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.

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.

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:

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

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**	$p(y\|x)$	(Gregory et al., 1999)
3.	**Reverse cond. probability**	$p(x\|y)$	(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**	$\max[p(x*)(p(y\|x)^2+p(\bar{y}\|x)^2)-p(*y)^2$ $+p(\bar{x}*)(p(y\|\bar{x})^2+p(\bar{y}\|\bar{x})^2)-p(*\bar{y})^2,$ $p(*y)(p(x\|y)^2+p(\bar{x}\|y)^2)-p(x*)^2$ $+p(*\bar{y})(p(x\|\bar{y})^2+p(\bar{x}\|\bar{y})^2)-p(\bar{x}*)^2]$	(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.

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

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**	$-\sum_z p(z\|C_{xy}) \log p(z\|C_{xy})$	(Krenn, 2000)
57. **Left context entropy**	$-\sum_z p(z\|C^l_{xy}) \log p(z\|C^l_{xy})$	(Shimohata et al., 1997)
58. **Right context entropy**	$-\sum_z p(z\|C^r_{xy}) \log p(z\|C^r_{xy})$	(Shimohata et al., 1997)
59. **Left context divergence**	$p(x*) \log p(x*) - \sum_z p(z\|C^l_{xy}) \log p(z\|C^l_{xy})$	
60. **Right context divergence**	$p(*y) \log p(*y) - \sum_z p(z\|C^r_{xy}) \log p(z\|C^r_{xy})$	
61. **Cross entropy**	$-\sum_z p(z\|C_x) \log p(z\|C_y)$	(Cover and Thomas, 1991)
62. **Reverse cross entropy**	$-\sum_z p(z\|C_y) \log p(z\|C_x)$	(Cover and Thomas, 1991)
63. **Intersection measure**	$\frac{2\|C_x \cap C_y\|}{\|C_x\|+\|C_y\|}$	(Lin, 1998)
64. **Euclidean norm**	$\sqrt{\sum_z (p(z\|C_x) - p(z\|C_y))^2}$	(Lee, 2001)
65. **Cosine norm**	$\frac{\sum_z p(z\|C_x) p(z\|C_y)}{\sum_z p(z\|C_x)^2 \cdot \sum_z p(z\|C_y)^2}$	(Lee, 2001)
66. ***L1* norm**	$\sum_z \|p(z\|C_x) - p(z\|C_y)\|$	(Dagan et al., 1999)
67. **Confusion probability**	$\sum_z \frac{p(x\|C_z) p(y\|C_z) p(z)}{p(x*)}$	(Dagan et al., 1999)
68. **Reverse confusion prob.**	$\sum_z \frac{p(y\|C_z) p(x\|C_z) p(z)}{p(*y)}$	
69. **Jensen-Shannon divergence**	$\frac{1}{2}[D(p(Z\|C_x) \\| \frac{1}{2}(p(Z\|C_x) + p(Z\|C_y)))$ $+D(p(Z\|C_y) \\| \frac{1}{2}(p(Z\|C_x) + p(Z\|C_y)))]$	(Dagan et al., 1999)
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**	$\sum_z p(z\|C_x) \log \frac{p(z\|C_x)}{p(z\|C_y)}$	(Dagan et al., 1999)
72. **Reverse KL divergence**	$\sum_z p(z\|C_y) \log \frac{p(z\|C_y)}{p(z\|C_x)}$	
73. **Skew divergence**	$D(p(Z\|C_x) \\| \alpha\, p(Z\|C_y) + (1-\alpha)\, p(Z\|C_x))$	(Lee, 2001)
74. **Reverse skew divergence**	$D(p(Z\|C_y) \\| \alpha\, p(Z\|C_x) + (1-\alpha)\, p(Z\|C_y))$	
75. **Phrase word coocurrence**	$\frac{1}{2}(\frac{f(x\|C_{xy})}{f(xy)} + \frac{f(y\|C_{xy})}{f(xy)})$	(Zhai, 1997)
76. **Word association**	$\frac{1}{2}(\frac{f(x\|C_y)-f(xy)}{f(xy)} + \frac{f(y\|C_x)-f(xy)}{f(xy)})$	(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**	$\omega_{l,C_e} = TF(u_l, C_e)$	
79. **in** TF·IDF **vector space**	$\omega_{l,C_e} = TF(u_l, C_e) \cdot IDF(u_l)$	
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**	$\omega_{l,C_e} = TF(u_l, C_e)$	
82. **in** TF·IDF **vector space**	$\omega_{l,C_e} = TF(u_l, C_e) \cdot IDF(u_l)$	

Table 3.5: Context-based association measures.

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 asociations. 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

A

B

C

D

E

F

G

H

Q

R

S