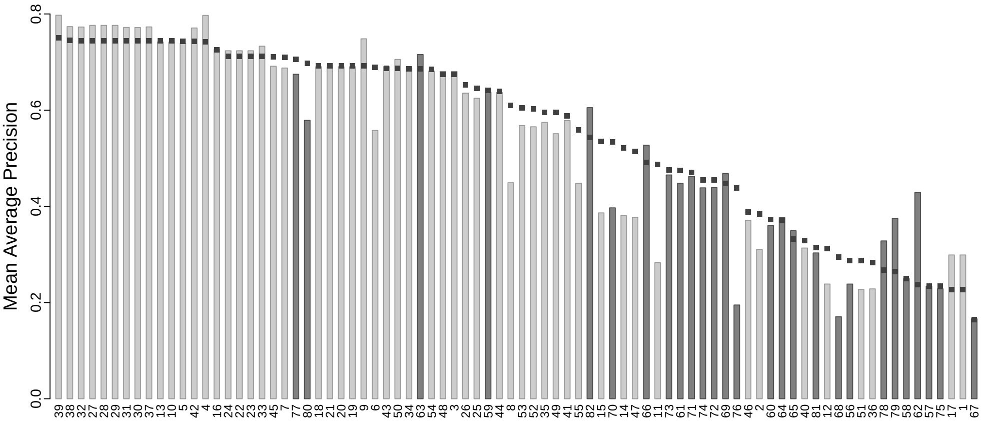

Figure 5.6: MAP scores of association measures computed on the *CNC-Surf* data set (bars) and sorted by the MAP obtained on the *PDT-Surf* data set (square points) in descending order.

The results are presented in Figure 5.6 and compared to those obtained on the *PDT--Surf* data set (again for a straightforward comparison). The effect of using a much larger data set is positive only for certain methods – surprisingly the most efficient ones. A significant improvement (4.5 absolute percentage points on average) is observed only for a few of the best performing association measures on *PDT-Surf* and also for some other less efficient methods. Performance of other association measures did not significantly change or it dropped down. The two absolute winners are *Unigram subtuple measure (39)* with MAP=79.74% and *Pointwise mutual information (4)* with MAP=79.71%, known to be very efficient on large corpora.

5.2.3 Swedish PAROLE Corpus

The *PAR-Dist* data set, on which we carried out this last experiment, differs in more aspects. It contains support-verb construction candidates extracted as distance bigrams (allowing up to three words occurring within the distance between components) from the 20 million word Swedish PAROLE corpus. Also, no frequency filter was applied to this data set. A baseline system ranking the *PAR-Dist* candidates randomly would operate with the expected precision of 7.59%, which is significantly lower than for the other data sets and thus the MAP of the association measures is expected to be lower.

Sorted MAP scores of the association measures are presented in descending order as the square points in Figure 5.7. The best performing measures evaluated on this data set are *Michael's coefficient (36)* with MAP=18.88%, *Piatersky-Shapiro's coefficient (51)* with MAP=18.87%, and *T-test (12)* with MAP=18.66%. Obviously, the scores are statistically indistinguishable (the paired Wilcoxon signed-rank test, α=0.05). The appearance of *T-test (12)* among the best measures is quite surprising because it per-

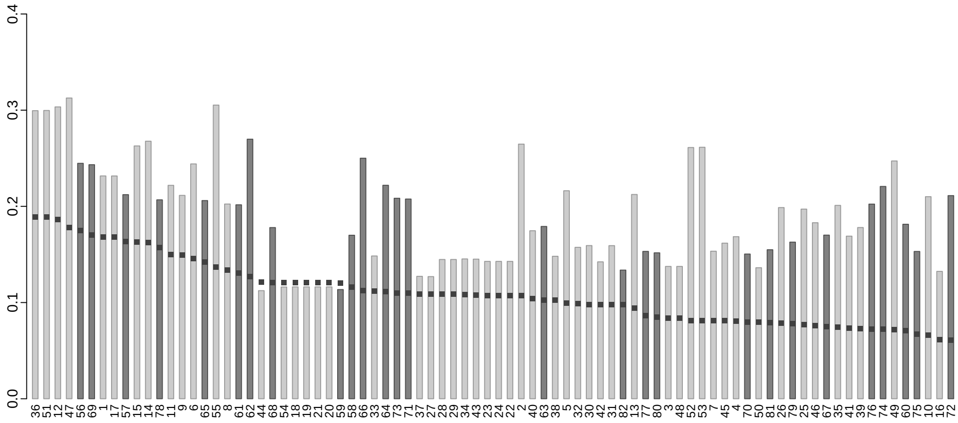

Figure 5.7: MAP scores of association measures computed on a subset of *PAR-Dist* (f > 5) (bars) and sorted by the descending scores of MAP obtained on the full *PAR-Dist* set (square points).

formed only slightly above the baseline precision on the other data sets. In fact, the results of other measures are also remarkably different and many of the best performing measures on other data sets appear in the tail (Figure 5.7).

The evaluation over the *PAR-Dist* data set might have been unfairly biased by the low frequency candidates that were not filtered out by the frequency filter as was the case with the other data sets. Hence, we applied the frequency filter to this set and preserved only the candidates appearing in the corpus more than five times (the same frequency threshold as for *PDT-Dep*, *PDT-Surf*, and *CNC-Surf*). The resulting set contains 5 530 candidates including 763 true collocations (the baseline precision is 13.79%). MAP scores of this reduced data set are visualized as bars and compared to the original ones (the square points) also in Figure 5.7.

Most of the association measures are indeed very sensitive to low frequency data and the MAP scores on the filtered and the full *PAR-Dist* data set do not correlate much. The best scores were achieved by *Gini index (47)*, MAP=31.27%, *Klosgen's coefficient (55)*, MAP=30.53%, and *T-test (12)*, MAP=30.34% (all insignificantly different). Surprisingly, *T-test (12)* is again among the best measures. Compared to the results on the full *PAR-Dist* set (18.87%), the MAP scores of the best measures are greater than what could be explained by the difference between the baseline precisions.

Figure 5.8 compares MAP scores on the full *PAR-Dist* data set and the *PDT-Dep* data set. It is evident that the performance of the individual measures varies to a large extent also in this case. While *Pearson's χ^2 test (10)* is the third worse method on *PAR-Dist*, it is among the best (statistically indistinguishable) methods on *PDT-Dep*. On the contrary, *T-test (12)* is in the group of the best (statistically indistinguishable) methods on *PAR-Dist*, but on *PDT-Dep*, it is among the methods with the lowest MAP.

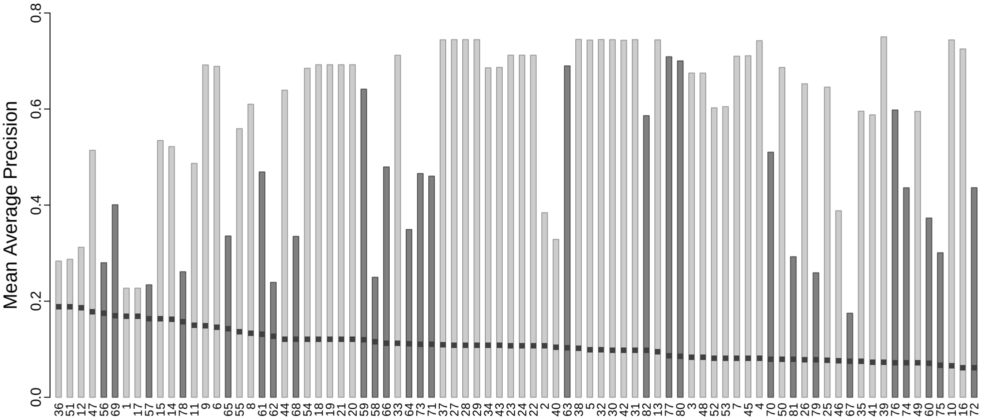

Figure 5.8: MAP scores of association measures obtained on the *PDT-Surf* data set (bars) and sorted by the descending scores of MAP measured on the *PAR-Dist* data set (square points).

5.3 Comparison

When comparing results on these data sets, we must be aware of the fact that the baseline MAP scores on these data sets are not equal (21.02% for *PDT-Dep*, 22.88% for *PDT-Surf*, 22.66% for *CNC-Surf*, and 7.59% for *PAR-Dist*) and their differences must be taken into account during the analysis of the MAP scores on different data sets. In most cases, these differences are relatively small compared to the differences in MAP of association measures that were observed in our experiments.

The complete results of all the experiments described in this chapter (including the significance tests) are presented in Appendix B. To make the picture even more complete, we have visualized how the results vary on the data sets by drawing their scatterplots in Figure 5.9. Each of the plots in the matrix contains the MAP of all association measures obtained on one data set plotted against the MAP on another data set. Each point represents two MAP scores of a particular association measure on two data sets. Fully correlated MAP scores on two data sets would appear on the diagonal of the corresponding plot. A certain correlation is observed between the results on the *PDT-Dep* and *PDT-Surf* data sets and also between *PDT-Surf* and *CNC-Surf* (which are most similar data set pairings). Significantly less correlated are the MAP scores on *CNC-Surf* and *PDT-Dep*, and basically no correlation is observed between the results obtained on the *PAR-Dist* and the other data sets.

Based on this observation, we can conclude that the performance of association measures on our data sets varies to a large extent and depends on every aspect of the task, such as the type of collocations being extracted, the way the candidates were obtained, the size of the source corpora, its language, etc.

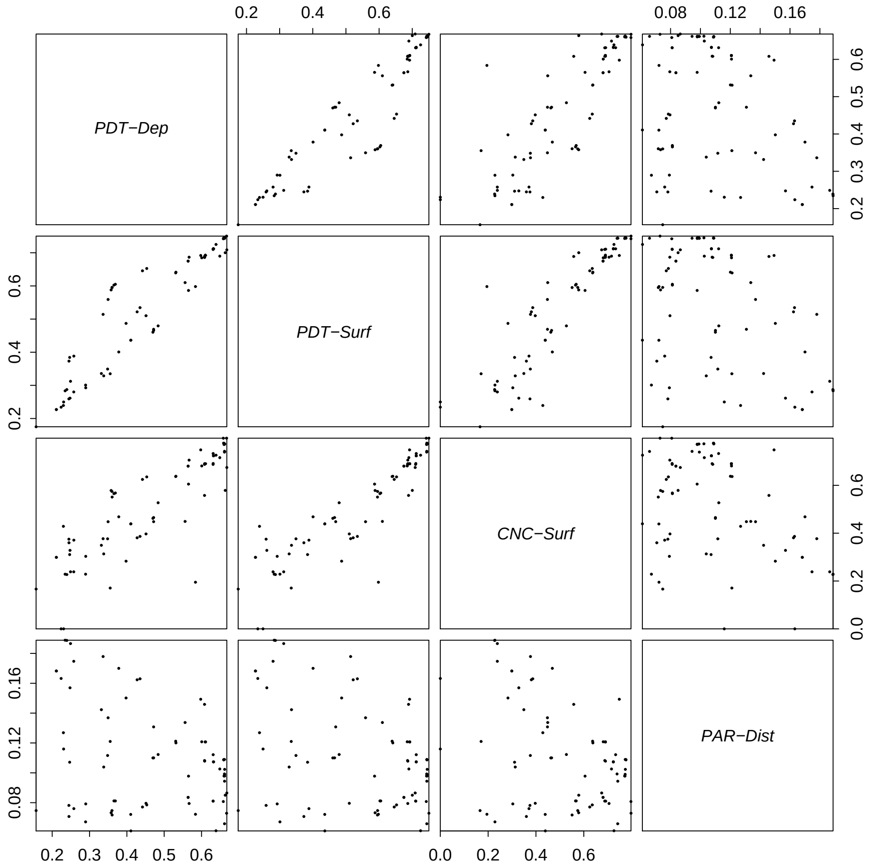

Figure 5.9: A matrix of scatterplots of MAP scores of all association measures computed on all the four data sets (*PDT-Dep*, *PDT-Surf*, *CNC-Surf*, and *PAR-Dist*). Each point in a particular scatterplot represents MAP scores of one measure obtained on the two corresponding data sets.

Although we are not able to recommend a measure (or measures) that perform successfully on any data (or task), the presented evaluation scheme can be effectively used to choose such a measure (or measures) for any particular task (assuming a manually annotated reference data set is available).

6
Combining Association Measures

In this chapter, we propose combining association measures into more complex statistical models that can exploit the potential of the individual association measures to discover different groups and types of associated words.

6.1 Motivation

It is quite natural to expect that the collocation extraction methods (especially those based on different extraction principles) rank collocation candidates differently. In the previous chapter, we used the mean average precision (MAP) as a measure of quality of such a ranking: methods that better concentrate true collocations at the top of the list were evaluated as more efficient than the others. Many measures achieved very similar MAP scores for a given data set and were evaluated as equally good. For example, *Cosine context similarity in boolean vector space (77)* and *Unigram subtuple measure (39)* performed on *PDT-Dep* with statistically indistinguishable scores of MAP=66.79% and 66.72%, respectively.

In a more thorough comparison by precision-recall (PR) curves, we observed that on *PDT-Dep*, the curve of *Cosine context similarity (77)* significantly predominates the curve of *Unigram subtuple measure (39)* in the first half of the recall interval and vice versa in the second half, as depicted in Figure 5.2. This is a case where MAP is not a suitable metric for comparing the performance of association measures. For a more detailed comparison, we should analyze not only their MAP but also their PR curves. Moreover, even if two methods have identical PR curves, the actual ranking of collocation candidates can still vary a lot and different association measures can prefer different types (or groups) of collocations above others. Such *non-correlated* measures could (perhaps) be combined and eventually improve the performance in ranking collocation candidates. An example of existence of such measures is shown in Figure 6.1. Association scores of *Pointwise mutual information (4)* and *Cosine context similarity (77)* seem independent enough to be (linearly) combined into one model and possibly achieve better performance.

6.2 Methods

Formally, each collocation candidate x^i can be empirically described by the **feature vector** $\mathbf{x}^i = (x^i_1, \ldots, x^i_{82})^T$ consisting of scores of all 82 association measures from Tables 3.4 and 3.5 in Chapter 3 and assigned a label $y^i \in \{0, 1\}$ which indicates whether

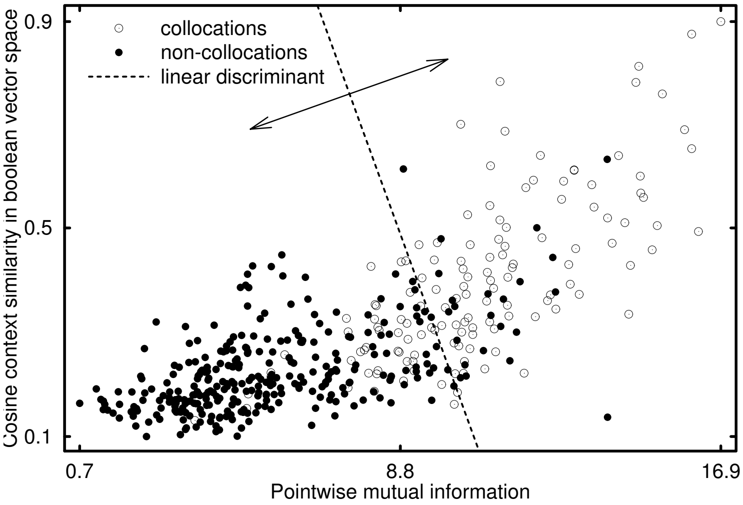

Figure 6.1: Visualization of scores of two association measures. The dashed line denotes a linear discriminant obtained by logistic linear regression. By moving this boundary, we can tune the classifier output or use it as a ranker (a 5% stratified sample of the evaluation data is displayed).

the bigram is considered to be a true collocation ($y = 1$) or not ($y = 0$). We look for a **ranker** function $f(\mathbf{x}^i) \to R$ that determine the strength of collocational association between components of collocation candidates (x^i) and hence can be used for their ranking in the same manner as the individual association measures. Performance of such a method could be evaluated in the same way as the individual association measures: MAP scores and PR curves. In this section, we briefly introduce several statistical-classification methods and demonstrate how we used them as such rankers. For further details on these methods, we refer to Venables and Ripley (2002).

6.2.1 Linear logistic regression

A generalized linear model (GLM) in a form of logistic regression is an additive model for a binary response represented by:

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n,$$

where $\text{logit}(\pi) = \log(\pi/(1-\pi))$ is a canonical link function for odds ratio and $\pi \in (0, 1)$ is a conditional probability of a positive response given a vector $\mathbf{x}$. The estimation of β_0 and β is computed by the maximum likelihood method which is solved by the *iteratively reweighted least squares* algorithm. The ranker function in this case is defined as the predicted value $\widehat{\pi}$ or equivalently (due to the monotonicity of the logit link function) as the linear combination $\widehat{\beta}_0 + \widehat{\beta}^T\mathbf{x}$.

6.2.2 Linear discriminant analysis

The basic idea of Fisher's linear discriminant analysis (LDA) is to find a one-dimensional projection defined by a vector $\mathbf{c}$ so that for the projected combination $\mathbf{c}^T\mathbf{x}$ the ratio of the *between variance* $\mathbf{B}$ to the *within variance* $\mathbf{W}$ is maximized. After the projection, $\mathbf{c}^T\mathbf{x}$ can be used directly as a ranker.

$$\max_{\mathbf{c}} \frac{\mathbf{c}^T \mathbf{B}\, \mathbf{c}}{\mathbf{c}^T \mathbf{W}\, \mathbf{c}}.$$

6.2.3 Support vector machines

For technical reasons, we now change the labels $y^i \in \{-1, +1\}$. The goal in support vector machines (SVM) is to estimate a function $f(\mathbf{x}) = \beta_0 + \beta^T \mathbf{x}$ and find a classifier $y(\mathbf{x}) = \text{sign}\big(f(\mathbf{x})\big)$ which can be solved through the following convex optimization:

$$\min_{\beta_0, \beta} \sum_{i=1}^{n} \left[1 - y^i(\beta_0 + \beta^T \mathbf{x}^i)\right]^+ + \frac{\lambda}{2}\|\beta\|^2.$$

with λ as a regularization parameter. The *hinge loss function* $L(y, f(\mathbf{x})) = [1 - yf(\mathbf{x})]^+$ is active only for positive values (i.e. bad predictions) and is therefore very suitable for ranking models with $\widehat{\beta}_0 + \widehat{\beta}^T \mathbf{x}$ as a ranker function. Setting the regularization parameter λ is crucial for both the estimators $\widehat{\beta}_0, \widehat{\beta}$ and further classification (or ranking). As an alternative to the often inappropriate grid search, Hastie et al. (2004) proposed an effective algorithm which fits the entire SVM regularization path $[\beta_0(\lambda), \beta(\lambda)]$ and provided an option to choose the optimal value of λ. As an objective function, we used the total amount of loss on training data rather than the number of false predicted training instances.

6.2.4 Neural networks

Assuming the most common model of neural networks (NNet) with one hidden layer, the aim is to find inner weights w_{jh} and outer weights w_{hi} for

$$y^i = \phi_0\big(\alpha_0 + \sum w_{hi}\phi_h(\alpha_h + \sum w_{jh}x_j)\big),$$

where h ranges over the units in the hidden layer. Activation functions ϕ_h and the function ϕ_0 are fixed. Typically, ϕ_h is taken as the logistic function $\phi_h(z) = \exp(z)/(1+\exp(z))$ and ϕ_0 as the indicator function $\phi_0(z) = I(z > \Delta)$ with Δ as a classification threshold. For ranking, we simply set $\phi_0(z) = z$. Parameters of the neural networks are estimated by the *backpropagation algorithm*. The loss function can be based either on *least squares* or *maximum likelihood*. To avoid problems with convergence of the algorithm, we used the former one. The tuning parameter of a classifier is then the number of units in the hidden layer.

method	*averaged precision at*			*MAP*	
	R=20	*R=50*	*R=80*	*R=⟨0.1,0.9⟩*	+%
Neural network (5 units)	91.00	81.75	70.22	80.87	21.08
Linear logistic regression	86.96	79.74	64.63	77.36	15.82
Linear discriminant analysis	85.99	77.34	61.44	75.16	12.54
Neural network (1 unit)	82.47	77.08	65.75	74.88	12.11
Support vector machine (linear)	81.33	76.08	61.49	73.03	9.35
Cosine similarity (77)	80.88	68.46	49.99	66.79	0.00
Unigram subtuples (39)	75.86	68.19	55.13	66.72	–

Table 6.1: Performance of methods combining all association measures on *PDT-Dep*: averaged (over the data folds) precision at fixed points of recall and mean average precision and its relative improvement (+%) compared to the best individual association measure (all values in %).

Training and application

The presented methods are originally intended for (binary) classification. For our purposes, they are used with a small modification: in the training phase, they are used as regular classifiers on two-class training data (collocations and non-collocations) to fit the model parameters, but in the application phase, no classification threshold applies and for each collocation candidate, the ranker function computes a value which is interpreted as the association score. Applying the classification threshold would turn the ranker back into a regular classifier. The candidates with higher scores would fall into one class (collocations), the rest into the other class (non-collocations).

6.3 Experiments

In this section, we describe experiments with the models presented above on the four reference data sets described in Chapter 4. The results will be evaluated by MAP scores and PR curves, and compared to the performance of the best individual measures evaluated in Chapter 5.

To avoid incommensurability of association measures in the experiments, we have used the most common preprocessing technique for multivariate **standardization**: the values of each association measure were centered towards zero and scaled to a unit variance. Precision-recall curves of all methods were obtained by vertical averaging in 6-fold crossvalidation on the same reference data sets as in the earlier experiments. Mean average precision was computed from the average precision values estimated on the recall interval ⟨0.1, 0.9⟩. In each crossvalidation step, five folds were used for training and one fold for testing.

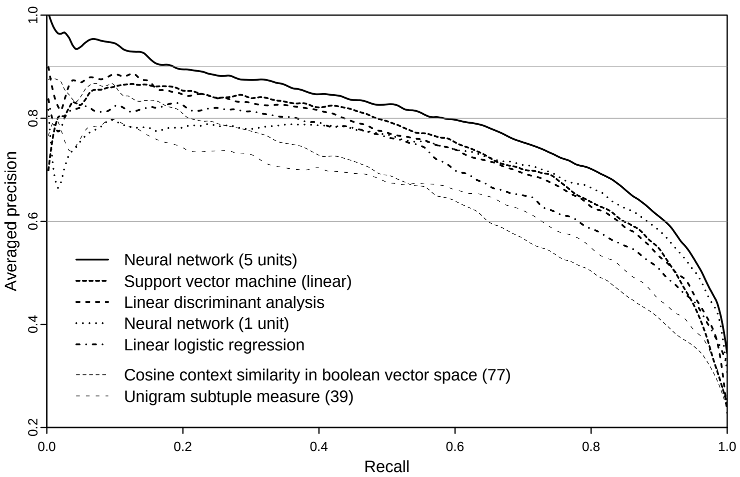

Figure 6.2: Precision-recall curves of selected methods combining all association measures on *PDT-Dep*, compared with curves of two best measures employed individually on the same data.

6.3.1 Prague Dependency Treebank

First, we studied the performance of the combination methods on the *PDT-Dep* reference data. All combination methods worked very well and gained a substantial performance improvement in comparison with individual measures. The best result was achieved by the neural network with five units in the hidden layer (NNet.5) with MAP=80.93%, which is 21.17% relative and 14.08% absolute improvement compared to the best individual association measures, such as *Cosine context similarity in boolean vector space (77)* and *Unigram subtuple measure (39)*. More detailed results are given in Table 6.1 and precision-recall curves are depicted in Figure 6.2. We observed a relatively stable improvement within the whole interval of recall. The neural network was the only method which performed better in its more complex variants (with up to five units in the hidden layer). More complex models, such as neural networks with more than five units in the hidden layer, support vector machines with higher order polynomial kernels, quadratic logistic regression, or quadratic discriminant analysis, overfitted the training data folds and better scores were achieved by their simpler variants.

The results on the *PDT-Surf* data set were similar. The best method was also NNet.5. It achieved even higher MAP=84.84% but compared to the best performing individual measure *Unigram subtuple measure (39)* with MAP=75.03%, the relative improvement was only 12.43%.

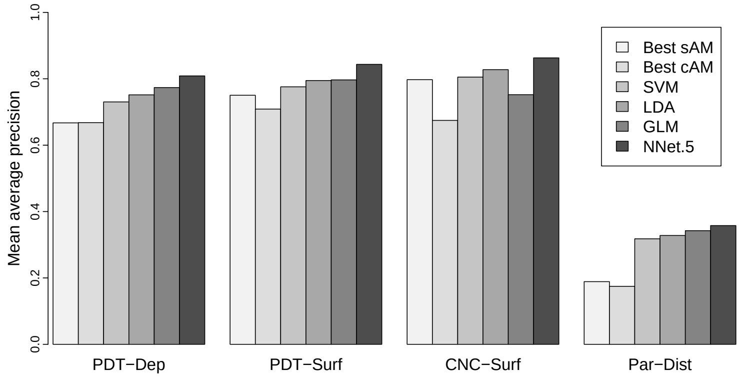

Figure 6.3: MAP scores of methods combining all association measures obtained on the reference data sets: *PDT-Dep*, *PDT-Surf*, *CNC-Surf*, and *PAR-Dist*. *Best sAM* and *Best cAM* refer to the best statistical association measure and context-based measure on each data set, respectively.

6.3.2 Czech National Corpus

The *CNC-Surf* data set provides a much better estimation of the occurrence probabilities of the collocation candidates and their components. Also the context information extracted for the candidates in this data set from the Czech National corpus is much more representative. The best individual association measures evaluated on *CNC-Surf* gained about 4.5% (absolute) compared to the results on *PDT-Surf* (the same collocation candidates but frequency and context information extracted from the much smaller Prague Dependency Treebank). The best method on *CNC-Surf*, *Unigram subtuple measure (39)*, achieved MAP=79.74% and NNet.5 combining all association measures then increased this score to a remarkable 86.3%.

By taking the *CNC-Surf* data set as a representative sample of all collocation candidates from the whole Czech National Corpus (filtered by the same part-of-speech and frequency filter) we can use this MAP score as an estimation of MAP that can be achieved by this method on the full population of candidates from this corpus (which is 1.5 million surface bigrams, see Table 4.12). *Any* portion of true collocations in this population can be extracted by this neural network with the expected precision 86.3%. If we limit ourselves to a specific recall, we can extract e.g. 20% of true collocations with an expected precision of 94.07%, 50% of true collocations with an expected precision of 88.09% and 80% of true collocations with an expected precision of 75.62% (these values are averaged precision scores at 20%, 50%, and 80% of recall obtained by NNet.5 on *CNC-Surf*, respectively).

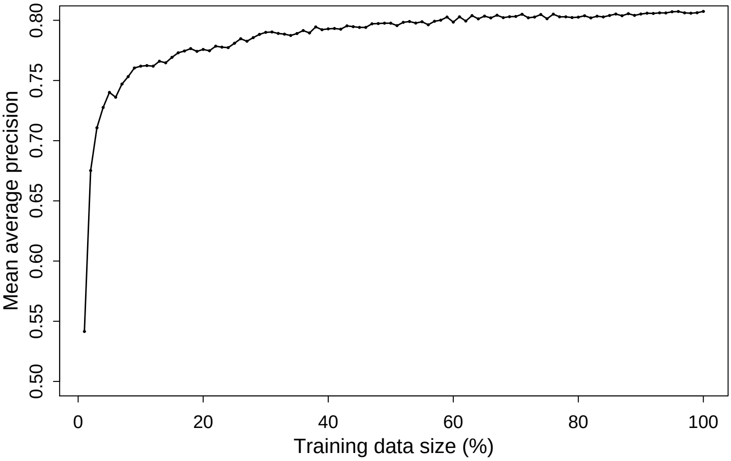

Figure 6.4: The learning curve (MAP with respect to training data size) of the neural network with 5 units in the hidden layer measured on the *PDT-Dep* reference data set.

6.3.3 Swedish PAROLE Corpus

The comparison of the performance of all the combination methods on all the reference data sets is depicted in Figure 6.3. NNet.5 was evaluated as the best performing method also on the *PAR-Dist* reference data set. It achieved MAP=35.78%, which is, compared to the best individual measure on the same data set, *Michael's coefficient(36)*, with MAP=18.88%, a substantial improvement of 89.5% (relative). Based on the suspicion that the evaluation on the (full) *PAR-Dist* data set (see also Section 5.2.3) might be biased by the low frequency candidates, we limited another experiment to the subset of candidates with frequency greater than five. The best MAP score of individual association measure *Gini Index (47)* was 31.27%. The same neural network model on this subset achieved MAP=52.15% which is also quite a substantial improvement of 66.76% (relative).

Learning curve analysis

Our next experiment is focused on the effect of using different amounts of data for training the combination models. The experiments presented so far in this chapter were based on 6-fold crossvalidation (see Section 6.3). They used five out of the six evaluation folds for training (fitting model parameters) and one fold for testing (predicting association strength). For example, in each crossvalidation step on *PDT-Dep*, 8 737 data instances (candidates labeled as collocations or non-collocations) were used for training and other 1 747 for testing. The first question is whether such an amount of training data is sufficient or whether we would profit from having more data available

method	*PDT-Dep*	*PDT-Surf*	*CNC-Surf*	*PAR-Dist*
NNet.5 (AM+POS+DEP)	**84.53**	–	–	–
NNet.5 (AM+POS)	82.79	**86.48**	**88.22**	–
NNet.5 (AM)	80.87	84.35	86.30	**35.78**
Best AM	66.72 (77)	75.03 (39)	79.74 (39)	18.88 (36)
Baseline	21.02	22.88	22.66	7.59

Table 6.2: Summarization of the results achieved on each data sets by the best individual association measure (Best AM) and the best combination method (NNet.5) using association measures (AM), information about part-of speech pattern (POS) and dependency type (DEP).

for training. In case we have enough data for training, the second question is whether its amount is not unnecessarily large and whether we can train a well-performing model on less data.

We have repeated the experiment with NNet.5 on *PDT-Dep* with a varying proportion of data used for training (the data used for testing did not change). The experiment ran over 100 iterations. It started with 1% of data used for training (87 instances) in each of the six crossvalidation steps and in every subsequent iteration, we added another 1% of the data for training. The MAP scores computed after each iteration of this experiment are depicted in Figure 6.4. The resulting curve is called a **learning curve** and is a commnon tool for the analysis of model performance in dependency on the size of the training data. The beginning of the curve obtained by NNet.5 on *PDT-Dep* is fairly steep and reaches 90% of its maximum value with only 5% of training data; with 15% of training data, it climbs up to 95%; 99% of the maximum MAP score can be achieved with about 50% of training data.

We expect the learning curve to stay flat even when using more data, and thus we can conclude that the amount of data we used in our experiments is sufficient. Moreover, we can use significantly less data and train a very well-performing system with as little as 15% of the original amount of the training data. The effect of using more than approximately 60% of the data is within the statistical error (by paired Wilcoxon signed-rank test).

6.4 Linguistic features

In the following experiment, we attempted to improve the combination methods by using some linguistic information extracted with the collocation candidates from the source corpora, namely part-of-speech patterns and dependency types. This information was incorporated into the models by *binarization* and *dummy variables* (Boros et al., 1997) for each possible value of the part-of-speech pattern and dependency type, indicating presence or absence of the value for each data instance (collocation candidate).

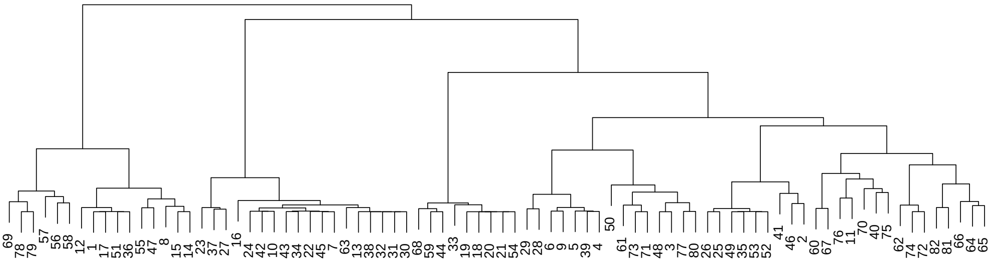

Figure 6.5: Dendrogram visualizing hierarchical clustering of association measures based on their (pairwise) correlation over the held-out data fold from the *PDT-Dep* data set.

The linguistic information contributed to the models quite significantly. The MAP scores of the best performing method (NNet.5) exploiting this kind of information on the reference data sets are shown in Table 6.2. Using POS information improved the MAP scores of NNet.5 approximately by 2% (absolute) on all Czech data sets (the Swedish *PAR-Dist* contains only verb-noun combinations). Additional 2% (absolute) were gained on *PDT-Dep* by exploiting the information on the dependency type (the only data set containing this kind of information) and the best performing method achieved MAP=84.53% which is a relative improvement of 25.94% compared to MAP of the best individual measure.

6.5 Model reduction

In the previous sections, we have demonstrated that combining association measures is generally very reasonable and significantly helps in the task of ranking collocation candidates. However, the methods which employ all 82 association measures in linear combination (or more complex models, such as the neural networks with multiple units in the hidden layer) are unnecessarily complex (in the number of the variables used), mainly for the following two reasons:

First, some of the association measures are too **similar** (analytically or empirically) – when combined they do not bring any new information and become redundant. Such highly correlated measures make the training (fitting the models) quite difficult and should be eliminated. After applying *principal component analysis* (see e.g. Jolliffe, 2002)) to the all 82 association scores of collocation candidates from the *PDT-*-*Dep* reference data, we observed that 95% of the total variance is explained by only 17 principal components and 99.9% is explained by 42 components. We expect to able to significantly reduce the number of variables in our models, possibly with a very limited degradation of their performance.

Second, some of the measures are **improper** for ranking collocation candidates at all – they do not determine well the strength of association, bring unnecessary noise to

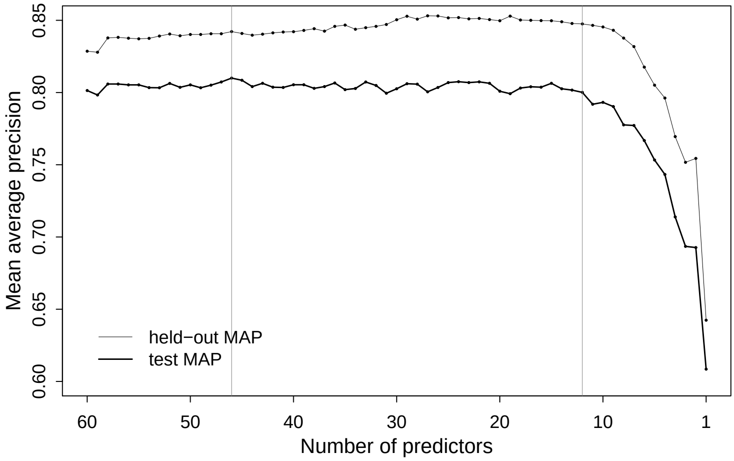

Figure 6.6: MAP scores obtained after each iteration of the model reduction of NNet.5 on *PDT-Dep* initiated with 60 variables (lower curve, scores were crossvalidated on the evaluation folds) and MAP scores on the held-out fold used to select the variables to be removed (upper curve).

the combination models, and eventually, they can also hurt their performance. Also such measures should be identified and removed from the model. In this section, we attempt to propose an algorithm, which reduces the combination models by removing such redundant (correlated) and useless (non-efficient) variables.

A straightforward, but in our case hardly feasible (due to the high number of the model variables), approach would be an exhaustive search through the space of all possible subsets of all the association measures. Another option is a heuristic *step-wise* algorithm iteratively removing one variable at a time until a stopping criterion is met. Such algorithms are not very robust: they are particularly sensitive to data and generally not recommended. That is why we tried to minimize these problems by initializing the algorithm with clustering variables with the same contribution to the model and choosing only one measure from each cluster as a representative. Thus we can reduce the highly correlated variables and continue with the step-wise procedure.

6.5.1 Algorithm

The proposed algorithm eliminates the model variables (association measures) based on two criteria: linear correlation with other variables and poor contribution to efficient ranking of collocation candidates.

First, a **hierarchical clustering** (Kaufman and Rousseeuw, 1990) is employed in order to group highly correlated measures into clusters. This clustering is based on the similarity matrix formed by the absolute values of *Pearson's correlation coefficient* computed for each pair of association measures estimated from the held-out data fold

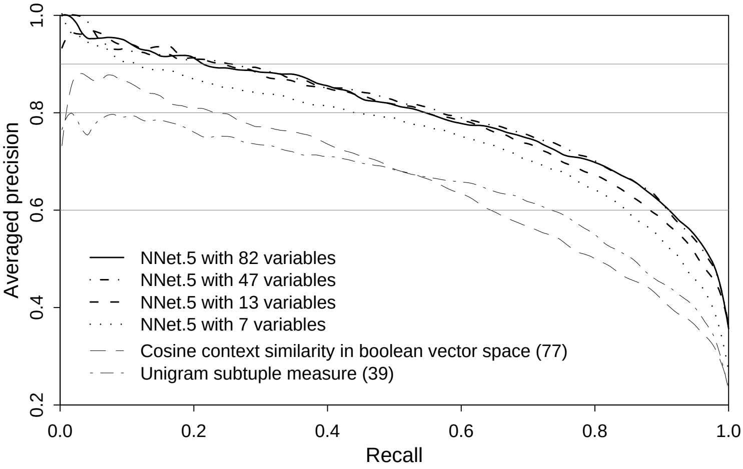

Figure 6.7: Precision-recall curves of the reduced NNet.5 models compared with the curves of the full model and two best individual methods on the *PDT-Dep* data set.

(independent from the evaluation data folds). This technique starts with each variable in a separate cluster and merges them into consecutively larger clusters based on the values from the similarity matrix until a desired number of clusters is reached or the similarity between clusters exceeds a limit.

An example of a complete hierarchical clustering of association measures is depicted in Figure 6.5. If the stopping criterion is set correctly the measures in each cluster have an approximately equal contribution to the model. Only one of them is selected as a representative and used in the reduced model (the other measures are redundant). The selection can be random or based e.g. on the (absolute) individual performance of the measures on the held-out data.

The reduced model at this point does not contain highly-correlated variables and can be more easily fit (trained) to the data. However, these variables are not guaranteed to have a positive contribution to the model. Therefore, the algorithm continues with the second step and applies a standard **step-wise** procedure removing one variable in each iteration, causing minimal degradation of the model's performance measured by MAP on the held-out data fold. The procedure stops when the degradation becomes statistically significant – e.g. by the paired t-test or paired Wilcoxon signed-rank test.

6.5.2 Experiments

We tested the model reduction algorithm with NNet.5 (as the best performing combination method) on the *PDT-Dep* reference data set as follows: The initial hierarchical clustering was stopped after merging the variables into 60 clusters (the number was

#	*association measure*	*MAP*
13.	Reverse cross entropy (62)	22.98
12.	First Kulczynsky coefficient (23)	63.21
11.	**S** cost (41)	35.77
10.	Left context entropy (57)	22.38
9.	Reverse confusion probability (68)	35.53
8.	Left context divergence (59)	53.14
7.	Phrase word cooccurrence (75)	28.94
6.	Right context entropy (58)	23.05
5.	Cosine context similarity in boolean vector space (77)	66.79
4.	Dice context similarity in TF vector space (81)	28.98
3.	Unigram subtuple measure (39)	66.72
2.	Dice context similarity in TF·IDF vector space (82)	56.51
1.	Log frequency biased Mutual Dependency (6)	60.81

Table 6.3: Association measures (with their individual MAP scores) of the final model of the reduction algorithm applied to NNet.5 on *PDT-Dep* in the order they would be further removed.

set experimentally). In each iteration step of the algorithm, we estimated performance of the current model reduced by each variable (one by one) on the held-out data fold: six crossvalidation models were trained as usual on five of the evaluation folds and tested not on the sixth one but on the *held-out* fold (so the MAP score was estimated from six different rankings of candidates from one data fold). The variable causing minimal degradation of this score was selected and removed from the model. The new model was evaluated as usual on all the *evaluation* folds and the obtained MAP score was tested to find out if it is significantly worse than the one from the previous step. The decision which variable to remove in each iteration was done independently of the performance evaluation of the intermediate models.

Figure 6.6 displays the MAP scores of the intermediate models from the whole process. It started with 60 variables, the best MAP was achieved by a model with 47 variables. The MAP scores further oscillated around the same value until the model had about 16 variables. Then, MAP dropped down a little after each iteration and with less than 13 variables this degradation became significant (the paired Wilcoxon signed-rank test, confidence level $\alpha = 0.05\%$) which is even smaller than the number of principal components that explain 95% of the sample variance as mentioned earlier.

Precision-recall curves for some intermediate models are shown in Figure 6.7. We can conclude that we were able to reduce the NNet.5 model to 13 variables without a statistically significant difference in performance, MAP=80.18%. The final model contained the association measures listed in Table 6.3 in the order in which they would be removed if the algorithm continued. They include measures across the entire spectrum, based on different extraction principles, and with very different individual performance. The precision-recall curves of these measures are depicted in Figure 6.8.

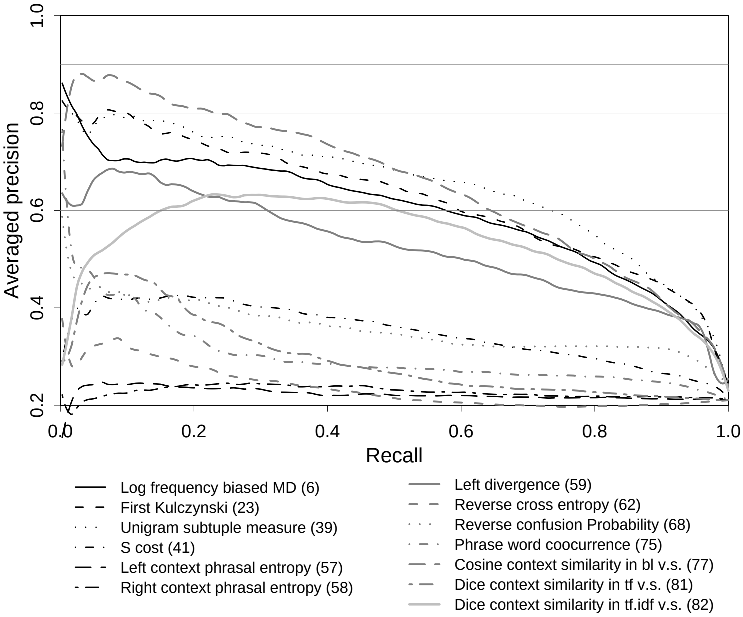

Figure 6.8: Averaged precision-recall curves of the 13 measures (applied individually on the *PDT-Dep* data set) included in the reduced combination model (NNet.5).

Some of the measures/variables of the final model (e.g. *57, 58, 62*) performed only very slightly above the baseline when employed individually, however their contribution to the model is perceptible – if any of them was removed from the model, the model's performance would drop significantly (measured by the paired Wilcoxon signed-rank test at the confidence level α = 0.05%). If we let the model reduction algorithm make one step more, it would remove the measure *(62)* with individual MAP=22.98% (which is less than absolute 2% above the baseline) and the model's MAP would drop to 79.37% (which was confirmed to be a significant difference by the paired Wilcoxon signed-rank test). If this difference (and the contribution of such poorly performing measures) was not interpreted as "practically" significant and we removed all measures with MAP less than 25% *(57, 58, 62)*, the model's MAP would drop to 76.54% – i.e. the three "poor" methods contribute to the model's MAP by almost 4% absolute.

It should be emphasized that the model-reduction algorithm is very sensitive to data and can very easily lead to different results depending on the task. However, we employed the reduced NNet.5 models with the 13 variables on the other data sets and it also performed very well, although in some cases, the differences are statistically significant (see Table 6.4).

	PDT-Dep		*PDT-Surf*		*CNC-Surf*		*PAR-Dist*	
model	*full*	*red*	*full*	*red*	*full*	*red*	*full*	*red*
NNet.5 (AM+POS+DEP)	84.53	84.16	–	–	–	–	–	–
NNet.5 (AM+POS)	82.79	82.51	86.48	86.33	88.22	87.58	–	–
NNet.5 (AM)	80.87	80.18	84.35	83.81	**86.30**	**85.01**	**35.78**	**33.19**
Best AM	66.72	(77)	75.03	(39)	79.74	(39)	18.88	(36)
Baseline	21.02		22.88		22.66		7.59	

Table 6.4: Comparison of the MAP scores of the full and reduced (13 variables) NNet.5 models on all the reference data sets. Significantly different scores are in bold.

7
Conclusions

In this work, we have studied lexical association measures and their application to collocation extraction. First, we have compiled a comprehensive inventory of 82 lexical association measures for two-word (bigram) collocation extraction based on three different extraction principles. These measures are divided into two groups: *statistical association* measures and *context-based association* measures.

Second, we have developed four reference data sets for the task of identifying collocation candidates. All of them consist of bigram collocation candidates. *PDT-Dep* and *PDT-Surf* were extracted from the Czech manually annotated *Prague Dependency Treebank* and differ only in the character of the bigrams: *PDT-Dep* consists of dependency bigrams while *PDT-Surf* comprises surface bigrams. Both the sets were filtered by the same part-of-speech pattern and frequency filters. Manual annotation was done exhaustively by three annotators, *true collocations* were indicated in all the data. The *CNC-Surf* reference data set was extracted from a substantially larger data from the *Czech National Corpus* and consists of the surface bigrams also appearing in *PDT-Surf*. This data set can be considered as a random sample from the full set of collocation candidates in this corpus filtered by the same part-of-speech pattern filter and frequency filter as the *PDT-Surf* reference data. The *PAR-Dist* reference data set is quite different. It consists of Swedish verb-noun combinations manually extracted from the Swedish *PAROLE* corpus in a nonexhaustive fashion with an indication of *true support-verb constructions*.

These four reference data sets were designed to allow comparison of effectiveness of the association measures in different settings. On *PDT-Dep* and *PDT-Surf*, we have compared two ways of extracting collocation candidates (dependency vs. surface bigrams). On *PDT-Surf* and *CNC-Surf*, we have explored the effect of using a much larger source corpus (1.5 million vs. 242 million tokens). *PAR-Dist* complements these three sets with the data that differs in more aspects: the language (Swedish vs. Czech), the way the candidates were obtained (distance vs. dependency or surface bigrams), the type of collocations being extracted (support-verb constructions vs. general collocations), the size of the source corpora (20 million vs. 1.5 million or 242 million tokens), and finally, the frequency filter (all candidates vs. those occurring more than five times).

We implemented all the 82 lexical association measures and evaluated their performance in ranking collocation candidates over the four reference data sets by averaged *precision-recall* (PR) curves and *mean average precision* (MAP) scores in 6-fold cross validation. The baseline scores were set as the expected MAP of a system that would rank the collocation candidates in each of the reference data sets randomly, which corresponds to the prior probability of a collocation candidate to be a true collocation: 21.02% for *PDT-Dep*, 22.88% for *PDT-Surf*, 22.66% for *CNC-Surf*, and 7.59% for *PAR-Dist*.

The best result on the *PDT-Dep* reference data was achieved by a context-based method measuring *Cosine context similarity in boolean vector space* with MAP=66.79% followed by 15 other association measures with statistically indistinguishable performance. Extracting collocations as surface bigrams was observed to be a more efficient approach (in terms of higher MAP). The results of almost all measures obtained over the *PDT-Surf* reference data significantly improved: the best MAP=75.03% was achieved with *Unigram subtuple measure* followed by 13 other measures with statistically insignificant differences in MAP. The experiments carried out on the *CNC-Surf* reference data showed that processing of a larger corpus had a positive effect on the quality of collocation extraction; the MAP score of the best measures, *Unigram subtuple measure* and *Pointwise mutual information*, increased up to 79.7%. The results on the *PAR-Dist* reference data set were remarkably different not only in the absolute MAP scores of the best methods (*Michael's coefficient*, *Piatersky-Shapiro's coefficient*, and *T-test* with statistically indistinguishable MAP=18.66–18.88%) but also in the relative difference of their performance over the other data sets. For example, *T-test*, one of the best measures on *PAR-Dist*, performed only slightly above the baseline across all *PDT-Dep*, *PDT-Surf*, and *CNC-Surf*. These results demonstrate that performance of lexical association measures strongly depends on the actual data and task. None of the measures can be selected as the "best" measure that would perform efficiently on any data set. However, the proposed evaluation scheme (based on MAP scores and eventually also on PR curves) can be effectively used to choose such a measure (or measures) for any particular task (if manually annotated data is available).

Further, we have demonstrated that by combining association measures, we can achieve a substantial performance improvement in ranking collocation candidates. The lexical association measures presented in this work and used as ranking functions provide scores that are uncorrelated to such an extent that a linear combination of all of them produces better association scores than any of the measures employed individually. All investigated combination methods (*linear logistic regression*, *linear discriminant analysis*, *support vector machines*, and *neural networks*) significantly outperformed all individual association measures on all the reference data sets. The best results were achieved by a simple neural network with five units in the hidden layer. Its MAP=80.87% that was achieved on the *PDT-Dep* data set represents 21.53% relative improvement with respect to the best individual measure on the same set. In the experiments on the *CNC-Surf* data set, the same neural network achieved MAP=86.30%. After adding linguistic features (information about part-of-speech and dependency type) to this model, the MAP score on *PDT-Dep* increased to 84.53% (25.94% relative improvement) and on *CNC-Surf* to 88.22%.

Moreover, we have observed that it is not necessary to combine all the 82 association measures, Even a small subset of about 13 selected measures which performs statistically indistinguishably from the full model (with the neural network with five units in the hidden layer, measured by MAP on *PDT-Dep*), is sufficient. This subset contains measures from the entire spectrum, based on different extraction principles, and with very different individual performance. Although the combination of the 13 measures is not guaranteed to be efficient also on other data sets, the proposed algorithm can be easily used to select the right measures for any specific data set and task (assuming manually annotated data is available).

Performance of lexical association measures in the task of ranking collocation extraction heavily depends on many aspects and must be evaluated on particular data and task. Combining association measures is meaningful and improves precision and recall of the extraction procedure and substantial performance improvements can be achieved with a relatively small number of measures combined in a relatively simple model.

A
MWE 2008 Shared Task Results

This appendix is devoted to our participation in the MWE 2008 "Towards a Shared Task for Multiword Expressions" evaluation campaign focused on ranking MWE candidates which was also described in Pecina (2008a). The system we used for this shared task slightly differed from the one described in this work, namely in the following aspects: we employed only the 55 statistical association measures (no context--based measures were used), the results were crossvalidated in 7-fold crossvalidation and compared by mean average precision (MAP) estimated on the *full* (not limited) interval of recall $\langle 0, 1\rangle$. We employed the same combination methods and again observed significant performance improvement by combining multiple association measures.

A.1 Introduction

Four gold-standard data sets were provided for the MWE 2008 shared task. The goal was to re-rank each list such that the "best" candidates are concentrated at the top of the list[1]. Our experiments were carried out over only three of the data sets – those provided with corpus frequency data by the shared task organizers:

1. German *Adj-Noun* collocation candidates from the *Frankfurter Rundschau* corpus,
2. German *PP-Verb* collocation candidates from the *Frankfurter Rundschau* corpus,
3. Czech general collocation candidates from the *Prague Dependency Treebank*.

In the following sections, for each of these data set, we present the best performing individual association measure and results of methods combining multiple association measures.

A.2 System overview

In our system, described also in our previous work (Pecina and Schlesinger, 2006; Pecina, 2005), each collocation candidate x^i is described by the *feature vector* $\mathbf{x}^i = (x^i_1, \ldots, x^i_{55})^T$ consisting of the 55 association scores from Table 3.4 (in Chapter 3 of this work) computed from the corpus frequency data, and assigned a label $y^i \in \{0, 1\}$ which indicates whether the bigram is considered as true positive ($y = 1$) or not ($y = 0$).

[1] http://multiword.sf.net/mwe2008/

	full set		*used subset*		*1–2*		*1–2–3*	
category	*items*	*%*	*items*	*%*	*items*	*%*	*items*	*%*
1	367	29.31	358	29.53	511	42.16	628	52.82
2	153	12.22	153	12.62				
3	117	9.35	117	9.65	701	57.84		
4	45	3.35	41	3.38			584	48.18
5	537	42.89	517	42.66				
6	33	2.64	26	2.15				
total	1 252	100.0	1 212	100.0	1 212	100.0	1 212	100.0

Table A.1: Category distribution in the full German *Adj-Noun* data (*full set*) and its subset of the items provided with frequency information from the corpus (*used subset*).

	full set		*all*		*infr.30*		*light.v*	
category	*items*	*%*	*items*	*%*	*items*	*%*	*items*	*%*
FVG	549	2.51	543	2.91	282	5.75	455	7.26
figur	600	2.75	589	3.16	280	5.71	286	4.56
TP	1 149	5.26	1 132	6.08	562	11.45	741	11.82
total	21 796	100.00	18 648	100.00	4 907	100.00	6 271	100.00

Table A.2: Category distribution in the full German *PP-Verb* data (*full set*) and its subsets of the candidates provided with frequency information from the corpus (*all*, *in.fr30*, *light.v*).

A part of the data is used to train standard statistical-classification models to predict the labels. These methods are modified so that they do not produce 0–1 classification but rather a score that can be used (similarly as for association measures) for ranking the collocation candidates (Pecina and Schlesinger, 2006). The following statistical-classification methods were used in experiments described in this appendix: *linear logistic regression* (GLM), *linear discriminant analysis* (LDA), *neural networks* with 1 and 5 units in the hidden layer (NNet.1, NNet.5), and *support vector machines* (SVM).

For evaluation, we followed a similar procedure that was described in Chapter 5 of this work. Before the experiments, each of the reference data sets was split into seven stratified folds of the same size, each containing the same ratio of true positives. *Average precision* (AP), corresponding to the area under the *precision-recall curve*, was estimated for each data fold and its mean was used as the main evaluation measure – *mean average precision* (MAP). The methods combining multiple association measures used six data folds for training (fitting the parameters) and one for testing in the 7-fold crossvalidation.

	1–2		*1–2–3*	
Baseline	42.12		51.78	
Best AM	**62.88**	(51)	69.14	(51)
GLM	60.88		70.62	
LDA	**61.30**		**70.77**	
SVM	57.95		64.24	
NNet.1	60.52		70.38	
NNet.5	59.87		70.16	

Table A.3: MAP scores of ranking the German *Adj-Noun* collocation candidates.

A.3 German Adj-Noun collocations

A.3.1 Data description

This data set consist of 1 252 German collocation candidates randomly sampled from the 8 546 different adjective-noun pairs (attributive prenominal adjectives only) occurring at least 20 times in the *Frankfurter Rundschau* corpus (Rundschau, 1994). The collocation candidates were lemmatized with the IMSLex morphology (Lezius et al., 2000), pre-processed with the partial parser YAC (Kermes, 2003) for data extraction, and annotated by professional lexicographers with the following categories (distribution is shown in Table A.1):

1. true lexical collocations, other multiword expressions,
2. customary and frequent combinations, often part of a collocational pattern,
3. common expressions, but no idiomatic properties,
4. unclear/boundary cases,
5. not collocational, free combinations,
6. lemmatization errors corpus-specific combinations.

A.3.2 Experiments and results

Frequency counts were provided for 1 212 collocation candidates from this data set. We performed two sets of experiments on this subset. First, only the categories 1–2 were considered true positives. There was a total of 511 such cases and thus the baseline precision was quite high (42.12%). The highest MAP=62.88% achieved by *Piatersky–Shapiro coefficient (51)* was not outperformed by any of the combination methods.

In the second set of experiments, the true positives comprised categories 1–2–3 (the total of 628 items). The baseline precision was as high as 51.78%. The best association measure was again *Piatersky–Shapiro coefficient (51)* but it was slightly outperformed by most of the combination methods. The best one was based on LDA and achieved MAP=70.77%. Detailed results are presented in Table A.3.

	all		*in.fr30*		*light.v*	
Baseline	2.91		5.75		7.26	
Best AM	18.26	(48)	28.48	(48)	43.97	(14)
GLM	28.40		26.59		41.25	
LDA	28.38		40.44		**45.08**	
SVM	14.15		27.51		32.10	
NNet.1	**30.77**		42.42		44.98	
NNet.5	30.49		**43.40**		44.23	

Table A.4: MAP of ranking the German *PP-Verb* support-verb construction (*FVG*) candidates.

A.4 German PP-Verb collocations

A.4.1 Data description

This data set comprises 21 796 German combinations of a prepositional phrase (PP) and a governing verb extracted from the Frankfurter Rundschau corpus (Rundschau, 1994) and used in a number of experiments (e.g. Krenn, 2000). PPs are represented by the combination of a preposition and a nominal head. Both the nominal head and the verb were lemmatized using the IMSLex morphology (Lezius et al., 2000) and processed by the partial parser YAC (Kermes, 2003). See Evert (2004) for details of the extraction procedure. The data was manually annotated as lexical collocations or non-collocations by Krenn (2000). In addition, a distinction was made between two subtypes of lexical collocations: *support-verb constructions* (*FVG*), and *figurative expressions* (*figur*), detailed statistics for the data are shown in Table A.2.

A.4.2 Experiments and results

Frequency data were provided for the total of 18 649 collocation candidates (*all*). We carried out several series of experiments on this subset. First, we focused on the support-verb constructions (*FVG*) and figurative expressions (*figur*) separately, then we attempted to extract the same items without making the distinction in their type (*TP*). Further, as suggested by the shared task organizers, we restricted ourselves to a subset of 4 908 candidate pairs that occur at least 30 times in the Frankfurter Rundschau corpus (*in.fr30*). In the similar manner, additional experiments were restricted to candidate pairs containing one of 16 typical *light verbs*. This step was motivated by the assumption that filtering based on such condition should significantly improve the performance of association measures. After applying this filter, the resulting set (*light.v*) contains 6 272 collocation candidates.

	all		*in.fr30*		*light.v*	
Baseline	3.16		5.70		4.56	
Best AM	14.98	(48)	21.04	(51)	23.65	(12)
GLM	**19.22**		15.28		10.46	
LDA	18.34		**23.32**		24.88	
SVM	7.95		15.70		13.29	
NNet.1	19.05		22.01		24.30	
NNet.5	18.26		22.73		**25.86**	

Table A.5: MAP scores of ranking the German *PP-Verb* figurative expression (*figur*) candidates.

Support-verb constructions

The baseline precision for ranking only the support-verb constructions in all the data (*all*) is as low as 2.91%, while the best MAP (18.26%) was achieved by *Confidence measure (48)*. Additional substantial improvement was achieved by all combination methods. The best score (30.77%) was obtained by the neural network with 1 unit in the hidden layer (NNet.1). When we focused on the candidates occurring at least 30 times (*infr.30* – baseline precision 5.75%), the best individual association measure was again *Confidence measure (48)* with MAP 28.48%. The best combination method was then neural network with 5 units in the hidden layer (NNet.5): MAP 43.40%. The best performing individual association measure on the light verb data (*light.v*) was *Poisson significance measure (14)* with MAP as high as 43.97% (baseline 7.25%). The performance gain achieved by the best combination method was not, however, so significant (45.08%, LDA). Details are shown in Table A.4.

Figurative expressions

Ranking figurative expressions is more difficult. The best individual association measure on all the data (*all*) is again *Confidence measure (48)* with MAP of only 14.98%, although the baseline precision is a little bit higher then in the case of support-verb constructions (3.16%). The best combination of multiple association measures is obtained by liner logistic regression (GLM) with MAP equal to 19.22%. Results for the candidates occurring at least 30 times (*in.fr30* – baseline precision 5.70%) are higher: the best AM (*Piatersky-Shapiro coefficient (51)*) with MAP 21.04% and LDA with MAP 23.32%. In the case of PP combinations with light verbs (*light.v*), the winning individual AM is *t test (12)* with MAP=23.65%, and the best combination method is NNet.5 with 25.86%. Details are given in Table A.5.

Support-verb constructions and figurative expressions

The last set of experiments performed on the German *PP-Verb* data aimed at ranking both support-verb constructions and figurative expressions without making any dis-

	all		*in.fr30*		*light.v*	
Baseline	6.07		11.45		11.81	
Best AM	31.17	(48)	43.85	(48)	63.59	(14)
GLM	44.66		47.81		65.37	
LDA	41.20		57.77		65.54	
SVM	–		51.91		55.10	
NNet.1	44.71		**60.59**		65.10	
NNet.5	**44.77**		59.59		**66.06**	

Table A.6: MAP scores of ranking the German *PP-Verb* candidates of both support-verb constructions and figurative expressions.

tinction between these two types of collocations. The results are shown in Table A.6 and are not very surprising. The best individual AM on all the candidates (*all*) as well as on the subset of frequent candidates (*in.fr30*) was *Piatersky-Shapiro coefficient (51)* with MAP 31.17% and 43.85%, respectively. *Poisson significance measure (14)* performed best on the candidates containing light verbs (*light.v*) (63.59%). The best combination methods were the neural networks with 1 and 5 units (NNet.1, NNet.5), respectively. The most substantial performance improvement obtained by combining multiple association measures was observed on the set of all candidates (no filtering applied).

A.5 Czech PDT-Dep collocations

A.5.1 Data description

The PDT data contains an annotated set of 12 232 normalized dependency bigrams occurring in the manually annotated Prague Dependency Treebank 2.0 more than five times and having part-of-speech patterns that can possibly form a collocation. Every bigram is assigned to one of the six categories described below by three annotators. Only the bigrams where all annotators agreed on them being collocations (of any type, categories 1–5) are considered true positives. The entire set contains 2 572 such items.

0. non-collocations,
1. stock phrases, frequent unpredictable usages,
2. names of persons, organizations, geographical locations, and other entities,
3. support-verb constructions,
4. technical terms,
5. idiomatic expressions.

Note: This data set is identical to the *PDT-Dep* reference data set described in Section 4.2.1 of this work. However, the evaluation was performed over all seven cross-validation folds (and thus the results are slightly different).

	AM	*AM+POS*
Baseline	21.01	
Best AM	65.63	(39)
GLM	67.21	77.27
LDA	67.23	75.83
SVM	**71.44**	74.38
NNet.1	67.34	77.76
NNet.5	70.31	**79.51**

Table A.7: MAP scores of ranking the Czech *PDT-Dep* data. The second column refers to experiments using combination of association measures and POS pattern information.

A.5.2 Experiments and results

The baseline precision on this data is 21.02%. In our experiments, the best performing individual association measure was *Unigram subtuple measure (39)* with MAP=65.63%. The best method combining all association measures was support vector machine with MAP equal to 71.44%. After introducing a new (categorical) variable indicating POS patterns of the collocation candidates and adding it to the combination methods, the performance increased up to 79.51% (for the best method – NNet.5).

A.6 Conclusion

The overview of the best results achieved by the individual association measures and by the combination methods on all the data sets (and their variants) is shown in Table A.8. With only one exception the combination methods significantly improved the ranking of collocation candidates on all data sets. Our results showed that different measures give different results for different tasks (data). It is not possible to recommend "the best general association measure" for ranking collocation candidates, as the performance of the measures heavily depend on the data/task. Instead, we suggest to use the proposed machine learning approach and let the classification methods do the job and weight each measure appropriately for each specific task and/or data. It seems that a neural network is probably the most suitable learner for this task, but the other combination methods also performed well.

data set	*variant*	*baseline*	*best AM*	*best CM*	*+%*
GR *Adj-Noun*	*1-2*	42.40	62.88	61.30	-2.51
	1-2-3	51.74	69.14	70.77	2.36
GR *PP-Verb* FVG	*all*	2.89	18.26	30.77	68.51
	in.fr30	5.71	28.48	43.40	52.39
	light.v	7.26	43.97	45.08	2.52
GR *PP-Verb* Figur	*all*	3.15	14.98	19.22	28.30
	in.fr30	5.71	21.04	23.32	10.84
	light.v	4.47	23.65	25.86	9.34
GR *PP-Verb* all	*all*	6.05	31.17	44.77	43.63
	in.fr30	11.45	43.85	60.59	38.18
	light.v	11.73	63.59	66.06	3.88
CZ *PDT-Dep*	*AM*	21.01	65.63	71.44	8.85
	AM+POS	21.01	65.63	79.51	21.15

Table A.8: Summary of the results obtained on all the data sets and their variants. The last two columns refer to the scores of the best methods combining multiple association measures (*best CM*) and the corresponding relative improvements (+%) compared to the best individual association measure (*best AM*). The last row refers to the experiment using a combination of association measures and information about part-of-speech (*POS*) patterns.

B

Complete Evaluation Results

This appendix contains an overview of the results of all the evaluation experiments performed in this work. For each data set, we present: 1) a barplot of the MAP scores of all individual association measures (sorted in descending order), 2) a visualization of the results of significance tests of difference between all individual association measures (by the paired Student's t-test and paired Wilcoxon signed-ranked test), the gray points correspond to p-values greater than $\alpha = 0.05$ and indicate pairs of measures with statistically indistinguishable performance, and 3) a table of the MAP scores obtained by combination of all association measures in different models and their relative performance improvement compared to the best individual measures.

B.1 *PDT-Dep*

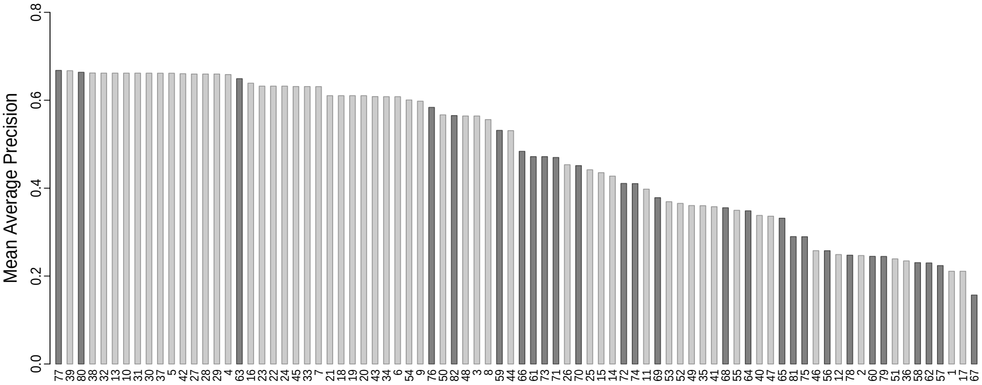

Figure B.1: MAP scores of all individual association measures (in descending order).

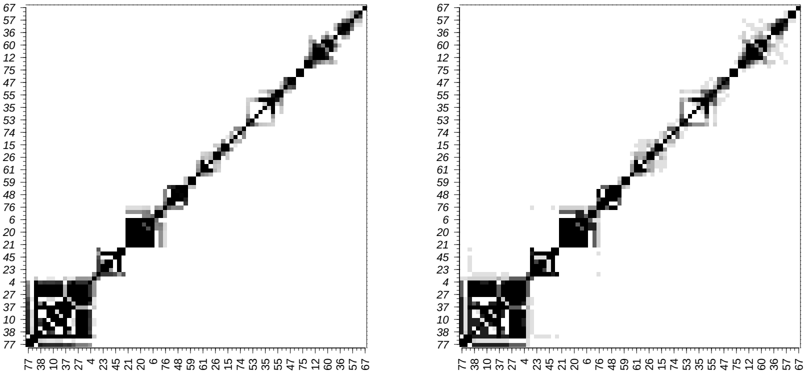

Figure B.2: Significance tests of difference between all individual association measures (the paired t-test on the left and paired signed-rank Wilcoxon test on the right, α=0.05).

	AM	+%	*AM+POS*	+%	*AM+POS+DEP*	+%
Baseline	21.01	–	21.01	–	21.01	–
Best AM (77)	66.79	0.00	66.79	0.00	66.79	0.00
GLM	77.36	15.82	79.77	19.43	82.07	22.88
LDA	75.16	12.54	78.00	16.79	82.07	22.88
SVM	73.03	9.35	77.55	16.10	79.01	18.29
NNet.1	74.36	11.33	78.28	17.20	82.01	22.79
NNet.5	**80.87**	**21.08**	**82.79**	**23.96**	**84.53**	**25.56**

Table B.1: MAP scores of combination of all association measures and their relative performance improvement (+%) compared to the best individual measure.

B.2 *PDT-Surf*

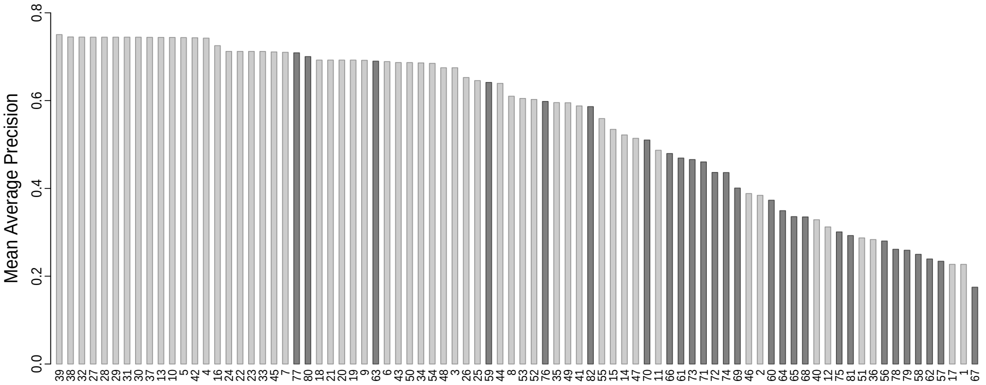

Figure B.3: Sorted MAP scores of all individual association measures.

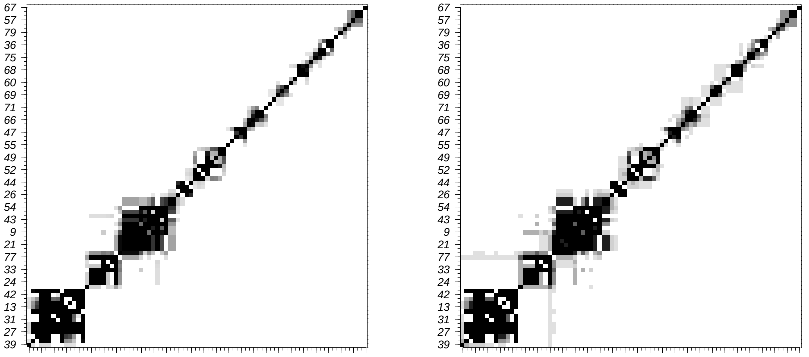

Figure B.4: Significance tests of difference between all individual association measures (the paired t-test on the left and paired signed-rank Wilcoxon test on the right, α=0.05).

	AM	+%	*AM+POS*	+%
Baseline	22.88	–	22.88	–
Best AM (39)	75.03	0.00	75.03	0.00
GLM	79.67	6.18	78.91	5.17
LDA	79.47	5.92	82.56	10.03
SVM	77.58	3.40	81.09	8.08
NNet.1	79.1	5.43	82.44	9.87
NNet.5	**84.35**	**12.43**	**86.40**	**15.15**

Table B.2: MAP scores of combination of all association measures and their relative performance improvement (+%) compared to the best individual measure.

B.3 *CNC-Surf*

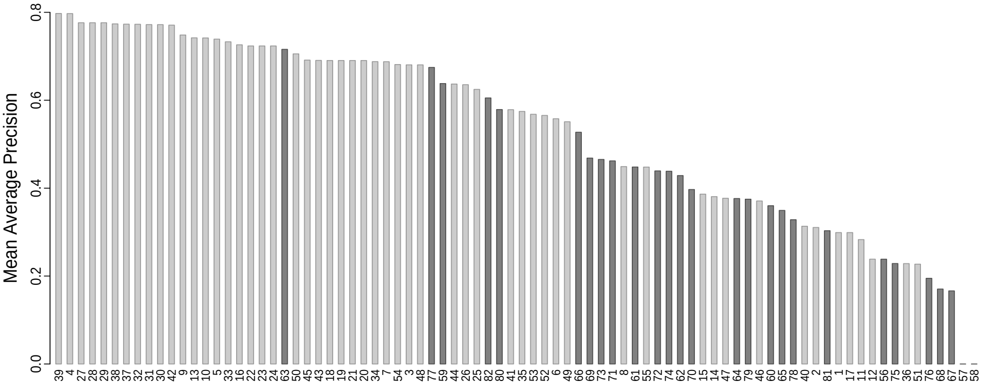

Figure B.5: Sorted MAP scores of all individual association measures.

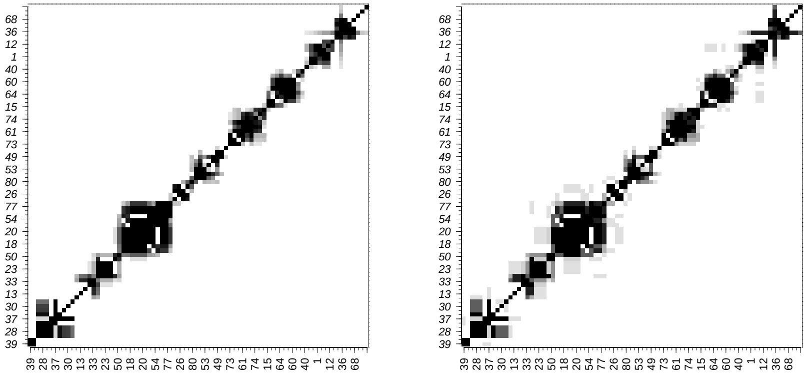

Figure B.6: Significance tests of difference between all individual association measures (the paired t-test on the left and paired signed-rank Wilcoxon test on the right, α=0.05).

	AM	+%	*AM+POS*	+%
Baseline	22.66	–	22.66	–
Best AM (39)	79.74	0.00	79.74	0.00
GLM	75.21	-5.69	85.13	6.76
LDA	82.75	3.77	84.54	6.01
SVM	80.51	0.97	81.41	2.10
NNet.1	83.07	4.17	85.26	6.92
NNet.5	**86.30**	**8.23**	**88.22**	**10.64**

Table B.3: MAP scores of combination of all association measures and their relative performance improvement (+%) compared to the best individual measure.

B.4 *PAR-Dist*

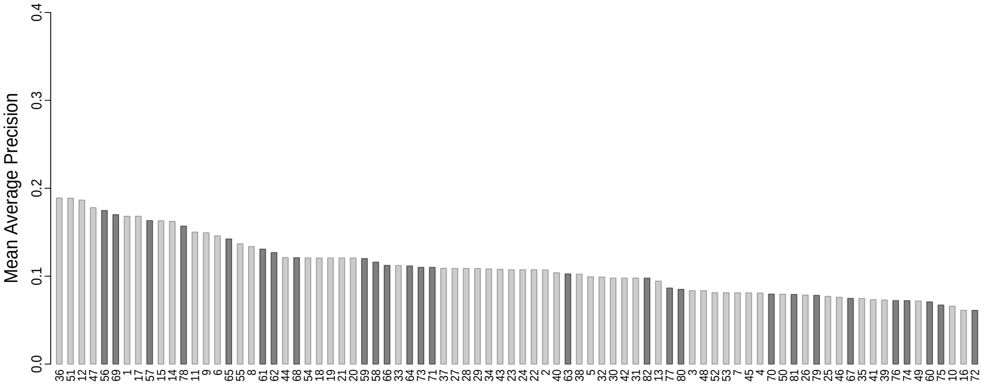

Figure B.7: Sorted MAP scores of all individual association measures.

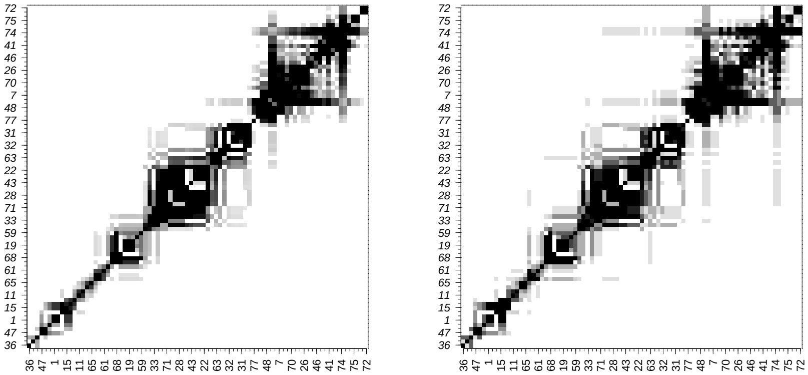

Figure B.8: Significance tests of difference between all individual association measures (the paired t-test on the left and paired signed-rank Wilcoxon test on the right, α=0.05).

	AM	+%
Baseline	7.59	–
Best AM (36)	18.88	0.00
GLM	34.24	81.35
LDA	32.79	73.68
SVM	31.94	69.17
NNet.1	34.52	82.82
NNet.5	**35.78**	**89.50**

Table B.4: MAP scores of combination of all association measures and their relative performance improvement (+%) compared to the best individual measure.

*PAR-Dist (*f > 5*)*

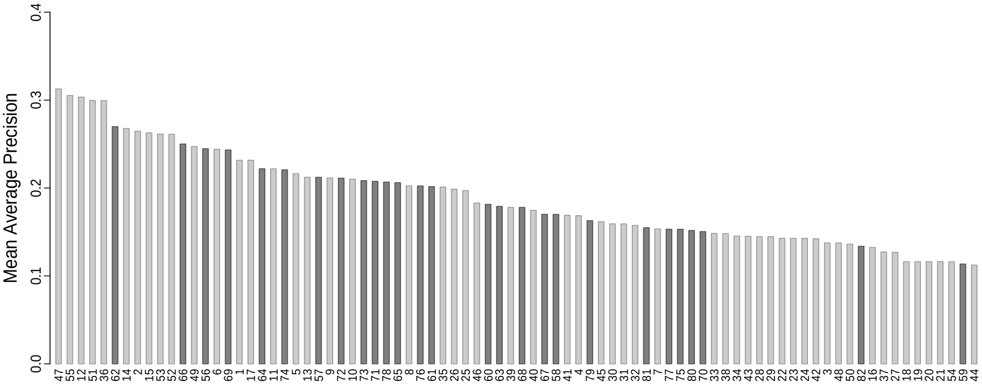

Figure B.9: Sorted MAP scores of all individual association measures.

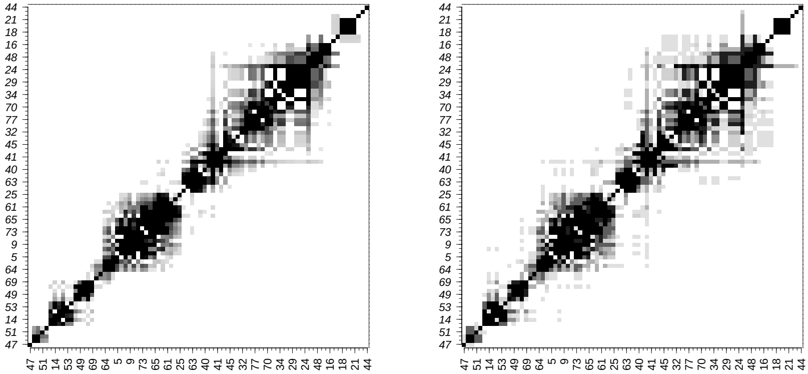

Figure B.10: Significance tests of difference between all individual association measures (the paired t-test on the left and paired signed-rank Wilcoxon test on the right, α=0.05).

	AM	+%
Baseline	13.79	–
Best AM (47)	31.27	0.00
GLM	47.87	53.09
LDA	48.11	53.85
SVM	47.12	50.68
NNet.1	48.28	54.39
NNet.5	**52.15**	**66.76**

Table B.5: MAP scores of combination of all association measures and their relative performance improvement (+%) compared to the best individual measure.

Summary

This work is devoted to an empirical study of lexical association measures and their application to two-word collocation extraction. We have compiled a comprehensive inventory of 82 lexical association measures and present their empirical evaluation on four reference data sets: Czech dependency bigrams from the manually annotated *Prague Dependency Treebank*, surface bigrams from the same source, instances of the latter from the substantially larger *Czech National Corpus* provided with automatically assigned lemmas and part-of-speech tags, and finally, Swedish distance verb-noun combinations from the automatically part-of-speech tagged *PAROLE* corpus. The collocation candidates in the reference data sets were manually annotated and labeled as collocations or non-collocations by educated linguists. The applied evaluation scheme is based on measuring the quality of ranking collocation candidates according to their chance to form collocations. The methods are compared by *precision-recall curves*, *mean average precision scores*, and appropriate tests of statistical significance. Further, we also study the possibility of combining lexical association measures and present empirical results of several combination methods that significantly improved state of the art in collocation extracting. Finally, we propose a model reduction algorithm that significantly reduces the number of combined measures without any statistically significant difference in performance.

Bibliography

Hiyan Alshawi and David Carter. Training and scaling preference functions for disambiguation. *Computational Linguistics*, 4(20):635–648, 1994.

Carmen Alvarez, Philippe Langlais, and Jian-Yun Nie. Word pairs in language modeling for information retrieval. In *Proceedings of the 7th Conference on Computer-Assisted Information Retrieval (RIAO)*, pages 686–705, Avignon, France, 2004.

Sophia Ananiadou. A methodology for automatic term recognition. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pages 1034–1038, Kyoto, Japan, 1994.

Ofer Arazy and Carson Woo. Enhancing information retrieval through statistical natural language processing: A study of collocation indexing. *Management Information Systems Quarterly*, 3(31), 2007.

Debra S. Baddorf and Martha W. Evens. Finding phrases rather than discovering collocations: Searching corpora for dictionary phrases. In *Proceedings of the 9th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS)*, Dayton, USA, 1998.

Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press/Addison-Wesley, 1999.

Jens Bahns. Lexical collocations: a contrastive view. *ELT Journal*, 1(47):56–63, 1993.

Timothy Baldwin. Compositionality and multiword expressions: Six of one, half a dozen of the other?, 2006. Invited talk, given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties.

Timothy Baldwin and Aline Villavicencio. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, Taipei, Taiwan, 2002.

Lisa Ballesteros and Bruce W. Croft. Dictionary-based methods for crosslingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pages 791–801, 1996.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. A statistical approach to the semantics of verb-particles. In Anna Korhonen, Diana McCarthy, Francis Bond, and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan, 2003.

Marco Baroni, Johannes Matiasek, and Harald Trost. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the ACL Workshop on Morphological and Phonological Learning*, pages 48–57, 2002.

Cesare Baroni-Urbani and Mauro W. Buser. Similarity of binary data. *Systematic Zoology*, 25: 251–259, 1976.

Sabine Bartsch. *Structural und Functional Properties of Collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence.* Gunter Narr Verlag Tübingen, 2004.

Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. Semi-automatic extraction of linguistic information for syntactic disambiguation. *Applied Artificial Intelligence*, 7:339–364, 1993.

Laurie Bauer. *English Word-Formation*. Cambridge University Press, 1983.

Doug Beefermam, Adam Berger, and John Lafferty. A model of lexical attraction and repulsion. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 373–380, 1997.

Morton Benson. Collocations and idioms. In Roberr Ilson, editor, *Dictionaries, Lexicography and Language Learning*, pages 61–68. Pergamon, Oxford, 1985.

Morton Benson, Evelyn Benson, and Robert Ilson. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam, Netherlands, 1986.

Godelieve L.M. Berry-Rogghe. The computation of collocations and their relevance in lexical studies. In *The Computer and Literal Studies*, pages 103–112, Edinburgh, New York, USA, 1973. University Press.

Chris Biemann, Stefan Bordag, and Uwe Quasthoff. Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 967–970, Lisbon, Portugal, 2004.

Don Blaheta and Mark Johnson. Unsupervised learning of multi-word verbs. In *Proceedings of the ACL Workshop on Collocations*, pages 54–60, 2001.

Endre Boros, Peter L. Hammer, Toshihide Ibaraki, and Alexander Kogan. Logical analysis of numerical data. *Mathematical Programming*, 79(1-3):163–190, 1997.

Josias Braun-Blanquet. *Plant Sociology: The Study of Plant Communities. Authorized English translation of Pflanzensoziologie.* New York: McGraw-Hill, 1932.

Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pages 33–40, Athens, Greece, 2000. ACM.

Ronald Carter. *Vocabulary: Applied linguistic perspectives*. Routledge, 1987.

František Čermák. Syntagmatika slovníku: typy lexikálních kombinací. In Zdeňka Hladká and Petr Karlík, editors, *Čeština - univerzália a specifika 3*, pages 223–232. Masarykova Univerzita, Brno, Czech Republic, 2001.

František Čermák. Kolokace v lingvistice. In František Čermák and Michal Šulc, editors, *Kolokace*. Nakladatelství Lidové noviny, 2006.

František Čermák and Jan Holub. *Syntagmatika a paradigmatika českého slova: Valence a kolokabilita*. Státní pedagogické nakladatelství, Praha, Czech Republic, 1982.

František Čermák and Michal Šulc, editors. *Kolokace*. Nakladatelství Lidové noviny, 2006.

František Čermák et al. *Slovník české frazeologie a idiomatiky*. Leda, Praha, Czech Republic, 2004.

Noam Chomsky. *Syntactic Structures*. The Hague/Paris: Mouton, 1957.

Yaacov Choueka. Looking for needles in a haystack or: Locating interesting expressions in large textual databases. In *Proceedings of the International Conference on User-Oriented Content-Based Text and Image Handling*, pages 609–623, Cambridge, Massachusetts, USA, 1988.

Yaacov Choueka, S.T. Klein, and E. Neuwitz. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, 4(1):34–38, 1983.

Kenneth Church and William Gale. Concordances for parallel text. In *Proceedings of the 7th Annual Conference of the UW Center for the New OED and Text Research*, Oxford, UK, 1991.

Kenneth Church and Patrick Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, pages 22–29, 1990.

Kenneth Church and Robert L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24, 1993.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Parsing, word associations and typical predicate-argument relations. In M. Tomita, editor, *Current Issues in Parsing Technology*. Kluwer Academic, Dordrecht, Netherlands, 1991.

Silvie Cinková and Veronika Kolářová. Nouns as components of support verb constructions in the Prague Dependency Treebank. In *Korpusy a korpusová lingvistika v zahraničí a na Slovensku*, 2004.

Silvie Cinková and Jan Pomikálek. Lempas: A make-do lemmatizer for the Swedish PAROLE corpus. *Prague Bulletin of Mathematical Linguistics*, 86, 2006.

Silvie Cinková, Petr Podveský, Pavel Pecina, and Pavel Schlesinger. Semi-automatic building of Swedish collocation lexicon. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1890–1893, Genova, Italy, 2006.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 1960.

Michael Collins. Discriminative training methods for Hidden Markov Models: Theory and experiments with Perceptron algorithms. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, USA, 2002.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, USA, 1991.

David A. Cruse. *Lexical Semantics*. Cambridge University Press, Cambridge, UK, 1986.

Ido Dagan and Kenneth Church. Termight: Identifying and translation technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP)*, pages 34–40, Stuttgart, Germany, 1994.

Ido Dagan, Lillian Lee, and Fernando Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1), 1999.

Robert Dale, Hermann Moisl, and Harold Somers, editors. *A Handbook of Natural Language Processing*. Marcel Dekker, 2000.

Jesse Davis and Mark Goadrich. The relationship between precision-recall curves and the ROC curve. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA, 2006.

Gaël Dias, Sylvie Guilloré, Jean-Claude Bassano, and José Gabriel Pereira Lopes. Combining linguistics with statistics for multiword term extraction: A fruitful association? In *Proceedings of Recherche d'Informations Assistee par Ordinateur (RIAO)*, 2000.

Harold E. Driver and Alfred Louis Kroeber. Quantitative expression of cultural relationship. *The University of California Publications in American Archaeology and Ethnology*, 31:211–256, 1932.

Ted E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

Philip Edmonds. Choosing the word most typical in context using a lexical cooccurrence network. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 507–509, Madrid, Spain, 1997.

David A. Evans and Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 17–24, Santa Cruz, California, USA, 1996.

Stefan Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, University of Stuttgart, 2004.

Stefan Evert and Hannah Kermes. Experiments on candidate data for collocation extraction. In *Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics (EACL)*, pages 83–86, 2003.

Stefan Evert and Brigitte Krenn. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 188–195, 2001.

Joel L. Fagan. Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods. Technical report, Cornell University, Ithaca, New York, USA, 1987.

Joel L. Fagan. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40:115–32, 1989.

Tom Fawcett. ROC graphs: Notes and practical considerations for data mining researchers. Technical report, HPL 2003–4. HP Laboratories, Palo Alto, California, USA, 2003.

Christiane Fellbaum, editor. *WordNet, An Electronic Lexical Database*. Bradford Books, 1998.

Olivier Ferret. Using collocations for topic segmentation and link detection. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, 2002.

John R. Firth. Modes of meanings. In *Papers in Linguistics 1934–1951*, pages 190–215. Oxford University Press, 1951.

John R. Firth. A synopsis of linguistic theory, 1930–55. In *Studies in linguistic analysis, Special volume of the Philological Society*, pages 1–32. Philogical Society, Oxford, UK, 1957.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

Thierry Fontenelle. Towards the construction of a collocational database for translation students. *Meta*, 1(39):47–56, 1994a.

Thierry Fontenelle. What on earth are collocations? *English Today*, 4(10):42–48, 1994b.

William B. Frakes and Ricardo A. Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms*, chapter Stemming algorithms. Prentice-Hall, Englewood Cliffs, NJ, 1992.

Pascale Fung and Kathleen R. McKeown. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, 1997.

Pascale Fung, Min yen Kan, and Yurie Horita. Extracting Japanese domain and technical terms is relatively easy. In *Proceedings of the 2nd International Conference on New Methods in Natural Language Processing*, pages 148–159, 1996.

Vincent E. Giuliano. The interpretation of word associations. In M. E. Stevens et al., editor, *Statistical association methods for mechanized documentation*, pages 25–32, 1964.

Vincent E. Giuliano. Postscript: A personal reaction to reading the conference manuscripts. In Mary E. Stevens, Vincent E. Giuliano, and Laurence B. Heilprin, editors, *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation*, volume 269 of *National Bureau of Standards Miscellaneous Publication*, pages 259–260, Washington, DC, USA, 1965.

Gregory Grefenstette and Simone Teufel. A corpus-based method for automatic identification of support verbs for nominalisations. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Dublin, Ireland, 1995.

Michelle L. Gregory, William D. Raymond, Alan Bell, Eric Fosler-Lussier, and Daniel Jurafsky. The effects of collocational strength and contextual predictability in lexical production. In *Chicago Linguistics Society (CLS)*, pages 151–166, University of Chicago, USA, 1999.

Jan Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, volume 1. Charles University Press, Prague, Czech Republic, 2004.

Jan Hajič, Jarmila Panevová, Eva Burňová, Zdeňka Urešová, and Alla Bémová. A manual for analytic layer tagging of the prague dependency treebank. Technical Report TR–1997–03, ÚFAL MFF UK, Prague, Czech Republic, 1997.

Michael A.K. Halliday. Lexis as a linguistic level. In C. Bazell, J. Catford, M. Halliday, and R. Robins, editors, *In Memory of J.R. Firth*, pages 148–162. Longman, London, UK, 1966.

Michael A.K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, UK, 1967.

Ute Hamann. Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Betrag zum System der Monokotyledonen. *Willdenowia*, 2:639–768, 1961.

Masahiko Haruno, Satoru Ikehara, and Takefumi Yamazaki. Learning bilingual collocations by word-level sorting. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, Copenhagen, Denmark, 1996.

Ruqaiya Hasan. Coherence and cohesive harmony. In J. Flood, editor, *Understanding Reading Comprehension*, pages 181–219. Newark, Del: International Reading Association, 1984.

Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5, 2004.

Ulrich Heid. Towards a corpus-based dictionary of german noun-verb collocations. In *Proceedings of the EURALEX International Congress*, volume 1, pages 301–312, Liège, Belgium, 1998.

David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, Pennsylvania, 1993.

David Hull and Gregory Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57, Zurich, Switzerland, 1996.

ICNC. Czech National Corpus – SYN2000, 2000. Institute of the Czech National Corpus Faculty of Arts, Charles University, Praha.

ICNC. Czech National Corpus – SYN2005, 2005. Institute of the Czech National Corpus Faculty of Arts, Charles University, Praha.

Diana Inkpen and Graeme Hirst. Acquiring collocations for lexical choice between near synonyms. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, pages 67–76, Philadelphia, Pennsylvania, 2002.

Paul Jaccard. The distribution of the flora in the alpine zone. *The New Phytologist*, 11:37–50, 1912.

Maojin Jiang, Eric Jensen, Steve Beitzel, and Shlomo Argamon. Effective use of phrases in language modeling to improve information retrieval. In *Symposium on AI & Math Special Session on Intelligent Text Processing*, Florida, USA, 2004.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, 2nd ed. Springer, New York, 2002.

John S. Justeson and Slava M. Katz. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 1:1–19, 1991.

John S. Justeson and Slava M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Sciences, 1990.

Hannah Kermes. *Off-line (and On-line) Text Analysis for Computational Lexicography*. PhD thesis, IMS, University of Stuttgart, 2003.

Christopher S. G. Khoo, Sung Hyon Myaeng, and Robert N. Oddy. Using cause-effect relations in text to improve information retrieval precision. *Information Processing and Management*, 37(1):119–145, 2001.

Adam Kilgarriff. *Polysemy*. PhD thesis, University of Sussex, UK, 1992.

Adam Kilgarriff and David Tugwell. WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Proceedings of the ACL 2001 Collocations Workshop*, pages 32–38, Toulouse, France, 2001.

Tibor Kiss and Jan Strunk. Scaled log likelihood ratios for the detection of abbreviations in text corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1228–1232, Taipeh, Taiwan, 2002a.

Tibor Kiss and Jan Strunk. Viewing sentence boundary detection as collocation identification. In S. Busemann, editor, *Tagungsband der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, pages 75–82, Saarbrücken, Germany, 2002b.

Kenji Kita and Hiroaki Ogata. Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer. *Computer Assisted Language Learning: An International Journal*, 10(3):229–238, 1997.

Kenji Kita, Yasuhiro Kato, Takashi Omoto, and Yoneo Yano. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing*, 1(1):21–33, 1994.

Goran Kjellmer. Aspects of english collocations. In W. Meijs, editor, *Corpus Linguistics and Beyond. Proceedings of the 7th International Conference on English Language Research on Computerised Corpora*, pages 133–40, Amsterdam, Netherlands, 1987.

Goran Kjellmer. *A mint of phrases*. Longman, Harlow, UK, 1991.

Goran Kjellmer. *A Dictionary of English Collocations*. Clarendon Press, 1994.

Aleš Klégr, Petra Key, and Norah Hronková. *Česko-anglický slovník spojení: podstatné jméno a sloveso*. Karolinum, Praha, Czech Republic, 2005.

Ron Kohavi and Foster Provost. Glossary of terms. *Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, 30(2/3):271–274, 1998.

Brigitte Krenn. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. PhD thesis, Saarland University, 2000.

Brigitte Krenn, Stefan Evert, and Heike Zinsmeister. Determining intercoder agreement for a collocation identification task. In *Proceedings of KONVENS'04*, pages 89–96, Vienna, Austria, 2004.

Stanisław Kulczynski. Die Pflanzenassociationen der Pienenen. *Bulletin International de L'Acad'emie Polonaise des Sciences et des Letters, Classe des Sciences Mathematiques et Naturelles, Serie B, Supplement II*, 2:57–203, 1927.

Julian Kupiec, Jan O. Pedersen, and Francine Chen. A trainable document summarizer. In *Research and Development in Information Retrieval*, pages 68–73, 1995.

Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. *Artificial Inteligence*, pages 65–72, 2001.

Michael Lesk. Word-word associations in document retrieval systems. *American Documentation*, 1(20):27–38, 1969.

Wolfgang Lezius, Stefanie Dipper, and Arne Fitschen. IMSLex - representing morphological and syntactical information in a relational database. In *U. Heid, S. Evert, E. Lehmann, and C. Rohrer (eds.): Proceedings of the 9th EURALEX International Congress*, Stuttgart, Germany, 2000.

Dekang Lin. Using collocation statistics in information extraction. In *Proceedings of the 7th Message Understanding Conference (MUC 7)*, 1998.

Dekang Lin. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic*, pages 317–24, College Park, Maryland, USA, 1999.

David M. Magerman and Mitchell P. Marcus. Parsing a natural language using mutual information statistics. In *Proceedings of the 8th National Conference on Artificial Intelligence*, pages 984–989, Boston, Massachusetts, USA, 1990.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, USA, 1999.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Diana Maynard and Sophia Ananiadou. Identifying contextual information for multi-word term extraction. In *Proceedings of 5th International Congress on Terminology and Knowledge Engineering (TKE)*, pages 212–221, 1999.

Diana McCarthy, Bill Keller, and John Carroll. Detecting a continuum of compositionality in phrasal verbs. In Anna Korhonen, Diana McCarthy, Francis Bond, and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, 2003.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Human Language Technologies and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, Canada, 2005.

Kathleen R. McKeown and Dragomir R. Radev. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *A Handbook of Natural Language Processing*. Marcel Dekker, 2000.

Dan I. Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

Ellis L. Michael. Marine ecology and the coefficient of association. *Journal of Animal Ecology*, 8: 54–59, 1920.

Rada Mihalcea and Ted Pedersen. An evaluation exercise for word alignment. In *Proceedings of HLT-NAACL Workshop, Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada, 2003.

Terry F. Mitchell. Linguistic 'goings on': Collocations and other lexical matters arising on the syntactic record. *Archivum Linguisticum*, 2:35–69, 1971.

Elke Mittendorf, Bojidar Mateev, and Peter Schäuble. Using the co-occurrence of words for retrieval weighting. *Information Retrieval*, 3(3):243–251, 2000.

María Begona Villada Moirón. *Data-driven identification of fixed expressions and their modifiability*. PhD thesis, University of Groningen, 2005.

Rosamund Moon. *Fixed Expressions and Idioms in English*. Clarendon Press, Oxford, UK, 1998.

Robert C. Moore. On log-likelihood-ratios and the significance of rare events. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, 2004.

Robert C. Moore, Wen tau Yih, and Andreas Bode. Improved discriminative bilingual word alignment. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 513–520, Sydney, Australia, 2006.

Václav Novák and Zdeněk Žabokrtský. Feature engineering in maximum spanning tree dependency parser. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue (TSD)*, Pilsen, Czech Republic, 2007.

Kumiko Ohmori and Masanobu Higashida. Extracting bilingual collocations from non-aligned parallel corpora. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 88–97, University College, Chester, England, 1999.

David S. Palermo and James J. Jenkins. *Word Association norms*. University of Minnesota Press, Mineapolis, Minnesota, USA, 1964.

Frank R. Palmer, editor. *Selected Papers of J.R. Firth 1952–1959*. Bloomington: Indiana University Press, 1968.

Harold E. Palmer. *A Grammar of English Words*. Longman, London, UK, 1938.

Harold E. Palmer and Albert S. Hornby. *Thousand-Word English*. George Harrap, London, UK, 1937.

Patrick Pantel and Dekang Lin. Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada, 2002.

Darren Pearce. A comparative evaluation of collocation extraction techniques. In *Proceedings of the 3rd International Conference on language Resources and Evaluation (LREC)*, Las Palmas, Spain, 2002.

Pavel Pecina. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, Ann Arbor, Michigan, USA, 2005.

Pavel Pecina. Machine learning approach to mutliword expression extraction. In *Proceedings of the 6th International Conference on Language Resources and Evaluation Workshop: Towards a Shared Task for Multiword Expressions*, Marrakech, Morocco, 2008a.

Pavel Pecina. Reference data for Czech collocation extraction. In *Proceedings of the 6th International Conference on Language Resources and Evaluation Workshop: Towards a Shared Task for Multiword Expressions*, Marrakech, Morocco, 2008b.

Pavel Pecina and Pavel Schlesinger. Combining association measures for collocation extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Sydney, Australia, 2006.

Pavel Pecina, Petra Hoffmannová, Gareth J.F. Jones, Jianqiang Wang, and Douglas W. Oard. Overview of the CLEF 2007 Cross-Language Speech Retrieval Track. *Evaluation of Multilingual and Multi-modal Information Retrieval (CLEF 2007), Revised Selected Papers. Lecture Notes in Computer Science*, 2008.

Ted Pedersen. Fishing for exactness. In *Proceedings of the South Central SAS User's Group Conference*, pages 188–200, Austin, Texas, USA, 1996.

Ted Pedersen. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Pittsburgh, Pennsylvania, USA, 2001.

Luboš Prchal. *Selected aspects of functional estimation and testing: Functional response in regression models and statistical analysis of ROC curves with applications*. PhD thesis, Charles Univeristy of Prague and Paul Sabatier Univeristy - Toulouse III, 2008.

Uwe Quasthoff and Christian Wolff. The Poisson collocation measure and its applications. In *Proceedings of 2nd International Workshop on Computational Approaches to Collocations*, Wien, Austria, 2002.

Reinhard Rapp. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, College Park, Maryland, USA, 1999.

Reinhard Rapp. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipeh, Taiwan, 2002.

Reinhard Rapp. Utilizing the one-sense-per-discourse constraint for fully unsupervised word sense induction and disambiguation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 951–954, Lisbon, Portugal, 2004.

Philip Resnik. Selectional preferences and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*, Washington, DC, USA, 1997.

Robert Robins. *A Short History of Linguistics*. Longman, London, UK, 1967.

David J. Rogers and Taffee T. Tanimoto. A computer program for classifying plants. *Science*, 132:1115–1118, 1960.

Ian C. Ross and John W. Tukey. Introduction to these volumes. In *Index to Statistics and Probability*, Los Altos, California, USA, 1975. The RandD Press.

Frankfurter Rundschau, 1994. The FR corpus is part of the ECI Multilingual Corpus I distributed by ELSNET.

P. F. Russel and T. R. Rao. On habitat and association of species of anopheline larvae in south-eastern madras. *Journal of Malaria Institute India*, 3:153–178, 1940.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin/Heidelberg, 2002.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Patrick Schone and Daniel Jurafsky. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 100–108, 2001.

Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 476–481, 1997.

George Gaylord Simpson. Mammals and the nature of continents. *American Journal of Science*, 241:1–31, 1943.

John Sinclair. Beginning the study of lexis. In C. Bazell, J. Catford, M. Halliday, and R. Robins, editors, *In Memory of J.R. Firth*, pages 410–430. Longman, London, UK, 1966.

John Sinclair. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford, UK, 1991.

Frank A. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19: 143–177, 1993.

Frank A. Smadja and Kathleen R. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistic (ACL)*, pages 252–259, 1990.

Frank A. Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

Robert R. Sokal and Charles D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.

Robert R. Sokal and Peter H. Sneath. *Principles of Numerical Taxonomy*. W. H. Freeman and Company, San Francisco, USA, 1963.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, Praha, Czech Republic, 2007.

Mary E. Stevens, Vincent E. Giuliano, and Laurence B. Heilprin, editors. *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation*, volume 269. National Bureau of Standards Miscellaneous Publication, Washington, DC, USA, 1965.

Matthew Stone and Christine Doran. Paying heed to collocations. In *Proceedings of the International Language Generation Workshop (INLG)*, pages 91–100, Herstmonceux Castle, Sussex, UK, 1996.

Raz Tamir and Reinhard Rapp. Mining the web to discover the meanings of an ambiguous word. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 645–648, Melbourne, Florida, USA, 2003.

Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the Eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

Takaaki Tanaka and Yoshihiro Matsuo. Extraction of translation equivalents from non-parallel corpora. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 109–119, 1999.

Pasi Tapanainen, Jussi Piitulainen, and Timo Jarvinen. Idiomatic object usage and support verbs. In *Proceedings of the 36th Annual Meeting of the Association for Compositional Linguistic and 17th International Conference on Computational Linguistics (COLING/ACL)*, pages 1289–1293, Montreal, Quebec, Canada, 1998.

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. A discriminative matching approach to word alignment. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, British Columbia, 2005.

Egidio Terra and Charles L. A. Clarke. Frequency estimates for statistical word similarity measures. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT/NAACL)*, pages 244–251, Edmonton, Alberta, Canada, 2003.

Aristomenis Thanopoulos, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of collocation extraction metrics. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, volume 2, pages 620–625, Las Palmas, Spain, 2002.

Jörg Tiedemann. Automated lexicon extraction from aligned bilingual corpora. Master's thesis, Otto-von-Guericke-Universität Magdeburg, 1997.

Keita Tsuji and Kyo Kageura. Extracting morpheme pairs from bilingual terminological corpora. *Terminology*, 7(1):101–114, 2001.

Rodham E. Tulloss. *Assessment of Similarity Indices for Undesirable Properties and New Tripartite Similarity Index Based on Cost Functions*. Parkway Publishers, Boone, North Carolina, USA, 1997.

Tem van der Wouden. *Negative contexts: collocations, polarity and multiple negation*. Routledge, London/New York, 1997.

Cornelis J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 1979.

Olga Vechtomova. *Approaches to using word collocation in Information Retrieval*. PhD thesis, City University, London, UK, 2001.

William N. Venables and B.D. Ripley. *Modern Applied Statistics with S. 4th ed.* Springer Verlag, New York, USA, 2002.

Jan Votrubec. Morphological tagging based on averaged Perceptron. In *Proceedings of Contributed Papers (WDS)*, Prague, Czech Republic, 2006. MFF UK.

Michael Wallace. What is an idiom? An applied linguistic approach. In R. Hartmann, editor, *Dictionaries and Their Users: Papers from the 1978 B. A. A. L. Seminar on Lexicography*, pages 63–70. University of Exeter, Exeter, UK, 1979.

Matthijs J. Warrens. *Similarity coefficients for binary data: properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients*. PhD thesis, Leiden University, 2008.

Marc Weeber, Rein Vos, and Harald R. Baayen. Extracting the lowest-frequency words: Pitfalls and possibilities. *Computational Linguistics*, 3(26):301–317, 2000.

Janyce M. Wiebe and Kenneth J. McKeever. Collocational properties in probabilistic classifiers for discourse categorization, 1998.

Hua Wu and Ming Zhou. Synonymous collocation extraction using translation information. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, pages 120–127, Sapporo, Japan, 2003.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, 1995.

Daniel Zeman, Jiří Hana, Hana Hanová, Jan Hajič, Emil Jeřábek, and Barbora Vidová-Hladká. A manual for morphological annotation, 2nd edition. UFAL technical report. Technical Report TR–2005–27, ÚFAL MFF UK, Prague, Czech Republic, 2005.

Chengxiang Zhai. Exploiting context to identify lexical atoms: A statistical view of linguistic context. In *Proceedings of the International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT)*, pages 119–129, 1997.

Georg Kingsley Zipf. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Cambridge, Massachusetts, USA, 1949.

Index