Multi-Word Tokenization for Natural Language Processing

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.) genehmigte Abhandlung

Vorgelegt von

Lukas Michelbacher

aus Heidenheim a. d. Brenz

Hauptberichter:	Prof. Dr. Hinrich Schütze
Mitberichter:	Prof. Dr. Sebastian Padó

Tag der mündlichen Prüfung:	14. Januar 2013

Institut für maschinelle Sprachverarbeitung
Universität Stuttgart

2013

Contents

List of abbreviations

AM	association measure
BNC	British National Corpus
EAT	Edinburgh Word Association Thesaurus
IR	information retrieval
LP	left-predictive
LSA	latent semantic indexing
MAP	mean average precision
MI	mutual information
MWE	multi-word expression
MWT	multi-word tokenization
MWU	multi-word unit
NC	non-compositional
NER	named entity recognition
NLP	natural language processing
NP	noun phrase
PMI	pointwise mutual information
POS	part of speech
PP	prepositional phrase
REC	recall
RP	right-predictive
SHR	semantic head recognition
SQN	square-root normalization
SVD	singular value decomposition
SWT	single-word tokenization
USF	University of South Florida

Abstract

Sophisticated natural language processing (NLP) applications are entering everyday life in the form of translation services, electronic personal assistants or open-domain question answering systems. The more voice-operated applications like these become commonplace, the more expectations of users are raised to communicate with these services in unrestricted natural language, just as in a normal conversation.

One obstacle that hinders computers to understand unrestricted natural language is that of *collocations*, combinations of multiple words that have idiosyncratic properties, for example, *red tape*, *kick the bucket* or *there's no use crying over spilled milk*. Automatic processing of collocations is nontrivial because these properties cannot be predicted from the properties of the individual words.

This thesis addresses multi-word units (MWUs), collocations that appear in the form of complex noun phrases. Complex noun phrases are important for NLP because they denote real-world entities and concepts and are often used for specialized vocabulary such as scientific or legal terms.

Virtually every NLP system uses tokenization, the partitioning of textual input into meaningful units, or tokens, as part of preprocessing. Traditionally, tokenization does not deal with MWUs which leads to early errors and error propagation in subsequent NLP tasks, resulting in poorer quality of NLP applications.

The central idea presented in this thesis is the proposition of *multi-word tokenization* (MWT), MWU-aware tokenization as a preprocessing step for NLP systems. The goal of this thesis is to drive research towards NLP applications that understand unrestricted natural language.

Our main contributions cover two aspects of MWT. First, we conducted fundamental research into *asymmetric association*, the phenomenon that lexical association from one component of an MWU to another can be stronger in one direction than in the other. This property has not been investigated deeply in the literature. We position asymmetric association in the broader

context of different types of word association and collected human syntagmatic associations using a novel experiment setup. We measured asymmetric association in human syntagmatic production and showed that it is a phenomenon that is indicative of MWUs. Furthermore, we created corpus-based asymmetric association measures and showed that asymmetry in word combinations can be predicted automatically with high accuracy using these measures.

Second, we present an implementation of MWT where we cast MWU recognition as a classification problem. We built an MWU classifier whose features address properties of MWUs. In particular, we targeted semantic non-compositionality, a phenomenon of unpredictable meaning shifts that occurs in many MWUs. In order to detect meaning shifts, we used features of contextual similarity based on distributional semantics. We found that context features significantly improve MWU classification accuracy but that there are unreliable aspects in the workings of such features. Additionally, we integrated MWT into an information retrieval system and showed that incorporating MWU information improves retrieval performance.

Zusammenfassung

Hoch entwickelte Anwendungen der maschinellen Sprachverarbeitung (NLP, von engl. *natural language processing*) erhalten Einzug in das tägliche Leben in Form automatischer Übersetzungs-, allgemeiner Frage-Antwort-Systeme sowie elektronischer persönlicher Assistenten. Mit der Etablierung sprachgesteuerter Anwendungen steigen die Erwartungen der Benutzer, diese Anwendungen mit unbeschränkter natürlicher Sprache zu bedienen, sich also ganz normal mit ihnen zu unterhalten.

Ein Hindernis, das es Computern erschwert, uneingeschränkte natürliche Sprache zu verstehen, sind Kollokationen, Kombinationen mehrerer Wörter mit besonderen Eigenschaften, wie zum Beispiel *toller Hecht*, *den Löffel abgeben* oder *wo gehobelt wird, da fallen Späne*. Die Automatische Verarbeitung von Kollokationen ist ein nicht-triviales Problem, weil deren besondere Eigenschaften nicht aus den Eigenschaften ihrer Bestandteile vorhergesagt werden können.

Die vorliegende Arbeit beschäftigt sich mit Mehrworteinheiten (MWUs, von engl. *multi-word unit*), Kollokationen, die als komplexe Nominalphrasen auftreten. Komplexe Nominalphrasen sind für NLP von besonderer Bedeutung, da sie Objekte und Konzepte der realen Welt bezeichnen und häufig in Fachbegriffen auftreten, so zum Beispiel in wissenschaftlichen oder juristischen Begriffen.

Beinahe jedes NLP-System beruht auf dem Vorverarbeitungsschritt der Tokenisierung, der Unterteilung textueller Daten in bedeutungstragende Einheiten, sogenannter Tokens. Für gewöhnlich beinhaltet Tokenisierung keine Behandlung von Mehrworteinheiten, was zu frühen Fehlern, Fehlerfortpflanzung und schlechterer Qualität in NLP-Anwendungen führt.

In der vorliegenden Arbeit schlagen wir Mehrwort-Tokenisierung (MWT, engl. *multi-word tokenization*) vor, Tokenisierung, die Mehrworteinheiten erkennt. Ziel unserer Arbeit ist, Forschung voranzutreiben, die es Anwendungen ermöglicht, uneingeschränkte natürliche Sprache verstehen. Die Hauptbeiträge decken zwei Bereiche ab, die für MWT relevant sind.

Erstens präsentieren wir Grundlagenforschung zu asymmetrischer Assoziation, dem Phänomen, das lexikalische Assoziation zwischen den Bestandteilen von MWUs unterschiedlich stark ausgeprägt sein kann. Diese Eigenschaft wurde bisher in der Literatur noch nicht tiefer gehend behandelt. Zum einen verorten wir asymmetrische Assoziation in einem breiteren Kontext verschiedener Typen von Wortassoziationen, zum anderen haben wir menschliche syntagmatische Assoziationen in einem dafür neu entwickelten Experiment gemessen. Wir zeigen, dass asymmetrische Assoziation ein Indikator dafür ist, dass eine Phrase eine MWUs ist. Außerdem haben wir korpus-basierte Assoziationsmaße entwickelt und gezeigt, dass Asymmetrie in Wortpaaren automatisch und mit hoher Genauigkeit vorhergesagt werde kann.

Zweitens präsentieren wir eine MWT-Implementierung, in der MWU-Erkennung als Klassifikationsproblem definiert wird. Dazu haben wir einen Klassifikator entwickelt, dessen Features auf MWU-Eigenschaften zugeschnitten sind. Dabei zielen wir insbesondere auf Nicht-Kompositionalität ab, das Phänomen unvorhersehbarer Bedeutungsverschiebungen, das in vielen MWUs auftritt. Zur Erkennung von Bedeutungsverschiebungen benutzen wir Features kontextueller Ähnlichkeit, die auf distributioneller Semantik aufbauen. Wir zeigen, dass diese Features MWU-Klassifikation entscheidend verbessern, Aspekte ihrer Funktionsweise jedoch unzuverlässig sind. Darüber hinaus haben wir MWT in ein Information-Retrieval-System integriert und gezeigt, dass das Einbeziehen von MWU-Informationen die Leistung des Systems verbessert.

List of Tables

List of Figures

Acknowledgements

First of all, I would like to thank my supervisor Hinrich Schütze for supporting and guiding my research.
I would like thank my collaborators Beate Dorow, Florian Laws, Uli Heid, Christian Scheible and Jason Utt (from the *WordGraph* project) and Steffen Koch (from the *Scalable Visual Analytics* project). Thanks to Stefan Evert for his work and comments on asymmetric association measures and to Alok Kothari, Christina Lioma and Martin Forst for their work on semantic head recognition. A thank you to my external examiner Sebastian Padó for his input, in particular on lexical semantics.
Thank you to the colleagues who made working at the IMS such a place pleasant experience. Special thanks is due to Florian, with whom I shared an office for the most part of the journey. Florian was an excellent office mate who always had an open ear and patiently answered countless questions. I would like to thank Kyle Richardson, Stephen Roller, Charles Jochim and Michael Walsh for the corrections and suggestions they provided. Special thanks to Florian Laws, with whom I shared an office for most of the journey. Florian was an excellent office mate who always had an open ear and patiently answered countless questions.

I would to thank my parents for their unconditional support.

And Caroline for staying silly.

Chapter 1

Introduction

Tokenization, the process of dividing up text into meaningful units, or tokens, is a fundamental preprocessing step in almost all applications of natural language processing (NLP). The standard tokenization approach is single-word tokenization where input is split into words using white space characters as delimiters. This approach introduces errors at an early stage into the NLP pipeline because it ignores multi-word units. Multi-word units are collocations, habitual word combinations with idiosyncratic properties. Because of these properties, multi-word units require special treatment that goes beyond straightforward processing of individual words.

In this thesis, we propose multi-word tokenization for natural language processing to address the problem of single-word tokenization and multi-word units. We present (i) a supervised classification method to identify multi-word units based on their idiosyncratic properties and (ii) fundamental research into asymmetric association, a new property of multi-word units.

This introduction is organized as follows: we first describe collocations and their idiosyncratic properties. We then introduce statistical association measures which play a central role in collocation research. We define multi-word units and introduce the property of asymmetric association and asymmetric association measures. In the context of NLP applications, we discuss the benefits of multi-word tokenization over standard single-word tokenization. We conclude the introduction by summarizing our contributions.

1.1 Collocations and Idiosyncratic Properties

Achieving full command of a language goes beyond knowing its morphology, syntax and the semantics of its individual words. Speakers who are limited to this knowledge cannot understand everything that is said and written in the language, nor can they use the language to its full potential. This is because part of the expressivity of a language rests upon certain habitual word *combinations*. These combinations have additional function or meaning that often cannot, or can only partly, be derived from their components.

The are many different kinds of habitual word combinations. Virtually every grammatical category and every part of speech can be involved: noun phrases (*hot dog*, *red tape*, *credit crunch*), verb-object constructions and phrasal verbs (*kick the bucket*, *give up*), adverbial constructions (*by and large*, *to and fro*), prepositional phrases (*out of the frying pan into the fire*) and complete sentences (*There's no use crying over spilled milk*).

These habitual combinations have been given different names in the literature: collocations, idioms, multi-word expressions or multi-word units:

- Choueka (1988) defines collocations as "sequences of words whose unambiguous meaning cannot be derived from that of their components, and which therefore require specific entries in the dictionary."

- Manning and Schütze (1999) give a similar definition: "a collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon."

- Sag et al. (2002) define multi-word expressions as "idiosyncratic interpretations that cross word boundaries (or spaces)."

There are no clear boundaries between the definitions; the terminology is used interchangeably in the literature. Throughout this thesis, we use the term collocation as the top-level term to refer to habitual word combinations. We distinguish this from another term, multi-word unit, below.

Collocations are pervasive in language. For example, Jackendoff estimates that "[...] their number is about of the same magnitude as [...] single words of the vocabulary." (Jackendoff, 1997, p.156). All collocations have in common that they comprise multiple words and that they exhibit idiosyncratic properties with regards to morphology, syntax or semantics. Manning and Schütze (1999) define the following, widely accepted properties of collocations:

non-compositionality: Collocations carry meaning that cannot be inferred from the meaning of their components, e.g. *kick the bucket* means *to die* and not to physically kick a bucket.

non-modifiability: Collocations are restricted in terms of morphological or syntactical modification, e.g. changing *kick the bucket* to *kick a bucket* or *kick the small bucket* does not preserve the idiomatic status.

non-substitutability: Collocations are restricted in terms of replacing components with semantically similar words, e.g. *pail* is semantically close to *bucket* but *kick the pail* is not semantically close to *kick the bucket.*

Native speakers share the knowledge about collocations and children pick up this knowledge during language acquisition. Humans learn the meaning of new collocations either through context or through feedback from other speakers. For native speakers, learning new collocations is a life-long task. For foreign language learners, mastering collocations is an important step in increasing their proficiency in that foreign language.

If we equipped computers with the knowledge of a language's morphology and syntax, and the semantics of its individual words, collocations would still be missing and natural communication with computers would not be possible. In the long run, sophisticated NLP applications, for example personal assistants in mobile computers, need proper treatment of collocations to enable natural human-computer interaction. This thesis deals with improving automatic treatment of collocations to help NLP applications move closer to human-computer interaction using natural language.

1.1.1 An Alternative Definition of Collocations

In this thesis, we adopted Manning and Schütze's framework for describing the idiosyncratic properties of collocations because it captures the phenomena relevant for multi-word tokenization in a straightforward way. Sag et al. (2002) propose an alternative approach to collocations. Although their work focuses on deep syntactical analysis of verb constructions, they offer a different perspective on the basic concepts that we are concerned with in this thesis. Below, we briefly introduce their approach and discuss differences to, and commonalities with, Manning and Schütze view.

Sag et al. focus on what they call multi-word expressions[1] (MWEs) which they define as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002, p. 2). They identify MWEs as one of two major obstacles to successful NLP (the other being ambiguity). They distinguish two main classes of MWEs, *lexicalized phrases* and *institutionalized phrases.* Lexicalized phrases are phrases that are syntactically or semantically idiosyncratic (*hot dog*, *kick the bucket*, *get in line*). They represent the bulk of the phenomena that are discussed in the paper and they are further divided into a number of hierarchical subclasses. Institutionalized phrases are defined as semantically and syntactically compositional but they are idiosyncratic from a statistical point of view, i.e. they are so frequent and conventionally used, that they exhibit increased lexical association (see Section 1.2). Examples for institutionalized phrases are *traffic light* and *telephone booth.* Semantically, *telephone closet* is as plausible as *telephone booth*, but Sag et al. argue that it is not used because *telephone booth* has been institutionalized.

Sag et al. (2002) approach MWEs based on the difficulties they pose for deep linguistic analysis with standard NLP approaches. For most of the problems described we can make a connection to Manning and Schütze's characterization of collocations.

Sag et al. (2002) emphasize problems with MWEs from two different perspectives, (i) a "words-with-spaces" approach that simply tries to list

[1] Their use of the term MWE differs from ours; we exclusively use it for verb constructions. For the remainder of this discussion we use their MWE interpretation unless otherwise noted.

MWEs in the lexicon and (ii) “general, compositional methods of linguistic analysis”, that corresponds to a classic NLP pipeline (see Figure 1.1). The problems they report are:

overgeneration problem: A general, compositional generation system has no way of preventing generalization from correct MWEs like *telephone box* or *telephone booth* to non-MWEs like *telephone cabinet* or *telephone closet*. This problem can be linked to non-substitutability because the overgeneration step replaces *booth* and *box* with semantically similar words and the results are not MWEs.

idiomaticity problem: The same approach cannot predict the meaning of an MWE if it is unrelated to the meanings of the MWE’s parts. This problem is the counterpart of non-compositionality.

lexical proliferation problem: This problem addresses the inadequacy of listing MWEs in the dictionary. Some productive MWE families exemplified by light verbs (*take a walk/hike/trip/flight/etc.*), cause loss of generality when listed individually.

flexibility problem: This problem pertains to the syntactic flexibility of verb constructions that, in general, cannot be handled by the words-with-spaces approach (e.g. *look up the tower* vs. *look the tower up*).

For the problems of lexical proliferation and flexibility there are no direct analogies in Manning and Schütze’s (1999) three-property classification. Lexical proliferation, the fact that exhaustive lists are not a general solution, is a general problem with collocations. The flexibility problem is a narrow problem that applies mainly to phrasal verbs. We did not find an explicit analog to non-modifiability in Sag et al.’s work.

Sag et al. (2002) reserve the term collocation for all constructions that fall under their MWE definition and “any statistically significant cooccurrence” (Sag et al., 2002, p. 8), e.g. the frequent co-occurrence of *sell* and *house*. Here, the high frequency is simply due to real-world facts rather than linguistic factors. Such constructions are not classified as multi-word units or collocations in our terminology.

1.2 Association Measures

Collocations are complex linguistic phenomena that are difficult to capture automatically. When building NLP systems, the most common approach to collocations involves counting events in corpora and then calculating statistics based on these counts to indicate that a particular combination is a collocation.

Given a corpus, the problem is to distinguish between random co-occurrence and actual statistical association. Raw frequency counts are often insufficient for this task. High frequency is not *per se* a reliable indicator that a combination is indeed a collocation. For example, the combination *last year* is very frequent in news corpora but exhibits none of the idiosyncratic properties typical for collocations (see Section 2.1.4). If the components themselves are frequent enough, the high frequency of the combination might just be a coincidence. This is why corpus counts have to be further interpreted in some statistically meaningful way. The most widely used technique for this purpose is the use of *association measures*. Association measures quantify the lexical association between words, i.e. the statistical association between the event of one word occurring together with another word.

In simple terms, association measures are functions that assign to each word pair an association score that represents the pair's amount of lexical association. The calculation of the score is based on the pair's distribution in the corpus.

Research on collocations is tightly interwoven with statistical association measures. However, association measures are not directly aimed at any of the idiosyncratic properties of collocations. Collocations and association measures are linked by the assumption that the idiosyncratic properties of collocations are – at least to a certain extent – reflected in the collocation's co-occurrence patterns and that the degree of lexical association can serve as a universal indicator for the "interestingness" of a word combination. The idea behind this is the following: collocations are habitual word combinations; they appear with certain regularity in natural language. This regularity is reflected in the co-occurrence patterns of words in the corpus.

The background of lexical association measures is indispensable for discussing our contributions to asymmetric association and multi-word tokenization. We give a thorough introduction to statistical association measures in Section 2.1.

1.3 Multi-Word Units

This thesis focuses on multi-word units. We define a multi-word unit (MWU) as a type of collocation, namely a noun phrase (NP) consisting of multiple contiguous words that have to be processed as a whole. We reserve the term multi-word *expression* (MWE) for collocations that involve verbs such as verb-NP, verb-PP constructions and phrasal verbs.[2] The bold noun phrases in (1.1) and (1.2) are MWUs. In both sentences they act as a single unit.

(1.1) He ate a **hot dog** in three big bites.

(1.2) There are regulations, laws and **red tape**.

MWUs are worth examining because they make up a considerable part of language. Nicholson and Baldwin (2008) found 345 MWUs in a 1000 sentence sample of the British National Corpus (BNC).[3] Noun phrases are used to refer to concepts and objects in the real world. In particular, MWUs are used to encode domain-specific terminology. Tanaka and Matsuo (1999) report that in a bilingual terminology dictionary of financial terminology of 105,000 entries, 30% of Japanese terms are noun compounds and 37% of English terms are either noun compounds or adjective-noun combinations.

In this dissertation, we make two main contributions to research on MWUs. The first contribution deals with a theoretical aspect. We investigate a property of MWUs that so far has received very little attention in

[2] The processing of MWEs is similar to that of MWUs but has its own unique challenges. MWE processing is beyond the scope of this thesis.

[3] The number refers to noun compounds, MWUs consisting only of nouns. The frequency of non-restricted MWUs including, for example, adjective-noun combinations should be higher.

the computational linguistics literature: the association between the components of many MWUs is asymmetric. Our second major contribution focuses on a practical problem, namely a new approach to the automatic detection of MWUs in the context of tokenization. We describe these two contributions in more detail in the next two sections.

1.4 Asymmetry in MWUs

As collocations, MWUs exhibit the three properties mentioned above. For example, *hot dog* is non-compositional because it does not refer to a hot animal but a sausage in a bun. *Red tape*, as in excessive bureaucracy, exhibits both non-modifiability and non-substitutability. First, it has its special meaning only when used in singular. Second, *red cassette* is not similar in meaning to *red tape* even though *cassette* is similar in meaning to *tape*.

These three criteria are well-understood and there are many publications addressing them in the literature.[4] One of the goals of this thesis is to gain a deeper understanding of the theoretical properties of MWUs. To this end, we want to emphasize another idiosyncratic property that has not received a lot of attention in the literature. This property is asymmetric association, as defined below.

asymmetric association: In some MWUs, lexical association between components is much stronger from one component to another than vice versa.

An example of asymmetry in MWUs is the phrase *wishful thinking*. There is a stronger association from *wishful* to *thinking* than there is from *thinking* to *wishful*. The first word strongly suggests that the following word is *thinking* while the effect is not as strong in the reverse direction. *Thinking* as the second word of a combination leaves a range of plausible first components, for example *quick*, *good* or *clear*.[5]

[4] See, for example, the on-going workshop series on multi-word expressions (`http://multiword.sourceforge.net/`)

[5] The resulting combinations are not necessarily MWUs.

Asymmetry in MWUs constitutes one out of four possible types of word association. First, word associations can be distinguished by being either paradigmatic or syntagmatic. Words are paradigmatically related if they can be substituted for each other. In contrast, words are syntagmatically related if they occur in sequence. Second, word associations can be either symmetric or asymmetric. This distinction is orthogonal to the previous one. In NLP, word associations and word similarity are commonly assumed to be symmetric relations. Experimental evidence suggests, however, that cognitively, similarity can also be asymmetrical. Table 1.1 summarizes the four types of word association that exist along the two axes *paradigmatic-syntagmatic* (first and second row) and *symmetric-asymmetric* (first and second column). The first cell contains *good-bad*, a word pair whose association is paradigmatic and symmetric. The second cell represents paradigmatic associations that are asymmetric, for example, the pair *bird-canary.* The third cell shows an example of a symmetric syntagmatic word association, *epileptic-seizure.*

The subject of our research are asymmetric MWUs which are located in the bottom right cell. MWUs – as well as all other types of collocations – fall into the syntagmatic category. Because the syntagmatic level emphasizes sequentiality, there is an inherent directional aspect in asymmetric MWUs. Consequently, syntagmatic asymmetry can be further distinguished by the direction of stronger lexical association. If the first word predicts the second, we have right-predictive MWUs. If the second predicts the first, we speak of left-predictive MWUs. The above example, *wishful thinking*, is right-predictive (RP); *high fidelity* is an example of a left-predictive (LP) MWU.

There is little prior work on this aspect of MWUs and how to measure it. We developed corpus-based measures that capture asymmetric association in MWUs. The measures can distinguish right-predictive from left-predictive MWUs. These *asymmetric association measures* extend classic association measures. They are pairs of functions that assign to each word pair two association scores: forward association (association from the first to the second word) and backward association (association from the second to the first word).

paradigmatic + symmetric	**paradigmatic + asymmetric**
good *bad*	*bird* *canary*
syntagmatic + symmetric	**syntagmatic + asymmetric**
epileptic seizure	*wishful thinking* (RP), *high fidelity* (LP)

Table 1.1: Four types of word association. We corroborate this classification with experimental evidence in Chapter 3.

Word association data sets investigated in the psycholinguistic literature predominantly capture paradigmatic associations. For our investigation of asymmetry in MWUs, we designed a novel experiment setup and captured syntagmatic associations. We found asymmetry in human syntagmatic word associations in the resulting data set. We showed that with asymmetric association measures we can predict the asymmetry in those associations with high accuracy, demonstrating the theoretical justification and the feasibility of these new measures.

1.5 Single-Word and Multi-Word Tokenization

As discussed above, our second contribution is a new method for the automatic detection of MWUs. This is an important practical problem for NLP because the first step in the processing pipeline of NLP systems is usually tokenization, i.e. the detection of the basic units that will be used in subsequent stages of processing. Standard tokenization in NLP returns a stream of single words. With an effective module for MWU recognition in place, this standard tokenization procedure can be replaced by a more useful module that recognizes single words as well as MWUs.

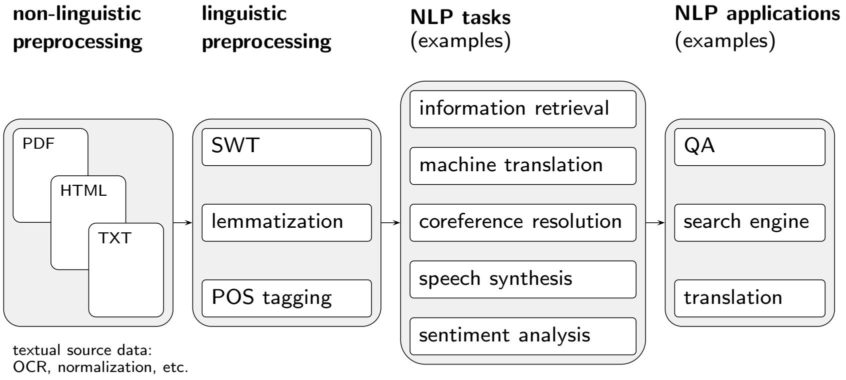


Figure 1.1: Typical NLP pipeline with single-word tokenization (SWT)

1.5.1 Single-Word Tokenization

Tokenization is an integral part of virtually every NLP task. In intuitive terms, tokenization is commonly understood as "splitting text into words." In this section, we describe how tokenization is typically handled in NLP and what problems arise with respect to the processing of MWUs. Motivated by the challenges we faced in two different research projects, we propose a new way of tokenization that is geared towards MWUs. We discuss our approach in the context of a typical NLP pipeline.

Figure 1.1 shows a typical NLP pipeline. The system is roughly divided into four modules: non-linguistic preprocessing, linguistic preprocessing, NLP tasks and NLP applications.

The first module deals with data acquisition.[6] This includes collecting data from different sources like web pages, databases, PDF documents or scanned books (OCR[7]) and converting them into a format that can be processed by standard NLP tools (e.g. UTF-8 or ASCII).

The second stage deals with linguistic preprocessing. Tokenization is the first step of this stage. Commonly, some form of lemmatization or stemming

[6] In this thesis, we discuss an NLP pipeline based on textual data.

[7] OCR, or optical character recognition, is the conversion of images of text into characters.

and part-of-speech (POS) tagging is performed. The de facto standard approach to tokenization in NLP is to split text into words using white space characters as delimiters and isolate punctuation symbols when necessary. This approach originates from the Penn Treebank project (Marcus et al., 1993) and has remained essentially unchanged for the last two decades (Dridan and Oepen, 2012). We refer to this tokenization approach as single-word tokenization (SWT). Tokenization is arguably the first step in preprocessing that is linguistically motivated because it identifies the basic units on which all other processing is based.[8]

NLP tasks, the main part of the pipeline, are the third step. Here, NLP algorithms are applied and combined to solve or assist in solving particular problems, e.g. speech synthesis, sentiment analysis, co-reference resolution, machine translation or information retrieval.

The final step consists of NLP applications. NLP applications integrate the results of NLP tasks with other components, e.g. the user interface, and make them accessible for end users. Typical NLP application inlcude search engines, question answering systems or translation services.

1.5.2 Multi-Word Tokenization

In this section we motivate, define and discuss multi-word tokenization, and introduce our approach to tokenization that is geared towards MWUs.

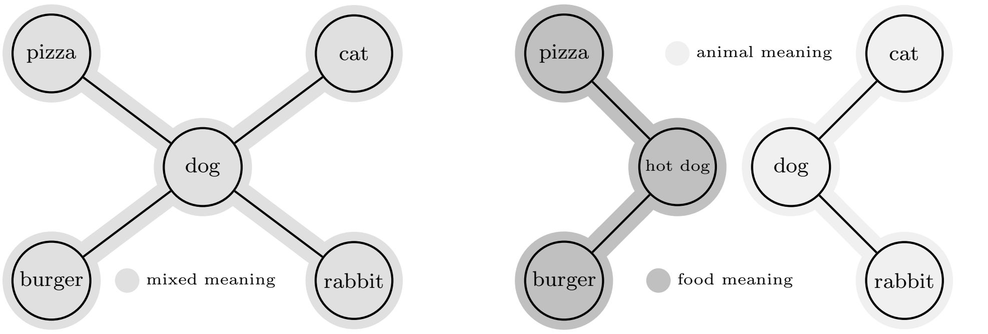
Motivation

From the perspective of automatic MWU processing, SWT is problematic. MWUs comprise multiple words and SWT offers no mechanism to process multiple words as the same unit. Mistakes made with SWT propagate to the following stages and result in poorer quality in the final application.

[8] It could be argued that, for example, converting German text encoded in UTF-8 into ASCII during the first stage and removing diacritics is a form of linguistic processing because linguistic information can get lost (e.g. converting *Äpfel* (apples) to *Apfel* (apple) turns plural into singular). We consider this to be a technical issue as opposed to linguistically-motivated transformation.

In the *WordGraph* project, we encountered this problem with SWT. One of the goals of the project was to develop a graph-based word-similarity measure for bilingual lexicon extraction (Laws et al., 2010). We created a graph of nouns, connecting two noun vertices if they appeared in the same coordination, e.g. *cats and dogs*. We used SWT to map surface text to noun nodes.

The word *cat* often appears in coordinations with *dog* and the word *burger* often appears with the MWU *hot dog*. With SWT-based mapping, *dog* and *hot dog* were unified into the same node *dog*. The similarity measure found the word semantically closest to *cat* to be *burger*. The reason for this lies in the structure of the graph and the recursiveness of the similarity measure which spreads similarity across the graph beyond first-order, i.e. direct, neighbors. In this scenario, the node *dog* acts as a semantic bridge between the *cat* and *burger* nodes resulting in these nodes becoming similar. With SWT, the animal meaning of *dog* and the food meaning of *hot dog* are put together in one node that ends up representing a mix of both meanings (Figure 1.2a).



(a) SWT: mixed meaning of *dog*, *hot dog* (b) MWT: Separate nodes for *dog*, *hot dog*

Figure 1.2: Effect of different tokenization techniques on graph structure and similarity spread

In the *Scalable Visual Patent Analysis* project, we worked on information retrieval for patent data. One of the goals of the project was to develop advanced visualization techniques that allow users to quickly get an overview of patents, for example lawyers conducting prior art searches. Patents contain a large number of domain-specific terminology, for example terms from electrical engineering or chemistry. Domain-specific vocabulary is commonly encoded in MWUs such as *free radical* or *absolute pressure transducer*. With SWT, there is no structured way to extract domain-specific MWUs for visual patent search.

The challenges regarding MWU processing we faced in the two projects encouraged us to explore a structured approach to MWU-aware tokenization. We developed multi-word tokenization which we define in the next section.

Definition

We define multi-word tokenization (MWT) as tokenization that automatically recognizes MWUs. We propose MWT as an intermediate step between basic linguistic preprocessing (which includes SWT) and higher-level NLP applications. MWT requires SWT to identify the words that are the building blocks for MWUs.[9]

The approach to MWT we suggest here works as follows: given input text, we collect MWU *candidates*, i.e. noun phrases from the input. MWUs are then selected automatically from the candidates. The result of MWT is tokenized text where MWUs are treated as single tokens. Figure 1.3 shows an overview of MWT. We describe the different stages with the help of an example.

preprocessed input: The input for MWT are single-word tokens coming out of SWT. Other linguistic preprocessing steps, for example lemmatization and POS tagging, are not essential to MWT but information from these steps can be incorporated into MWT for candidate and feature extraction.

[9] The proposed approach focuses on the English language. Because of its orthography, English compound nouns are often divided by spaces which makes the language susceptible to errors caused by SWT.

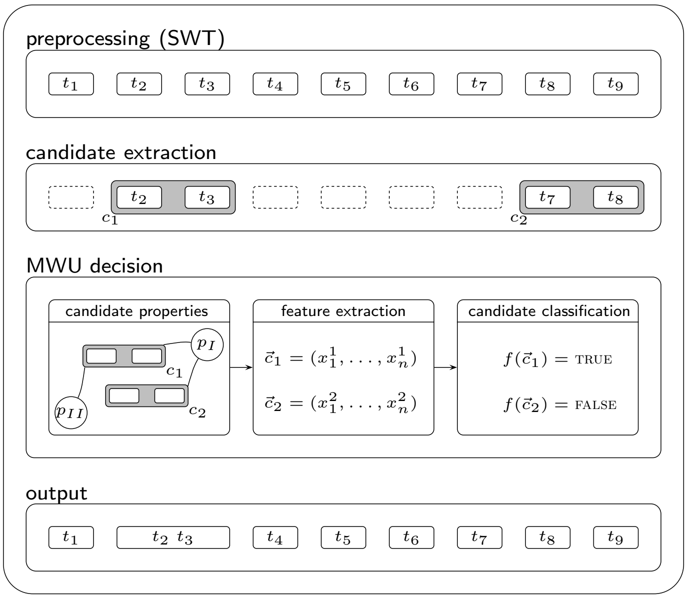


Figure 1.3: Stages of multi-word tokenization (MWT)

In the example, we have an input of nine tokens. The tokens represent the input *My hot dog was stolen by a small child*

candidate extraction: MWU candidates are extracted from the tokens. Candidates can be extracted with shallow linguistic filters, e.g. groups of adjacent tokens or POS patterns that match noun phrases depending on what information is available.

We extract two NPs as candidates, c_1 = *hot dog* and c_2 = *small child.*

MWU decision: We see MWT as a classification problem. The goal is to translate the idiosyncratic properties of MWUs into features and classify candidates into MWUs and non-MWUs based on feature values. Features are based on the candidate's distributional behavior across a text corpus.

We consider two properties, p_I and p_{II}. Suppose p_I represents commonness and p_{II} non-compositionality. Candidate c_1 has both properties and c_2 has only the former, i.e. *hot dog* and *small child* both are common candidates; *hot dog* is non-compositional whereas *small child* is not. Features that translate p_I and p_{II} are frequency of occurrence and semantic similarity measures, respectively.

Candidates are then turned into feature vectors $\vec{c_1}$ and $\vec{c_2}$. Feature vectors contain numerical representations of MWU properties, e.g. the number of times *hot dog* appeared in the data or a quantification of the degree of non-compositionality of *hot dog*.

A binary classification function $f(\vec{c}) \rightarrow \{\text{TRUE}, \text{FALSE}\}$ takes a feature vector as input and decides if the candidate is an MWU (TRUE) or not (FALSE).

In our example, *hot dog* is classified as an MWU, while *small child* is not.

output: This step integrates the classification results with the input token stream. The output is tokenized text containing multi-word tokens. A multi-word token is a token that contains multiple tokens from SWT. The output provides the input to subsequent processing, i.e. NLP tasks.

The tokens t_2 and t_3 are now merged into a single multi-word token (*hot dog*).

The features we use for MWU decision in this thesis are based on association measures as a general purpose tool to detect lexical association and semantic similarity to address non-compositionality. Asymmetric association measures offer another possible feature for candidate classification. The phrase *high fidelity*, for example, exhibits relatively low lexical association according to traditional association measures. In a feature, this information would not help the classifier to identify it as an MWU. The phrase is, however, asymmetric in the sense described above. Assume p_{III} represents asymmetry. With asymmetric association measures, we can translate p_{III} into a feature value, for example the ratio of forward and backward association.[10]

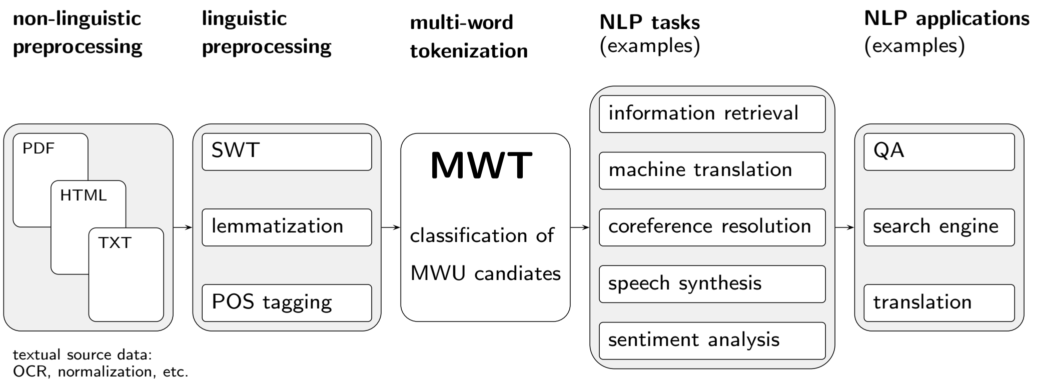


Figure 1.4: Multi-word tokenization (MWT) integrated into NLP pipeline

Discussion

We argue that MWT helps to eliminate early errors committed by SWT. Consider the graph similarity example Figure 1.2. With MWT, we can map *dog* and *hot dog* to two separate nodes. This would eliminate the unintended semantic bridge between *cat* and *burger* and separate the animal meaning of *dog* from the food meaning of *hot dog* as shown in Figure 1.2b. In the case of the patent search engine, MWT facilitates presenting to users the important domain-specific terminology contained in a certain document and encoded in MWUs.

The benefit of using MWT is that it results in MWU-aware input for all following steps in the NLP pipeline. Figure 1.4 shows an NLP pipeline that includes MWT. Running NLP tasks on data preprocessed with MWT will result in NLP applications of higher quality for end users. We give additional NLP tasks that we believe would benefit from MWT:

- In information retrieval, treating MWUs as single units can help return documents more relevant to a query. For example, the query *buy red tape* should not return documents about bureaucracy but about tape of red color and the query *reduce red tape* should not return documents

[10] This feature is unimplemented.

about tape of red color but about initiatives to eliminate bureaucracy.

- Machine translation would benefit from correct MWU tokenization because simple word-to-word translation fails when MWUs are involved (Melamed, 1997; Callison-Burch et al., 2006), e.g. in German, the literal translation of *red tape* is unrelated to bureaucracy.

- The knowledge that multiple words form a semantic unit can be a valuable source of information for prosody generation in text-to-speech synthesis, e.g. "...[i]n normal adjective noun patterns (e.g. *large site*) the stress normally goes on the second word, but in cases such as *web site* it goes on the first, as would be the case in single words such as parasite." (Taylor, 2009, pp. 62, 137).

- Lexical resources for sentiment analysis are usually based on unigrams (Pang and Lee, 2008). This approach fails when a negative sentiment item is part of an MWU which in turn has neutral polarity. For example, if a movie review uses the MWU *bad blood* to describe the relation between two characters, a sentiment analysis system that uses SWT picks up the negative item *bad*.

- Green et al. (2011) have shown that merging MWUs into single tokens improves performance for a variety of NLP tasks, e.g. syntactic parsing and language generation.

Generally, MWT is of use for those NLP tasks that process information represented as noun phrases, e.g. information extraction, semantic role labeling or coreference resolution. In addition, MWT increases the accuracy of word space models, an important building block in NLP.

To a certain extent, MWT could be achieved with static MWU lists. If we had an all-encompassing list of MWUs, we could simply look up each candidate and mark it as an MWU if it is on the list. Kulkarni and Finlayson (2011) presented jMWE, a toolkit that marks known MWUs in text. The toolkit relies solely on an external list of MWUs. Such a resource is expensive to create and maintain, even more so for specialized domains. Another shortcoming of MWU lists is currentness of data. The lists have to be constantly expanded to keep up with new vocabulary.[11]

In contrast, the goal of MWT is to recognize MWUs dynamically by exploiting MWU properties for classification. From a textual resource, we can extract candidates and features whose values encode the candidate's behavior. If we work on current data, new vocabulary will naturally be among the candidates. MWT classifies candidates automatically and does not rely on human judges making every single MWU decision by hand, as is the case with MWU lists. To our knowledge, there is currently no off-the-shelf system devoted to the problem of automatic multi-word tokenization. The ultimate goal of MWT is to fill this gap.

The decision whether a unit is an MWU or not is often hard to make, in large part because it can be context-dependent. For example, *black box* is an MWU in the context of aviation but not in the context of storage. Similarly, *red tape* can refer to actual tape that has the color red. The approach to MWT that we propose in this thesis makes one global MWU decision for each candidate. The decision is independent of the context in which the MWU appears. This means our approach will overgenerate (e.g. when *black box* is used in the storage setting) or undergenerate, e.g. if it was trained to prefer the compositional reading of *red tape.*

Consequently, when used in applications, the system will misclassify candidates even if carefully tuned to the domain in question. The same problem holds for SWT, which is an approach that operates on the assumption that all word combinations are compositional.[12] In the long run, with MWT we are striving for a solution that combines SWT with conservative, application-dependent recognition of the MWUs that are missed by SWT.

[11] For example, the MWU *web site* is common today but does not appear in the BNC, which was compiled in the 1990s.

[12] As we will describe in Chapter 2 and Chapter 4, our approach is distributional, i.e. ultimately, the MWU decision is influenced by the choice of the training corpus. Each distributional approach has the drawback that it learns from usage patterns of words and phrases typical for the training data. The learned patterns can be appropriate for one scenario but not for another.

1.6 Thesis Contributions and Structure

In this thesis, we make two contributions to the field of MWU research. First, we investigate a new theoretical property in MWUs, asymmetric association. Second, we address a practical issue with MWUs by proposing multi-word tokenization, a new approach to tokenization that automatically detects MWUs.

Asymmetric association in MWUs We conducted a study that investigates asymmetric association in both corpora and human production and created novel asymmetric association measures. Our contribution in this area is fundamental research with potential use for MWT. Our investigation of asymmetry in MWUs covers three aspects:

(i) asymmetry in corpus data: We created corpus-based measures of asymmetric lexical association. These *asymmetric association measures* quantify lexical association in word pairs in two directions unlike classic association measures that quantify lexical association as a whole. (ii) asymmetry in human production: We captured asymmetric, syntagmatic association in human production in an experiment. We designed a novel experiment setup because existing experiments and the resulting data sets predominantly cover paradigmatic and symmetric relations. (iii) We showed that the corpus-based measures can predict asymmetry in human production with high accuracy.

Multi-word tokenization We propose MWT, an approach to the detection of multi-word units in tokenization for NLP. We highlighted the problems that state-of-the-art tokenization faces with respect to MWUs. We argue that MWT will benefit a wide range of NLP tasks and applications. We see MWT as a classification problem where features are engineered to capture idiosyncratic MWU properties. In particular, we present:

(i) an implementation of MWT, (ii) an intrinsic evaluation of its performance and (iii) an extrinsic evaluation that incorporates MWU information into an information retrieval experiment. Our MWT implementation includes different feature types that address different MWU properties, e.g. non-compositionality. Furthermore, we propose a cascaded classification approach for the MWU decision step that handles candidates of arbitrary length.

The remainder of this thesis is structured as follows. In Chapter 2, we give the theoretical background necessary for the following chapters. We explain the backgrounds of association measures, measures of semantic similarity and supervised machine learning. Chapter 3 covers our study on asymmetric association in MWUs. Chapter 4 presents our MWT implementation and its two-fold evaluation. In Chapter 5, we summarize this thesis and present our conclusions.

Chapter 2

Theoretical Background

In this chapter, we introduce the theoretical background necessary for the research on asymmetric association measures and MWT. The chapter is divided into three sections. In Section 2.1, we explain statistical association measures and introduce the ones used in this thesis. Association measures are fundamental for the definition of asymmetric association measures and they are the basis for the basic features for MWT. In MWT, we need features to detect non-compositionality which requires some form of computational meaning representation. In Section 2.2, we present an approach to meaning representation based on distributional semantics. Finally, we provide the background in supervised machine learning. Machine learning is central to the MWU decision step as part of MWT. We present the necessary background in Section 2.3.

2.1 Association Measures

In an influential study on collocation extraction, Choueka (1988) writes about his search for a "measure of the degree with which the different words [in a collocation] 'attract each other'." His idea captures association measures (AMs) in a nutshell. In this section, we introduce a range of such measures and the theoretic considerations behind them. We have already sketched association measures as functions that quantify lexical association by assigning association scores to word pairs; now we detail this process. We provide the mathematical foundations of AMs and describe how frequency counts are collected from corpora and stored in contingency tables, the basis for computation of all measures. We introduce and discuss the concrete measures used in this thesis and how AMs fit into NLP.

2.1.1 Lexical Association and Independence

Intuitively, speakers and researchers have long been aware of the fact that, in natural language, words are not combined randomly into phrases and sentences. Consider the following example. Based on the frequencies of the word *red* an the word *tape* in the British National Corpus, the expected number of occurrences of the phrase *red tape* throughout the corpus is two. In reality, *red tape* occurs 129 times in this sample. This means that *red tape* appears far more often than what we would expect under complete randomness.

The concept of lexical association captures this notion that there is more to the co-occurrence of words than pure chance. Informally, lexical association can be thought of as the "glue" in word combinations. Formally, lexical association can be defined with the aid of statistical independence. In probability theory, two events x and y are independent if $P(x \cap y) = P(x)P(y)$, or if the probability of the two events occurring together is the product of the probabilities of the individual events. In the example, x and y are the occurrences of *red* and *tape* and the joint event $x \cap y$ is the occurrence of the phrase *red tape*. This means that the less independent the co-occurrence of two words is the more lexical association there is between them. The basic assumption of most corpus-based studies is that collocations exhibit a higher degree of lexical association than random word combinations.

With the concept of independence, we can define association measures more precisely. They are essentially formulas that, given co-occurrence data, return an association score, a real number which is used as an indicator of how much evidence there is against independence. The more evidence, the higher the lexical association.

All association measures have a common goal, namely, to quantify the amount of lexical association between the two words in question. Although there are different families of measures whose underlying theories come from very different backgrounds, they all share one property: their score is computed based on frequency counts. These counts are collected in contingency tables which we describe in the next section.

2.1.2 Frequency Counts and Contingency Tables

We have seen that we can draw conclusions about the amount of lexical association in a word pair by analyzing co-occurrence patterns in text samples. In this section, we are going to detail the process of obtaining frequency counts from corpora and how association scores are computed from frequency counts. Once the counts have been obtained they are stored in contingency tables of observed and expected frequencies. Observed frequencies represent the actual frequency counts in the sample data while expected frequencies represent the counts we would expect under statistical independence. The extraction of frequency counts consists of annotation of linguistic information on the token-level, the specification of a co-occurrence relation and a filter step.

During linguistic preprocessing, tokens are annotated with basic linguistic information (see Figure 1.4 on page 33). During this stage, (single-word) tokens are enriched or annotated with different layers of basic linguistic information. Typically, there at least two layers, lemmata and part-of-speech information. Both are important sources of information for collecting frequency counts.

During lemmatization each word is annotated with its lemma, or canonical form, which is used to represent all realizations of the same lexeme (e.g. *does*, *did* and *done* are lemmatized to *do* and *mouse* and *mice* are lemmatized to *mouse*[1]). This step abstracts over surface forms, reducing the number of

[1] A similar normalization step is stemming where words are truncated and mapped to the same root. The Porter stemmer (Porter, 1980), for example, normalizes *weakness* and *weakly* to the same root *weak*. Stemming is the standard approach in information retrieval. However, it ignores parts of speech which makes it less suitable for linguistically

pairs for which frequencies will be counted.

Part-of-speech annotation adds to each token its part of speech, e.g. noun, adjective or verb. For English, the most commonly used tag set is the Penn Treebank tag set (Marcus et al., 1993) which we use in our experiments.

An important detail is that in the process, surface tokens are mapped to *types*. For the rest of the computation, only types are considered. Typically, lemmata serve as types.[2]

An important decision for the collection of frequency counts is the definition of a co-occurrence relation. The co-occurrence relation dictates which kind of collocation ends up being counted. An appropriate relation for MWU extraction is "noun modified by adjective or noun". Part-of-speech filters are an effective way of selecting co-occurrence relations for collocation extraction, in particular for MWUs (Justeson and Katz, 1995; Daille, 1996). MWUs consist of groups of adjacent tokens and are easily captured by straightforward morpho-syntactic patterns.

The compilation of corpora includes non-linguistic preprocessing, i.e. the acquisition and preparation of texts from different sources such as web pages. These steps introduce errors and noise into the data, for example, left-over markup code and other formatting errors. To reduce noise in the final data set, heuristic filters such as frequency thresholds and word-shape filters (e.g. only extract tokens that consist of letters) are commonly applied.

The remaining word pairs are the final data set and frequency counting is performed on this set. The set represents a normalized and cleaned up version of the underlying language sample in the corpus. Its elements are tuples (u, v) where u and v are realizations of the random variables U and V which represent the pair's first and second component.

The frequency data are collected in contingency tables. In the tables, we not only collect the frequency of the pair in question but also the frequency of the individual components with other words. Altogether, we populate the contingency table of observed frequencies with four counts, O_{11}, O_{12},

motivated tasks. For collocation extraction and the majority of NLP task, lemmatization is the preferred approach.

[2] Throughout this thesis, we use the term *word* to refer to surface tokens as well as types. The distinction should be clear from the context.

observed frequencies

$$O_{11} = |\{(u,v) \mid U = u \wedge V = v\}| \quad O_{12} = |\{(u,v) \mid U = u \wedge V \neq v\}|$$
$$O_{21} = |\{(u,v) \mid U \neq u \wedge V = v\}| \quad O_{22} = |\{(u,v) \mid U \neq u \wedge V \neq v\}|$$

marginal frequencies

$$R_1 = |\{(u,v) \mid U = u\}| \quad R_2 = |\{(u,v) \mid U \neq u\}|$$
$$C_1 = |\{(u,v) \mid V = v\}| \quad C_2 = |\{(u,v) \mid V \neq v\}|$$

Table 2.1: Computation of observed and marginal frequencies

O_{21} and O_{22}. Marginal frequencies R_1, R_2, C_1 and C_2 are required for the computation of expected frequencies. The quantities are counted as shown in Table 2.1 and the final tables are shown in Table 2.2.

The input for the computation of association scores is data from the contingency tables of observed and expected frequencies. The measures differ in which information from the tables they incorporate and how this information is put in relation. Note that in NLP, association measures are commonly applied to combinations consisting of exactly two words.

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

(a) observed frequencies

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

(b) expected frequencies

Table 2.2: Contingency tables

2.1.3 Hypothesis Testing

An important question in statistics is if some observed data came about coincidentally or if they are the (expected) manifestation of the properties of an underlying distribution. Hypothesis test are functions that take observed data as input and tell us if the observations are compatible with a hypothesis that we proposed about the underlying distribution. The following paragraphs give an example of hypothesis testing and relate the procedure to association measures. The example was adapted from Bücker (1999, p.208).

A company that makes light bulbs has changed its production process and wants to know if the change resulted in an increase of the light bulbs' average lifetime. Translated into the language of hypothesis testing, the question is if the mean μ of the random variable X that represents the lifetime of a light bulb has increased. The company knows that their old production process yielded light bulbs with an average lifetime of 1000 hours with a standard deviation σ of 100 hours.

We begin with the formulation of the null hypothesis H_0 and the alternative hypothesis H_1:

H_0: $\mu = 1000$ (lifetime did not increase)
H_1: $\mu > 1000$ (lifetime did increase)

Next, they draw a random sample of $n = 100$ light bulbs from the new production line and test the lifetime of these bulbs. The sample bulbs show an average lifetime of $\bar{x} = 1020$ hours. We want to know if this sample was drawn from a population with mean 1000 hours (H_0) or with a greater mean (H_1). A suitable test for this scenario is the Gauß test. It is used to check if a sample was drawn from a population with a certain mean when the standard deviation is known. Its test statistic T is shown in Equation 2.1.

$$T = \sqrt{n}\,\frac{\bar{x} - \mu_0}{\sigma} \tag{2.1}$$

To summarize, the following information is available:

$$\mu_0 = 1000 \quad \sigma = 100 \quad \bar{x} = 1020 \quad n = 100$$

With Equation 2.1, we get

$$T = 10 \, \frac{1020 - 1000}{100} = 2$$

The question remains whether the observed data provide enough evidence to reject H_0 and assume that lifetime has increased. This question is answered by checking if the value of T is greater than some critical value c. This value can be looked up in statistical tables. The test statistic is normally distributed which means that we can look up the critical value in a standard normal distribution table. The particular critical value depends on the parameter α which specifies the risk of falsely rejecting H_0 even though it is true. A typical value is $\alpha = .05$, i.e. a 5% chance of falsely rejecting H_0. The value $c_{.05}$ is 1.65 which is smaller than 2. The end result is that the company can be 95% $(1 - \alpha)$ sure that their new process has increased the average lifetime of their light bulbs.

Hypothesis testing has been applied to co-occurrence data for collocation extraction (e.g. Giuliano, 1965). In the context of collocation extraction, the null hypothesis is that a combination is not a collocation. In other words, with association measures, we try to disprove the hypothesis that two words occur together independently of each other, i.e. the two of them occurring together is mere chance.

For finding collocations, we use hypothesis testing with these hypotheses:

H_0: $P(u \cap v) = P(u)P(v)$ (u and v are independent)
H_1: $P(u \cap v) \neq P(u)P(v)$ (u and v are not independent)

The idea is that for collocations there is enough evidence in the corpus to reject H_0.

More precisely, when we speak of lexical association, we have to distinguish positive and negative association. If a word pair exhibits positive association, it occurs more often than chance; it exhibits negative association, it occurs less often than chance. In NLP, we are traditionally interested in discriminating between positive association or not – the latter could be no association or negative association.

Statistical tests are subdivided into different groups. Parametric tests check a certain parameter of a population, goodness-of-fit tests check if a certain distribution fits a set of observations and independence test check if events are independent of each other. However, they are all united by the concept of the null hypothesis against which empirical evidence is collected. For association measures in NLP, we use hypothesis tests as a means of quantifying evidence against the co-occurrence of certain word pairs being random events. This is what we do when we use co-occurrence data to detect high lexical association. Our primary concern is not the study of hypothesis tests but their application to language data.

2.1.4 Association Measures Used in this Thesis

We use eight different association measures in this thesis. We give the theoretical background of the measures when necessary and provide the simplest formula to compute the respective association score in terms of observed and expected frequencies.

The first four measures, t-score, z-score, chi-square and log-likelihood are based on hypothesis tests. They measure the amount of evidence there is against the null hypothesis of independence. The remaining four measures are the Dice coefficient, pointwise mutual information, symmetrical conditional probability and raw frequency. These measures cannot be subsumed under a common idea like measures based on hypothesis tests and their background will be explained in the corresponding subsections.

By convention, we define high association scores to indicate high (positive) lexical association. Not all measures are defined in such a way by default. We follow Evert's (2004) definition of the measures in order to ensure this property.

t-score

Student's t-test is used for the same purpose as the Gauß test in the example above. It compares the observed and expected mean to determine if a sample was drawn from a population with a certain mean. The difference is that for the t-test we use the sample standard deviation s instead that of the population σ. It has the following test statistic, known as the t-score which has a Student's t-distribution:

$$t = \sqrt{n}\,\frac{\bar{x} - \mu_0}{s}$$

We have to extend the t-test for use with co-occurrence data. We think of the corpus as a sequence of bigrams and an indicator variable that takes the value 1 each time the bigram is (u, v) and 0 each time it is a different bigram.

The occurrence probabilities of u and v are the maximum likelihood estimates $P(u) = \frac{C_1}{N}$ and $P(v) = \frac{R_1}{N}$, respectively. Now the sequence of bigrams is a Bernoulli trial with success probability $p = P(u)P(v)$. In a large enough corpus p will be very small. This means that for the distribution $\mu = p$ and $s \approx p$.

Usually, the standard deviation is approximated by $\sqrt{O_{11}}$ yielding the following formula for the t-score association measure:

$$\text{t-score} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

Essentially, the t-score specifies how far (positively or negatively) the observed frequency deviates from the expected frequency. The deviation is expressed in multiples of the sample variance.

z-score

The z-score measure is closely related to the t-score measure. The difference lies in the way the sample variance is approximated. For z-scores, expected frequencies are used as the scaling factor:

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

The z-score measure is used in Smadja's (1993) Xtract system.

chi-square

In mathematical statistics, Pearson's χ^2 test is the standard test for independence of two random variables. It is a hypothesis test with the test statistic

$$X^2 = \frac{N((O_{11}O_{22} - O_{12}O_{21})^2)}{R_1R_2C_1C_2}$$

which is asymptotically χ^2 distributed with one degree of freedom.

We use an equivalent form that is expressed in terms of observed and expected frequencies:

$$\text{chi-square} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Church and Gale (1991) applied a slight variation of the χ^2-test to the problem of extracting translation pairs from parallel text. Their version, called Φ^2, is used for ranking and differs only in that the test statistic is multiplied by N.

The X^2 test statistic is known to give a bad approximation of the limiting distribution when the contingency table contains small numbers. Consequently, the test tends to overestimate of rare events (Manning and Schütze, 1999).

log-likelihood (G^2)

Likelihood ratio tests belong to an entirely different family of hypothesis tests. With a likelihood ratio test we try to find out which of two hypothesis better explains the observed data or, to be more exact, how much more likely one hypothesis is than the other. Applied to co-occurrence data, the observed data are the contingency tables for a certain pair (u, v) and the hypotheses refer to the underlying model that generated them.

The first hypothesis

$$\mathbf{H_0} : P((u, v)) = P((u, v')) = p; v' \neq v$$

represents independence as the explanation of the observed data, i.e. the probability of u occurring is the same regardless whether the next word is v or not.

The second hypothesis

$$\mathbf{H_1} : P((u,v)) = p_1 \neq p_2 = P((u,v')); v' \neq v$$

represents dependence as the explanation of the observed data, i.e. that the probability of u occurring when it is followed by v is different from when it is followed by a word other than v.

Formally, a hypothesis can be thought of as a subspace of the of space of unknown parameters of the underlying statistical model. Here, p_1 and p_2 form the parameter space. Under the null hypothesis this space is restricted to a one-element space where $p = p_1 = p_2$.

At the core of the test is the likelihood ratio

$$\lambda = \frac{\max L(H_0)}{\max L(H_1)}$$

where the likelihood function L gives the likelihood of the observed data under one of the hypotheses. We define L as follows:

$$L(k,n,p) = p^k(1-p)^{n-k}$$

with maximum likelihood estimates

$$p = \frac{R_1}{N},\ p_1 = \frac{O_{11}}{C_1},\ p_2 = \frac{O_{12}}{C_2}$$

for the parameters.

With this, we get

$$\lambda = \frac{L(O_{11}, C_1, p)\ L(O_{12}, C_2, p)}{L(O_{11}, C_1, p_1)\ L(O_{12}, C_2, p_2)}$$

Finally, the actual likelihood ratio test is carried out using

$$-2\log\lambda$$

The likelihood ratio has the property that $-2\log\lambda$ is asymptotically χ^2

distributed and that convergence is approached very quickly.

Dunning (1993) proposed the log-likelihood measure as an alternative to measures that assume co-occurrence data to be normally distributed. In particular, he criticizes pointwise mutual information (Section 2.1.4), z-scores and Pearson's χ^2 test for yielding inaccurate results for rare words. He proposes a likelihood ratio test for the statistical analysis of text because these tests do not make normality assumptions and should thus be more suitable for both rare and common phenomenon. Dunning casts the counting of words as Bernoulli trials with a binomial distribution[3] and very low positive outcome probability and argues that in these cases approximation with a normal distribution is inaccurate. His claim is supported by the Zipfian nature of language data – in their totality, rare words make up a large proportion of language. For example, words with a frequency of 1 in 50,000 make up 20-30% of English newswire text. Additionally, content-bearing words and technical vocabulary are often rare words (Dunning, 1993).

There is a form of the log-likelihood measure that is easy to express in terms of observed and expected frequencies.

$$\text{log-likelihood} = 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

It is also referred to as the G^2 statistic.

Dice coefficient

The Dice coefficient was introduced by Dice (1945). It can be thought of as a similarity measures for sets. For two sets A and B, the coefficient is defined as

$$\frac{2|A \cap B|}{|A| + |B|}$$

For example, the similarity of the two documents A and B measured with the Dice coefficient is the ratio of the amount of words that appear in both

[3] Hence the resemblance of the likelihood function L to the probability mass function of the binomial distribution.

documents (times two) to the total number of words in both documents.

For co-occurrence data, the Dice measure is defined as

$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1}$$

In contrast to measuring the amount of evidence against independence like the measures based on hypothesis tests, the Dice coefficient focuses on the magnitude of association. It is therefore suited to identify pairs with a very high degree of lexical association (Smadja et al., 1996).

Pointwise Mutual Information

The background for the association measure known as pointwise mutual information is the concept of mutual information (MI) from information theory (Shannon, 1948). The idea is to measure for two random variables X and Y how much information about Y is contained in X or in other words the reduction in uncertainty about Y when knowing X. Shannon and Weaver (1949) define mutual information, denoted $I(X;Y)$, as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual information incorporates a ratio of the true probability of the joint event and the expected joint probability under independence. If the variables are independent, $I(X;Y)$ is 0, i.e. knowing one variable tells us nothing about the other.

In NLP, pointwise mutual information (PMI), a derived concept, is commonly used. Fano (1961) proposed PMI, denoted $I(x,y)$, as mutual information between particular points $x \in X$ and $y \in Y$:

$$I(x,y) = \frac{p(x,y)}{p(x)p(y)}$$

This concept was adapted for language data by Church and Hanks (1990) with words serving as points.[4] They used maximum likelihood estimates for

[4] They called the measure *association ratio.* Nowadays, the term PMI is preferred over

the probabilities $p(x)$ and $p(y)$. In terms of observed and expected frequencies, the computation translates to the following formula which is equivalent to Fano's:

$$\text{PMI} = \frac{O_{11}}{E_{11}}$$

The PMI measure has a major shortcoming as Manning and Schütze (1999) demonstrate. For pairs that almost exclusively occur together, mutual information increases the rarer the pair becomes. This behavior overestimates low-frequency events and ignores the intuition that frequent occurrence should be a good indicator of high lexical association (see also e.g. Evert and Krenn, 2001). The measure is, however, a good indicator for independence.

Symmetrical conditional probability

The conditional probability of an event A under the condition that B has already occurred is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

Statistical independence can be expressed in terms of conditional probability as follows:

$$P(A|B) = P(A)$$

Silva and Lopes (1999) proposed the symmetrical conditional probability measure for the correlation between u and v. The measure is the "mirrored" product of the conditional probability of u given v and vice versa.

$$P(u|v)P(v|u) = \frac{P((u,v))^2}{P(u)P(v)}$$

In terms of observed and expected frequencies this translates to

$$\text{scp} = \frac{{O_{11}}^2}{R_1 C_1}$$

association ratio in NLP (e.g. Turney, 2001).

Schone and Jurafsky (2001) selected this measure as one of the statistical measures in their re-ranking experiment with multi-word units.

Frequency

The frequency measure is the most straightforward association measure. It consists of only the observed frequency of the pair. The idea behind this measure is quite simply that a pair that occurs often must have some of the properties that make a collocation. The measure produces a large number of false positives, e.g. highly frequent co-occurrences such as *last year*[5]. Depending on the application, however, it is not as bad as its simplicity would suggest (e.g. Krenn and Evert, 2001).

$$\text{frequency} = O_{11}$$

Previous work has shown that co-occurrence frequency performs surprisingly well in collocation and terminology extraction tasks (e.g. Daille, 1996; Krenn and Evert, 2001; Wermter and Hahn, 2006).

2.1.5 Applications of Association Measures

Association measures are the classic approach to corpus-based collocation extraction (Choueka, 1988; Church and Hanks, 1990; Smadja, 1993; Dunning, 1993). Association scores are used to rank pairs by association strength and do further processing based on these lists. With co-occurrence data extracted from corpora, we can perform tests for thousands of word pairs automatically and rank them association strength. This is the traditional approach of collocation extraction where an n-best list of collocation candidates is created based on association measures. Typically, the list will be given to lexicographers or other domain experts who will classify the list into true and false positives one by one. Another approach is to set a threshold for association

[5] The combination *last year* does not exhibit the properties by which we define a collocation. It is compositional, modifiable (*over the last two years*) and components can be substitute with similar words, resulting in predictable changes of meaning, e.g. *next year* or *last week*.

scores above which pairs are considered to be collocations.

Other applications of association measures in NLP include sentence boundary detection (Kiss and Strunk, 2006), resolving PP-attachment (Hindle and Rooth, 1993), extract typical predicate-argument structures from corpora (Church et al., 1989), dimensions weights in word space models (Rapp, 1999; Padó and Lapata, 2007), modeling human association norms (Church and Hanks, 1990; Rapp, 2002) (cf. Chapter 3), natural language generation (Edmonds, 1997), and feature selection for automatic document classification (Manning et al., 2008).

2.1.6 Summary

Association measures are a means to quantify lexical association between words. Lexical association is the "glue" between words, their tendency to occur together more often than chance. Association measures compute association scores, indicators of association strength, based on frequency counts collected from corpora. Association measures are the basis for *asymmetric* association measures which we develop for our corpus-based analysis of asymmetry in human association presented in Chapter 3. They are also important for MWT, because they are a universal tool for MWU detection. The amount of lexical association that they capture is used as a feature in the MWU decision step Chapter 4.

2.2 Distributional Models of Meaning

We have seen in the introduction that the processing of MWUs faces challenges because of their idiosyncratic properties. One nut that is particularly hard to crack is the issue of semantic non-compositionality. In a compositional world, the meaning of a complex phrase would simply be the transparent composition of the meanings of its parts. Unfortunately, this is not the case for non-compositional phrases for which the meaning of the whole phrase is not a transparent composition of the meanings of its parts. For such phrases, given the meanings of the parts, we are unable to predict the meaning of their combination.

As we have shown, association measures target lexical association, a general quality of collocations. We need a measure of word similarity to address the specific property of non-compositionality. As part of MWT, we want to be able to address this problem automatically. To this end, we require a computational model of meaning. The model should be both theoretically justified and computationally feasible.

In this thesis, we build meaning representations within the framework of distributional semantics. The theoretical foundation of distributional semantics is given by the distributional hypothesis. Meaning representations based on distributional semantics can be computed from text corpora automatically.

In this section, we will introduce two variations of distributional semantics: the geometric word space model and a graph-theoretic approach. Both models are implementations based on the distributional hypothesis according to which the meaning of a word is derived from the context in which it appears. For both models, words and contexts are taken from corpora.

2.2.1 The Distributional Hypothesis

Imagine you are reading a text and you come across the unknown word *tezgüino*[6]. Throughout the text, there are several mentions of the word as shown in (2.2).

(2.2) A bottle of **tezgüino** is on the table.

Tezgüino makes you drunk.

We make **tezgüino** out of corn.

Over the course of the text, more and more information about *tezgüino* is revealed and you develop a clearer idea of its meaning. Considering how *tezgüino* is used in the examples, there is good reason to believe that it is a kind of alcoholic beverage produced from corn. What is the justification for

[6] The *tezgüino* example is taken from Lin (1998).

this assumption? Based on the knowledge that we have about things that are stored in bottles, make people drunk and are made out of corn we derive that *tezgüino* must be a substance similar to the things fitting the aforementioned categories, for example Whiskey.

This reasoning is captured in what has become known as the distributional hypothesis. The idea is usually attributed to Zellig Harris who wrote:

> The fact that, for example, not every adjective occurs with every noun can be used as a measure of meaning difference. More than that: if we consider words [...] A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris, 1954, p. 156)

Harris formulated the idea in terms of difference meaning deriving from difference in contexts. Today, the distributional hypothesis is formulated in terms of similarity (Miller and Charles, 1991). We define it as follows:

The distributional hypothesis: *Words that occur in the same contexts tend to have similar meanings.*

2.2.2 The Word Space Model

In the word space model, the meaning of a word is a point in a high-dimensional Euclidean vector space, the word space. In the word space, the idea of similar meaning through similar contexts is expressed as spatial proximity. Words that have similar meanings lie close to each other. Figure 2.1 shows a 2-dimensional word space illustrating this idea. *Cat* and *dog* are closer to each other than they are to *car* because they appear in similar (animal-related) contexts and less in car-related contexts. The rest of this section describes how word spaces are created automatically from textual resources.

The vector space model from information retrieval provides the mathematical framework in which the distributional hypothesis is embedded. It

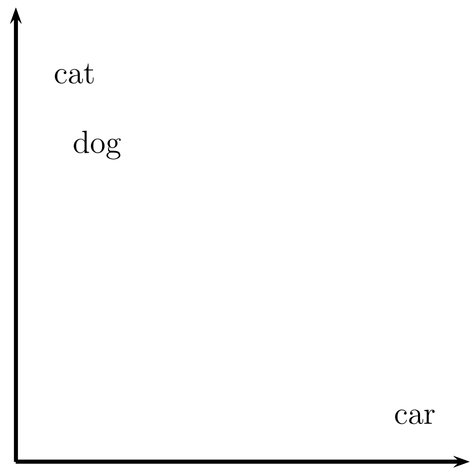


Figure 2.1: Example of a 2-dimensional word space

was originally developed by Salton et al. (1975) for the SMART information retrieval system. It is generally regarded as the first model of its kind (Manning et al., 2008). In this model, documents and queries are represented as vectors of words. The relevance of a document to a particular query is determined by the similarity of the document vector and the query vector.

The vector space model is readily adapted for the computation of word similarities. The initial step is the creation of a word-context matrix that records for each word the words that appear in its context (Schütze, 1992, 1993). Consider the sample text in (2.3).

(2.3) My old car broke down.
I need a car to drive to work.

We are going to build the word-context matrix **X** for these two sentences. We define context as occurring in the same sentence. **X** is populated by entering the number of times word w_j appears in the context of word w_i at position x_{ij}. Figure 2.2 shows the resulting word-context matrix.

	I	**My**	**a**	**broke**	**car**	**down**	**drive**	**need**	**old**	**to**	**work**
I	0	0	1	0	1	0	1	1	0	1	1
My	0	0	0	1	1	1	0	0	1	0	0
a	1	0	0	0	1	0	1	1	0	1	1
broke	0	1	0	1	1	1	0	0	0	0	0
car	1	1	1	1	0	1	1	1	1	1	1
down	0	1	0	1	1	1	0	0	0	0	0
drive	1	0	1	0	1	0	1	1	0	1	1
need	1	0	1	0	1	0	1	1	0	1	1
old	0	1	0	1	1	1	0	0	0	0	0
to	1	0	1	0	1	0	1	1	0	0	1
work	1	0	1	0	1	0	1	1	0	1	1

Figure 2.2: Word-context matrix for example (2.3)

The order of the rows is arbitrary as is the choice to represent words as rows and contexts as columns. The i-th row vector $\vec{x}_i$ represents the word w_i. The vector $\vec{x}_i$ represents the entirety of contexts in which w_i appears. We call $\vec{x}_i$, which reflects the distribution of w_i throughout the complete text sample, the semantic vector of word w_i. This is in the spirit of Harris who wrote:

> The distribution of an element will be understood as the sum of all its environments. (Harris, 1970, p. 775)

The word's semantic vector is its meaning because it determines its location in the word space.

Word similarities are computed between semantic vectors. There are a number of vector similarity measures that have been applied in the context of word spaces. The Jaccard and Dice coefficients, two measures originating in set theory, compute the overlap over all context words and can be straightforwardly be used for similarity computation. There are distance metrics like the Manhattan distance and Euclidian distance[7].

[7] Any distance measure D can trivially be used as a similarity measure by using its reciprocal: $sim(\vec{x}, \vec{y}) = \frac{1}{D(\vec{x}, \vec{y})}$.

$$\begin{array}{rcccccccl} & & \textbf{pet} & \textbf{drive} & \textbf{walk} & \textbf{cute} & \textbf{cheap} & \textbf{old} & \\ \vec{x}_{car} = & (& 1 & 76 & 7 & 2 & 66 & 53 &) \\ \vec{x}_{cat} = & (& 31 & 1 & 10 & 38 & 3 & 43 &) \\ \vec{x}_{dog} = & (& 22 & 2 & 33 & 27 & 3 & 44 &) \end{array}$$

Figure 2.3: Semantic vectors

The standard measure used for similarity in word spaces is cosine similarity. It computes the cosine of the angle between two semantic vectors:

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|}$$

where the norm, or length normalization, of a vector with length k is given by

$$|\vec{x}| = \sqrt{\sum_{i=1}^{k} x_i^2}$$

and the dot product of two vectors of length k is defined as

$$\vec{x} \cdot \vec{y} = \sum_{i=1}^{k} x_i y_i$$

The cosine measure takes on values between 0 (minimum similarity) and 1 (maximum similarity).[8]

Figure 2.3 shows three example semantic vectors for the words *car*, *cat* and *dog* for a small number of exemplary context words. The cosine similarities between the vectors reflect the spatial metaphor: *cat* and *dog* are more similar (closer) to each other than they are to *dog* (cf. Figure 2.1).

[8] The codomain of the cos function is of course $[-1, 1]$ but the elements of semantic vectors are non-negative restricting similarity scores to $[0, 1]$.

$$cos(\vec{x}_{cat}, \vec{x}_{dog}) = .91$$

$$cos(\vec{x}_{cat}, \vec{x}_{car}) = .38$$

$$cos(\vec{x}_{dog}, \vec{x}_{car}) = .38$$

2.2.3 Aspects of Word Space Model Implementation

The introduction in the previous section showed one possible way of constructing a word space. There are, however, a variety of decisions to make and several parameters to tweak in the course of getting from corpora to word spaces. The choices greatly influence the characteristics of the word space.

Above all, there is the definition of context. The distributional hypothesis leaves this question open. Text structure is one source for context. We can define co-occurrence as co-occurrence in the same document, paragraph or sentence. Another widely-used approach is the co-occurrence window. Here, the co-occurrences of a target word are defined as all the words that appear around the target in an n-word window regardless of structural or other linguistic boundaries. Generally, the larger the window, the broader the scope of semantic similarity becomes. Models using this approach are called *bag-of-words* models because they treat all context words equally and do not capture relations (e.g. word order) between context words.

In contrast to contexts based on the text structure and co-occurrence windows, deep linguistic relations are another basis for context definitions (Padó and Lapata, 2007). With this approach, dimensions are enriched with a linguistic relation between the target and the context word, e.g. adjectival modification or a direct-object relation.

Another influential factor is the weighting scheme. Initially, semantic vectors contain raw frequency counts. In order to dampen the impact of highly frequent words, logarithmic functions can be applied to the counts. Another direction in weighting schemes uses lexical association measures to replace raw frequency counts with association scores of target and context words. Log-likelihood and PMI and have been used for this purpose.

Two prominent issues of the word space model are data sparseness and scalability. Like all statistical approaches to natural language processing, the word space model suffers from the problem that even huge amounts of sample data do not provide reliable estimates for all reasonable events, i.e. some positions in the word-context matrix will be zero although they potentially represent plausible co-occurrence. With increasing number of dimensions, similarity computation becomes more expensive. The dilemma is that the solution to the former problem is more data (find evidence for more plausible events by processing ever larger amounts of text) but the solution to the latter problem is less data (reduce the number of dimensions of the word-context matrix to speed up similarity computation).

The simplest way to reduce the number of dimensions is to remove possible context words. Successful filters include part-of-speech filters, filters based on dimension weights or stop word lists. Filters of this kind are applied before the creation of the word-context matrix.

A more sophisticated dimensionality reduction technique is latent semantic analysis[9] (LSA) (Landauer and Dumais, 1997). The basic idea is to approximate the high-dimensional word-context matrix with a low-dimensional matrix. We will see that apart from reducing dimensionality, LSA can be seen to alleviate data sparseness and discover latent semantic structures in the word space.

LSA takes advantage of singular value decomposition (SVD), a matrix factorization technique. Factorization is the process of decomposing a matrix into a set of matrices, called factors. The process is reversible; the product of the factors is the original matrix. We will only go as deep into the mathematical details of SVD as is needed to explain its effects on the word space. In the case of SVD, one of the factors is a matrix that contains the linearly independent dimensions (which are characterized by singular values) of the original space sorted by variance ("importance") in descending order. The trick of LSA is to truncate the singular value matrix before restoring the original word-context matrix. By leaving out dimensions with low variance, we force SVD to approximate the original matrix as good as possible

[9] The technique was first applied under the name of latent semantic indexing (LSI) in information retrieval (Deerwester et al., 1990).

using only the remaining important dimensions. The new matrix contains the information stored in the original matrix in condensed form since less dimensions are available to approximate the same information.

The procedure obviously results in a matrix with less dimensions achieving the desired simplification of similarity computation. However, the effects of LSA can also be interpreted from different angles (Turney and Pantel, 2010). LSA can be said to discover latent, or hidden, relationships in the data. A side-effect of SVD is that it reveals higher-order co-occurrences, i.e. in the new word space, there is similarity not only between words that appear in the *same* contexts but also in *similar* contexts. For example, the original matrix might contain two dimensions for two overlapping contexts, one dimension for the contexts of *boat* and one for the contexts of *ship*. LSA discovers that a more compact representation using only one dimension is possible without much loss of information. LSA can also be interpreted as a smoothing technique. The new *ship-boat* dimension unifies all contexts the two words appear in. This means that *ship* has access to the contexts that *boat* appeared in even if *ship* did not appear in these contexts itself and vice versa.

Dimensionality reduction techniques like LSA add complexity to standard word space models and they make word space creation more expensive. Another drawback of LSA is that newly discovered dimensions can become hard to interpret linguistically, e.g. when unrelated concepts are merged into one dimension. For our experiments on non-compositionality we need clearly interpretable word spaces for model introspection and error analysis. We decided to use standard word space models without LSA in our experiments.

2.2.4 A Graph-Theoretic Implementation

An entirely different implementation of the distributional hypothesis is given by a graph-based model. A graph consists of nodes and links. Nodes represent arbitrary objects and links between nodes represent relationships between objects. For example, a graph could be used to represent a social network with users as nodes and the friendship relation as links. Graphs are a very flexible framework because they can be easily adapted to different domains.

In a graph-theoretic implementation of distributional semantics, words become nodes and contexts become links. For example, Figure 2.4 shows a word graph with nouns and verbs as nodes and links representing the linguistic context of the direct-object relation.

Constructing word graphs is similar to constructing word spaces. The same information that underlies the word-context matrix underlies the word graph. However, the way this information is represented is different. In the graphs, there is no inherent representation of spatial distance as there is in vector spaces. Consequently, word graphs have a different way to express word similarity. In the word space model, similarity computation relied on the spatial metaphor and the angle between semantic vectors. In word graphs, similarity computation rests on the graph's link structure. The semantic knowledge gathered from words and contexts is encoded in the way the nodes are connected with each other.

A possible node-similarity for computing similarity in word graphs is the recursive SimRank algorithm (Jeh and Widom, 2002). The basic idea is that two nodes are similar if their neighbor nodes are similar. SimRank straightforwardly uses the similarity between neighboring nodes to compute node similarity. The definition of SimRank for the similarity S_{ij} between two nodes i and j with neighbors $N(i)$ and $N(j)$, is given by:

$$S_{ij} = \frac{c}{|N(i)|\ |N(j)|} \sum_{k \in N(i), l \in N(j)} S_{kl}$$

Initially, every node is only similar to itself ($\forall i\ S_{ii} = 1$). By iteration this initial mass of similarity spreads over the whole graph. The dampening factor c guarantees convergence. For our example, we get the following similarities (among others). It shows that the more common neighbors (verbs) the nouns have, the higher the similarity.

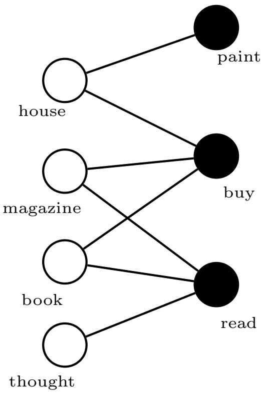


Figure 2.4: Word graph representing verb-object relations between nouns (white nodes) and verbs (black nodes)

$$S(\mathit{book}, \mathit{magazine}) = .58$$

$$S(\mathit{house}, \mathit{magazine}) = .44$$

$$S(\mathit{house}, \mathit{thought}) = .29$$

Initially, the word space model and the graph-based model seem equally suited for our purpose since they both implement the distributional hypothesis and both work on the same input data. We compared the two approaches. We manually evaluated the related words that each approach finds for a set of test words and explore how filters and weighting strategies influence the quality of the results. We came to the conclusion that the word-space model is the better choice for our experiments. See Appendix A for the detailed explanation of our decision.

2.2.5 Semantic Similarity Based on Structured Lexical Resources

Text corpora are usually unstructured except for document boundaries or section headings. There is no information about the relationships between the words in the corpus[10]. The word space model gives us semantic similarity between words because it applies the distributional hypothesis to large amounts of text.

A different approach is to use structured lexical resources instead of corpora as the basis for similarity computation. Possible sources are dictionaries, thesauri or WordNet and other semantic networks. In this line of work, lexical resources are interpreted as graphs and similarity is measured based on properties of paths in the graph, e.g. the shorter the path from one concept, the higher the similarity. See Budanitsky and Hirst (2006) for an overview and Pedersen et al. (2004) for off-the-shelf implementations of several WordNet-based measures. The advantage of these approaches is that they are similarity computation on a relatively small data set is less costly compared to building a word space model from corpora.

We ran experiments for MWT on the physics domain (Chapter 4) and therefore needed a similarity function for physics terminology. Hand-made lexical resources are expensive to build in particular for specialized domains. Given corpora, word space models can easily be built for special domains with existing tools.

2.2.6 Recent Developments in Distributional Semantics

This section discusses recent developments in distributional semantics and it is based on a survey that reviews word space models of word and phrase meaning (Erk, 2012).

Distributional Memory Baroni and Lenci (2010) propose a generalized framework of distributional semantics called distributional memory (DM). A DM stores the entirety of distributional information in a corpus as a third-order tensor. From this representation, semantic spaces (second-order tensors, i.e. matrices) can be generated.

[10] At least not explicitly. Hearst (1992), for example, extracted hyponymy relations from unstructured corpora using search patterns such as *X is a Y* or *Y such as X*.

The third-order tensor stores arbitrary linguistic relations r between two words w_1 and w_2 as weighted $\langle w_1, r, w_2 \rangle$ tuples, covering dependency relations ($\langle man, subj, sleep \rangle$) or lexico-syntactic patterns ($\langle animal, such\_as, dog \rangle$).

The motivation for DM derives from the observation that semantic spaces are commonly created in an ad hoc manner to address a particular NLP task. They argue in favor of a more general approach that extracts distributional information into a repository and enables on-demand creation of adequate semantic spaces from this repository via a matricization operation. In an empirical evaluation on a range of semantic tasks, DM performs competitively against state-of-the-art ad hoc solutions.

The appeal of DM lies in its flexibility. For example, models for computing the similarities between single words or relational similarity between word pairs are readily available from the DM thanks to a well-defined matricization operation.

Polysemy In word space models, words are by default represented in a prototype-based way: all occurrences of the word are conflated in the same vector, obscuring the different senses the word might have. Several approaches have tried to deal with polysemy in word space models. In one group of models the idea is to start out with the global word vector and modify it according to the context in which it is used (e.g. Erk and Padó, 2008; Mitchell and Lapata, 2008). Modification here means to modify the word vector in question incorporating the vectors of surrounding words. For that purpose, vectors need to be composed to yield the modified vector. Different composition functions have been used, for example vector addition, component-wise multiplication and tensor products.

Another approach to dealing with polysemy is clustering of a word's individual co-occurrence vectors. Clustering of individual occurrences reveals different word senses (Schütze, 1998; Reisinger and Mooney, 2010). In contrast to the prototype-based approach, the clustering approach uses exemplar vectors where each exemplar represents a context (e.g. a sentence) in which the target word occurred. Erk and Padó (2010) propose an exemplar-based approach that computes clusters dynamically by activating only a subset of exemplars based on the context of the target word.

Representations for phrases and sentences Analogous to word similarity, Mitchell and Lapata (2010) raised the question of how to compute phrase similarity, for example the similarity between *old man* and *elderly lady*.

One possibility is to represent phrases atomically treating it just like a standard word vector. However, the longer the phrases, the more severe the sparse data problem becomes. The solution is to look for ways to compose complex structures from simple vectors representing individual words.

Mitchell and Lapata (2008) introduced a general framework for vector composition which they use for two-word phrases. The composition functions mentioned above are used to combine vectors to represent phrases. In particular, the representation of noun phrases consisting of head noun and modifier has received attention in this area (e.g. Baroni and Zamparelli, 2010; Reddy et al., 2011; Hartung and Frank, 2011).

Going further, the ability to measure the similarity of larger constructs, sentences in particular, would be a useful tool for NLP tasks, e.g. paraphrase detection. For two-word noun phrases, a composition function that takes two basic elements such as adjectives and nouns is sufficient. Moving to sentences, the questions arise of how to represent and combine complex syntactic structures; e.g a desirable quality of such a model is to be able to distinguish *John loves Mary* from *Mary loves John*.

Several approaches deal with these questions. One group represents words and predicates as tensors (e.g. a matrix representing an intransitive verb) and applies tensor products to construct a meaning representation for a complete sentence. Models differ in whether they encode syntactic structure explicitly (e.g. Clark and Pulman, 2007) or implicitly (e.g. Coecke et al., 2010; Grefenstette and Sadrzadeh, 2011). In the former model, syntactic structure is visible in the resulting tensor; in the latter, it is not. A major shortcoming of these models is that they can only compute similarities of sentences that have identical syntactic structure.

Socher et al. (2011) presented a model that does not have this restriction. In their model, each constituent is represented by a vector, resulting in a tree of vectors for each sentence. Sentence similarity is then computed by computing pairwise similarities between all nodes from both trees. The resulting similarity matrix is the input to a classifier that decides whether the second sentence is a paraphrase of the paraphrase of the first.

In a third kind of approach, the goal is to combine symbolic, logic-based semantics (Montague, 1973) with distributional semantics (Garrette et al., 2011; Clarke, 2012). In these works the basic idea is to use the best of both worlds: the formal, symbolic machinery (which has no mechanism for word or phrase similarity) for the composition of complex structures and the distributional representation (which has no mechanism for sentence construction) to model the meaning of the simple elements.

Work on distributional approaches to representing sentence meaning is a new field. There are still problems with some approaches, e.g. computing similarities of sentences of different structure. A fundamental problem of compositional distributional semantics is that not all syntactic categories have obvious representations, e.g. it is unclear how to represent prepositional phrases (see paragraph *Non-Compositionality* below for a discussion of how these new models relate to MWT.

Non-Compositionality Biemann and Giesbrecht (2011) launched a shared task on compositionality grading where the (non-)compositionality of phrases has to be graded on a fine-grained numerical and a coarse-grained categorical scale.

The submitted systems consist mostly of traditional approaches based on statistical association measures and basic word space models. Reddy et al. (2011) submitted system that used prototype-based and an exemplar-based word space models. The prototype-based system slightly outperformed the exemplar-based one.

As we will show in Section 4.2.4, detection of semantic non-compositionality is a key ingredient of MWT. In principle, each of the models that try to represent phrases and sentences could be used for non-compositionality detection. The basic approach would be to detect meaning shifts between the components of a phrase and the phrase as a whole. Here, the meaning of the whole phrase would be computed using a certain composition function. However, the treatment of non-compositional semantics has not been a central question in this rather new branch of models. In Michelbacher et al. (2013)[11], we experimented with a number of composition functions for compositionality grading of adjective-noun combinations.

Non-Distributional Contexts There are models of word meaning that are based on contexts which are not gathered from corpus co-occurrences. For example, feature norms (e.g. McRae et al., 2005), i.e. data bases of objects and the features ascribed to them by human subjects, or non-textual content (e.g. images in Feng and Lapata (2010)). Both kinds of data can be used to populate word vectors. In this thesis we are only concerned with word space models built from corpus co-occurrences.

2.2.7 Summary

In this section, we introduced the word space model, a framework that enables us to measure semantic similarity between words and phrases. We use similarity computation in MWT for non-compositionality detection (see Section 4.2.4). The word space model is a model of distributional semantics which means that the meaning of a word is determined by the context in which it appears. In the word space model, words are represented in a high-dimensional vector space whose elements are words and dimensions represent contexts; semantic similarity is understood as spatial proximity in the space. We compared this implementation of distributional semantics with a graph-based approach that uses the same underlying information but represents contexts and computes similarity in a different way. In a comparison study we found the word space model to achieve better results.

[11] The work focuses on adapting unsupervised features to a new domain and goes beyond the scope of this thesis.

2.3 Supervised machine learning

2.3.1 Machine Learning in NLP

The beginnings of natural language processing happened during the 1940s and 1950s in connection with work on machine translation. For political reasons, there was strong demand for automatic translation, especially from Russian to English and vice versa. Many of the early approaches were of a probabilistic nature. However, high expectations, the absence of a quick breakthrough and the devastating verdict of the 1966 ALPAC report about the prospects of machine translation lead to an abandonment of statistical approaches.

The new direction of research was rule-based systems. Motivated by Chomsky's criticism of n-gram models (Chomsky, 1957) a lot of effort was spent on describing language as a set of formal rules. These rules were supposed to account for all language use.

In the late 1970s, statistical methods made a comeback. With more computing power and language resources available, researchers at IBM achieved a breakthrough in automatic speech recognition using probabilistic methods. Applying the noisy channel model from information theory to speech signals, recognition accuracy skyrocketed (Jelinek, 2009). The same techniques, Hidden Markov models and n-gram language models, were successfully adapted to part-of-speech tagging (DeRose, 1988; Church, 1988) and machine translation (Brown et al., 1990). The new methods proved very successful and the *statistical revolution* in NLP (Lee, 2004) was underway. Nowadays, NLP and statistical methods are inextricably linked. Standard approaches for part-of-speech and named-entity tagging, parsing, machine translation, relation extraction, sentiment analysis to name but a few are based on machine learning.

2.3.2 General concepts

This introduction is based on Friedman et al. (2001). Machine learning starts with inputs and outputs which are measurable observations in the real world.

Commonly, several input variables have influence on an output variable; we think of the former predicting the latter. For example, the prefixes and suffixes of a word predict its part of speech. The challenge of supervised machine learning is to discover rules, or a model, that – when presented with input variables, predict the correct output variables.

We denote input variables, or features, with X. We denote output variables with Y. We think of X as an input feature vector. We can access the components of X by subscripts X_i. Uppercase letters are used to address the input and output variables in general. For concrete observations lowercase letters x_i and y_i are used.

X is a p-dimensional vector. The number p is the number of features (parameters), of the model. If we use only the prefix and the suffix of a word to determine its part of speech, p is 2. In supervised learning, we have a set of training examples $\{(x_i, y_i)\}, i = 1, \ldots, N$ from which we learn the model. Afterwards, given a new input vector X, we predict the output Y.

If the output variable is categorical, i.e. a finite number of distinct classes, as in the part-of-speech example, the task is classification. The complementary task, where the output variable is continuous, is called regression. The application of machine learning in this thesis is MWT, which is a classification task. The classification method we use is logistic regression.

2.3.3 Logistic Regression

Logistic regression is a classification method. It models categorical outputs predicted by continuous input variables. In addition, logistic regression returns a probability for the predicted class. We will explain the ideas behind logistic regression for the case of a binary output variable.

Logistic regression models $P(Y|X)$. It is assumed that $P(Y|X)$ follows the form of the logistic function $Y = \frac{1}{1+\exp(-x)}$. For a Boolean classification problem with $Y = \{\textsc{true}, \textsc{false}\}$ the probabilities for each class are given by

$$P(Y = \textsc{true}|X) = \frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j)}$$

and

$$P(Y = \textsc{false}|X) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j)}$$

The probabilities sum to 1.

For the input X we are interested in the Y_i that maximizes $P(Y = Y_i|X)$. This means that if

$$1 < \frac{P(Y = \textsc{true}|X)}{P(Y = \textsc{false}|X)}$$

we assign label $Y = \textsc{true}$. If we substitute the class probabilities, we get

$$1 < \exp(\beta_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j)$$

Taking the natural logarithm on both sides, we can reduce the classification rule to a linear model.

$$0 < \beta_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

This linear equation cannot be fitted like a standard linear model because the necessary assumptions are not met. For logistic regression, an iterative method (e.g. Newton's method) has to be used for parameter estimation.

In NLP, multinomial logistic regression models are sometimes referred to as maximum entropy models (Manning and Klein, 2003).

		predicted class	
		TRUE	FALSE
real class	TRUE	TP (True Positive)	FN (False Negative)
	FALSE	FP (False Positive)	TN (True Negative)

Table 2.3: Confusion matrix of classification outcomes

2.3.4 Evaluation

The standard evaluation metrics for classification are precision and recall. These metrics are easily defined in terms of possible classification outcomes. The possible outcomes are summarized in Table 2.3. A true positive (TP) is a classification decision that assigns TRUE when the real class label is TRUE. If the label is TRUE and the classifier does not recognize it, we have a false negative (FN). The classifier predicting TRUE even though the label is FALSE is called a false positive (FP). Finally, a true negative occurs when the label is FALSE and the classifier predicts FALSE. We can now define precision and recall.

Precision is the ratio of true positives to all classification decisions.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall, on the other hand, is defined as the ratio of true positives of a label to all instances of that label.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision is a measure for how “noisy” a classifier is, i.e. the proportion of correct decisions when it claims to have found an instance of a particular class. Recall tells us how “thorough” a classifier is, i.e. how many of the instances of a class it identifies correctly. The two qualities of a classifier measured by precision and recall compete with each other. Increasing the one usually comes at the expense of decreasing the other. In practice, it is often the goal to identify which quality is more desirable and then finding a good trade-off.

A measure that evaluates the overall performance is accuracy. It measures the proportion of correct classification decisions.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

For our experiments, we use the accuracy measure as the primary evaluation metric.

2.3.5 Annotator Agreement

For supervised learning, we would like to assess how reliable the labels created by human annotators are. High reliability promises good models and low reliability (meaning that not even humans can solve a task consistently) could discourage the training of a model altogether. A common way in NLP to assess the reliability of human annotations is to compute inter-annotator agreement statistics (Artstein and Poesio, 2008). We discuss annotator agreement in the chapters presenting our experiments (Chapter 3 and Chapter 4, respectively).

2.3.6 Summary

In this chapter, we presented the principles of supervised machine learning. We covered logistic regression, a learning algorithm for classification. We explained how classifiers are commonly evaluated with precision, recall and accuracy. See Chapter 4 for how we incorporate the classifier into the MWU decision step.

Chapter 3

Asymmetric Association in Multi-Word Units

3.1 Introduction

In this chapter, we investigate asymmetric association in MWUs. The work presented here is based on joint work with Stefan Evert (Michelbacher et al., 2011a).

Apart from being relevant for MWT, asymmetric association is a linguistic and cognitive phenomenon worth studying in its own right. We will show that asymmetric association can be found in both human production and corpus data. We describe an elicitation study with human subjects that we conducted to measure asymmetry in syntagmatic word combinations. Additionally, we designed asymmetric association measures that capture asymmetry in corpus data. We show that these asymmetric measures can predict asymmetry in human production with high accuracy.

The remainder of this chapter is structured as follows. Section 3.2 places asymmetric association and the elicitation study into the broader context of word associations and research using elicited data. Section 3.3 introduces corpus-based measures of asymmetry based on classic association measures and describes asymmetry effects in corpus data. Section 3.4 contains the experimental design for measuring asymmetry in human associations. In

paradigmatic + symmetric	**paradigmatic + asymmetric**
good *bad*	*bird* *canary*
syntagmatic + symmetric	**syntagmatic + asymmetric**
epileptic seizure	*wishful thinking* (RP), *high fidelity* (LP)

Figure 3.1: Four types of word associations

Section 3.5, we analyze the results obtained for a sample of adjective-noun and noun-noun combinations. We show that the directedness of human association is accurately predicted by the corpus data and that asymmetric association can be regarded as a property of MWUs. In Section 3.6, we summarize our findings.

3.2 Background

3.2.1 Four Types of Association

We place the notion of asymmetry in syntagmatic word combinations into the broader context of different types of association between words. The question of how closely two words are related to each other or how strong the association between them is is relevant throughout NLP, for example for smoothing language models. Notions of relatedness and association are often defined as catch-all categories that mix together many different ways in which two words can be related (see Section 3.2.3). Rather than taking association as an atomic notion, we distinguish between four types of association that can be classified along the dimensions syntagmatic-paradigmatic and symmetric-asymmetric. Before we discuss the four types let us briefly recapitulate syntagmatic and paradigmatic relations.

The ideas of syntagmatic and paradigmatic relations between words have their origin in the work of Ferdinand de Saussure. Traditionally, the rela-

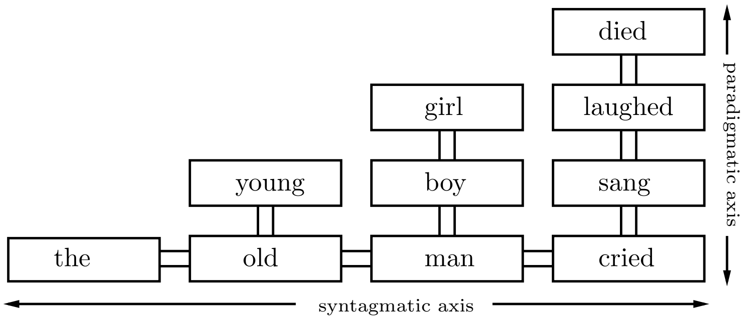


Figure 3.2: Syntagmatic and paradigmatic relations

tionship between two words is called syntagmatic if they occur in sequence:

> Combinations based on sequentiality may be called *syntagmas*. The syntagma invariably comprises two or more consecutive units [...]. In its place in a syntagma, any unit acquires its value simply in opposition to what precedes, or to what follows, or to both.
>
> (Saussure, 1966, 121)

We use the term syntagmatic in the sense of morphosyntactic relations, specifically noun-noun compounds and prenominal adjectives.

In contrast, paradigmatic relations are orthogonal to the sequential syntagmatic axis. Two words are said to be paradigmatically related if they can be substituted for each other. Such words usually have the same part of speech. Figure 3.2 illustrates the orthogonal relation of the syntagmatic and the paradigmatic axis.

Many prototypical paradigmatic association pairs like *good–bad* or *girl–boy* are *symmetric*, by which we mean that they prime each other with similar strength in free association experiments. We will see that apart from symmetric associations, there are paradigmatic as well as syntagmatic associations that are asymmetric; they consist of two elements where one strongly predicts the other, but not vice versa. We give examples for each of the four possible types of association in Figure 3.1.

The pairs *good–bad* and *epileptic–seizure* are both examples of symmetric combinations, the former being paradigmatic and the latter syntagmatic. Their elements prime each other with about the same strength. For example, in the USF word association norms (see Section 3.2.3), 75% of subjects give *good* as a response for *bad*, and 76% give *bad* as a response for *good*. In our elicitation experiment, we found that *epileptic–seizure* received a forward score of .541 and a backward score of .462 which supports our symmetry claim (see Section 3.5.1 for a definition of these scores).

Moving on, we now turn to asymmetry. Kjellmer (1991) investigated asymmetry in word combinations. He writes:

> A large part of our mental lexicon consists of combinations of words that customarily co-occur. The occurrence of one of the words in such a combination can be said to predict the occurrence of the other(s). (Kjellmer, 1991, 112)

The pairs *bird–canary* (paradigmatic), *high–fidelity* and *wishful–thinking* (both syntagmatic) are examples of asymmetric combinations. In the first pair, asymmetry is captured in the USF data set. 69% of subjects give *bird* as a response for *canary*, but only 6% give *canary* as a response for *bird*. In the context of syntagmatic combinations, asymmetry refers to the phenomenon that lexical association between two words can be stronger in one direction than in the other. With the inherent sequentiality of syntactic combinations, there are two sub types of asymmetric syntagmatic combinations which Kjellmer calls right-predictive (RP) and left-predictive (LP). In right-predictive combinations such as *wishful thinking*, *bonsai tree* or *wellington boots*, the first component suggests (or predicts) the second, but not the other way around. For left-predictive combinations, the opposite is the case: the second components of *high fidelity*, *deadly nightshade*, and *arms akimbo* suggest the first components, but not vice versa. We found evidence backing the asymmetry claim for both directions in our experiment. For example, *wishful–thinking* receives a forward score of .952 and a backward score of .006. Conversely, *high–fidelity* receives a forward score of 0.017 and a backward score of 0.696.

Sinclair (1991) introduced the notion of *upward* and *downward collocation*. In his terminology, a collocation – “the occurrence of two or more words within a short space of each other in a text” (Sinclair, 1991, 170) – consists of a *base* word and a *collocate*. In an upward collocation, the collocate is more frequent than the base; in a downward collocation, the collocate is less frequent than the base. Based on the assumption that *new* is more frequent than *tree*, which is in turn more frequent than *bonsai*, *new tree* is an instance of upward collocation and *bonsai tree* is an instance of downward collocation with base *tree* and collocates *new* and *bonsai*, respectively. In relation to Kjellmer’s notions, we expect stronger predictiveness from collocate to base in the case of downward collocation, and vice versa for upward collocation.

Kjellmer and Sinclair recognize the existence of asymmetry in syntagmatic association. However, they did not try to measure the effect systematically. This study picks up where Kjellmer and Sinclair left off by measuring asymmetry in both human production and corpus data. We demonstrate that, for adjective-noun and noun-noun combinations, these asymmetry effects are characteristic of human linguistic performance and can be accurately predicted from corpus data asymmetric association measures. Most studies in the literature are concerned with paradigmatic relations, either derived from free association norms or from large corpora using measures of statistical association and semantic relatedness. In contrast, we investigate the syntagmatic relation between words.

3.2.2 Motivation for Studying Asymmetry

Tversky (1977) argued that similarity is an asymmetric relation, criticizing the inherently symmetric aspect of metric-based models of similarity. He backed his view with a number of rating experiments in which subjects had to assess the similarity between different kinds of objects, for example figures, letters and countries. *North Korea*, for example, is judged more similar to *China* than vice versa. According to Tversky, the reason for this lies in the subjects’ feature representation of the two words. A large number of features are used to represent the concept *China*, only some of which are also included in the representation of *North Korea*. Conversely, a small number of features are used for *North Korea*, many of which are part of *China*'s representation.

Tversky showed that asymmetry in similarity is a cognitive phenomenon, but it can also be measured in corpus data. In the context of estimating co-occurrence probabilities for unseen events in language models, several measures of distributional similarity were discussed (Dagan et al., 1999). While most of the studied measures are symmetric, one asymmetric measure has received further attention: the alpha skew divergence s_α (Lee, 1999, 2001). It is a weighted version of the asymmetric Kullback-Leibler divergence (Kullback and Leibler, 1951). Lee (1999) mentions the subject of asymmetry in similarity, but does not investigate it further.

Weeds (2002) emphasizes the asymmetric aspect of the skew divergence and its potential usefulness in capturing asymmetry in similarity. She links asymmetric substitutability to the hypernymy relation and proposes that *fruit* and *apple* are similar to each other but *fruit* is more similar to *apple* than *apple* is to *fruit*. Applied to hypernymy, this would be reflected in $s_\alpha(hyper(x), x)$ being lower than $s_\alpha(x, hyper(x))$ (a lower score means higher similarity). In an initial experiment, Weeds was able to predict hypernyms and hyponyms in 156 preselected word pairs in over 90% of the cases using the above formula.

Kotlerman et al. (2010) have investigated asymmetric similarity measures.[1] Their focus is on lexical entailment, the task of deciding whether or not u entails v in an entailment pair. An entailment pair $u \rightarrow v$ is a relation between a word u and a word v so that the meaning of v follows from the meaning of u, for example, *bread* $\rightarrow$ *food*. The asymmetric element here is that $u \rightarrow v$ can be a valid entailment while $v \rightarrow u$ is not: *food* $\not\rightarrow$ *bread* (in this case v is a hypernym of u). Their goal is a directional distributional term-similarity measure. They formulate the properties of such a measure and present a distributional measure for lexical inference based on feature inclusion that meets these criteria. The underlying assumption is analog to Tversky's in that it is assumed that a term with a broader meaning contains the features of a term with a narrower meaning, but not vice versa, and

[1] They refer to their measures as directional rather than asymmetric.

that the narrower term entails the broader, but not vice versa. For the proposed measure, feature vectors contain grammatical dependency relations. The measure is evaluated directly on entailment pairs and in two tasks and it outperforms symmetric measures. The main difference between our asymmetric measure targeted at general word associations and that of Kotlerman et al. is that their goal is a measure for finding relations between narrower and broader terms.

In general, asymmetric measures of similarity are potentially an important factor in all NLP tasks that benefit from better treatment of mutual substitutability, for example reducing data sparseness in language models (Dagan et al., 1999) or the automatic acquisition of selectional preferences (Resnik, 1996).

For example, query expansion, a popular application of association measures in natural language processing, is an asymmetric task. It is appropriate to rewrite the query *fruit* as *fruit OR apple* since documents about apples are necessarily about fruit, but it is not appropriate to rewrite the query *apple* as *apple OR fruit.* Clearly, corpus-based measures of association are only useful in this context if they take such asymmetry into account.

3.2.3 Research with Elicited Data

This section contains a brief description of available data sets of human word associations. Two kinds of experiments are prevalent: free association and rating experiments. We describe the setups of the experiments, the characteristics of the resulting data sets and why existing data sets are not suitable for our experiment.

First, we describe free association experiments and the corresponding data sets which are called (free) association norms. Free association experiments are the standard way of eliciting word associations from human subjects.

Free association norms The first two norms we look at are known as *The Minnesota norms*, and *Palermo and Jenkins* norms. Both norms are closely related. The Minnesota norms were collected by presenting 100 stimulus words[2] to 1,008 college students of introductory psychology classes in 1952 (Wettler and Rapp, 1993). The well-known Palermo and Jenkins data set was presented in *Word association norms: Grade school through college* (Palermo and Jenkins, 1964). It is an extension of the previous experiment including students of different age groups. In addition to the 100 original words, another 100 words more suitable for young speakers were added. A variety of parts-of-speech including nouns, adjectives, verbs, adverbs and prepositions were used. In both studies, each stimulus word was presented with a blank line to the right of it and subjects were asked to write what first came to their mind on the line. 1,000 subjects ranging from 4th graders to undergraduate students took part in the study.

The *University of South Florida Word Association Rhyme and Word Fragment Norms* (USF) is a collection of word associations compiled by Nelson et al. at the University of South Florida. Data collection started in 1973 and went on for two decades. More stimulus words were added over the course of time. The finished data set was published in 1998. On average, each stimulus word was presented to around 150 subjects and each subject had to complete a booklet of 100 to 200 words. In total, the database contains 5,019 stimulus words. The elicitation procedure was almost identical to the one used by Palermo and Jenkins (1964). More than 6,000 participants produced nearly 750,000 responses.[3]

The *Edinburgh Word Association Thesaurus* (EAT) was created by Kiss et al. (1973). It contains 8,400 stimulus words including the stimuli used by Palermo and Jenkins (1964). Each stimulus was presented to 100 different subjects. The elicitation procedure was, again, very similar to Palermo and Jenkins, namely, that subjects were presented a list of stimuli without context and were asked to write down the first word they could think of. Subjects were urged to complete the task as quickly as possible.[4]

In psycholinguistics, researchers have been studying association norms for

[2] The same stimuli were previously used by Kent and Rosanoff (1910).

[3] The full database with detailed information about every stimulus-response pair is available for download at `http://web.usf.edu/FreeAssociation/`.

[4] An interactive version of the data is available online at `http://www.eat.rl.ac.uk/`.

over a century to explore the organization of the mental lexicon and how information is retrieved from it during language production and comprehension (e.g. Clark, 1971). We refer the reader to Mollin (2009) for a more detailed discussion of word association norms in psycholinguistics. Association norms have also been used as benchmarks for models of human semantic knowledge (Griffiths et al., 2007).

In the field of computational linguistics, association norms received increased attention when Church and Hanks (1990) observed that corpus-based statistical association measures can be used to model human responses given in free association and rating tasks. A number of studies have confirmed that human associations can be predicted from corpora (e.g. Spence and Owens, 1990; Rapp, 2002; Sahlgren, 2006; Michelbacher et al., 2007). However, the data sets utilized in these studies (for example, USF or EAT – see Section 3.2.3) do not distinguish between different types of semantic relatedness or association and they only contain a small portion of syntagmatic combinations. Table 3.1 gives examples of the various relationships between stimulus and response that occur in these data sets. The table was compiled by Hutchison (2003) who classified each stimulus and response pair of Palermo and Jenkins' norms. Almost all relations are paradigmatic, but three comprise syntagmatic pairs: on the one hand, the groups that Hutchison called *forward* and *backward phrasal associates* and that we refer to as syntagmatic combinations in our terminology. On the other hand, the group labeled *associated properties* can also be thought of as syntagmatic, for example in adjective coordinations (*a deep, dark hole*). In total, only 16.7% of the pairs were classified into these relations.[5] Washtell and Markert (2009) report higher numbers of syntagmatic relations in free associations. For two data sets, Kent and Rosanoff (1910) and Russell and Jenkins (Jenkins, 1970), they found 27% and 39%, respectively. Apart from the fact that they used different data sets than Hutchison, a likely cause for the higher number lies in Washtell and Markert's definition of syntagmatic. It is more lax than ours covering meronymy, holonymy and other "harder-to-classify topical or idiomatic relationships (*family–Christmas*, *rock–roll*)" (Washtell and Markert,

[5] Since many pairs fall into several categories, the total percentage exceeds 100%.

Association Type (and Example)	Percentage Rate
Synonyms (*afraid–scared*)	14.1
Antonyms (*day–night*)	24.3
Natural category (*sheep–goat*)	9.1
Artificial category (*table–chair*)	5.1
Perceptual only (*pizza–saucer*)	0.0
Supraordinate (*dog–animal*)	5.6
Perceptual property (*canary–yellow*)	11.1
Functional property (*broom–sweep*)	12.1
Script relation (*orchard–apple*)	6.1
Instrument (*broom–floor*)	6.1
Forward phrasal associate (*baby–boy*)	11.6
Backward phrasal associate (*boy–baby*)	4.1
Associated properties (*deep–dark*)	1.0
Unclassified (*mouse–cheese*)	5.1

Table 3.1: Common relationships between stimulus and response words in Palermo and Jenkins' association norms (classification by Hutchison (2003))

2009, 1).

Rating experiments Rating experiments use a different setup for collecting human association judgements. In those experiments (e.g. Rubenstein and Goodenough, 1965), the subjects are presented two words simultaneously. The task is two rate the relatedness between the two words with a number from a fixed scale. The resulting data set is a set of word pairs each associated with a corresponding relatedness score. There has been a large body of work on evaluating corpus-derived measures of semantic relatedness (including Miller and Charles (1991); Resnik (1996); Finkelstein et al. (2002); Gurevych (2005); Budanitsky and Hirst (2006); Strube and Ponzetto (2006); Gabrilovich and Markovitch (2007); as well as Lapata et al. (2001) and Keller and Lapata (2003) for syntagmatic combinations). These studies often use the data set by Rubenstein and Goodenough or similar data. Their methodology obscures any possible asymmetry effect because both words are presented to the user simultaneously.

In cognitive linguistics, there is a general consensus about the necessity to support hypotheses about linguistic phenomena and theories with usage-based evidence. It remains unclear, however, which methodological approach to corpus data is most suitable for obtaining such evidence, and whether different techniques are needed for different phenomena and hypotheses. Some recent studies use data elicited in psycholinguistic experiments in order to evaluate different methods of analyzing corpus data. It has been found, for example, that aspects of human language processing can be modeled with association measures and that different association measures vary in their ability to predict human intuitions (e.g. Wiechmann, 2008; Gries et al., 2005).

We see our study as a further step in this direction. In accordance with the approaches sketched above, we use a number of statistical measures and compare their predictions with data obtained from human subjects. However, we move our focus to a phenomenon that has not been considered in previous studies, namely the asymmetry of word associations.

We have established that word association norms such as the USF or EAT and rating experiments are not suitable for the purpose of studying syntagmatic asymmetry because (i) only a low number of stimulus-response pairs are suitable for such an investigation; (ii) they mix together different types of associations and (iii) they do not measure asymmetry.

3.3 Asymmetric Association Measures

This section describes the measures we designed to detect left and right-predictive combinations. The measures are based on classic association measures (see Section 2.1).

3.3.1 Corpus Data

For the investigation, we focus on two-word adjective-noun and noun-noun combinations. These combinations occur in uninterrupted sequence, which makes them suitable for elicitation experiments common noun followed by common noun:

```
[NN] [NN]
```

adjective or common noun followed by a proper noun:

```
([ADJ] ([COMMA | CONJ | ADJ | ADV]* [ADJ])? | [NN]) [NP]
```

adjective or proper noun followed by a common noun:

```
([ADJ] ([COMMA | CONJ | ADJ | ADV]* [ADJ])? | [NP]) [NN]
```

Figure 3.3: Part-of-speech patterns that capture the co-occurrence relation of adjective-noun and noun-noun modification. Square brackets represent tokens, e.g. `[NN]` is a noun token. Parentheses are used for grouping. A '?' marks optional matches, '∗' indicates zero or more matches and '|' is used for disjunction.

The corpus associations used in this work are based on data extracted from the British National Corpus (BNC) (Aston and Burnard, 1998). The corpus consists of British English language samples from a variety of sources (mostly written English with a small proportion of transcribed conversations). The sample size is about 100M words. Our co-occurrence relation is adjective-noun and noun-noun modification. Proper nouns were only allowed in conjunction with a preceding adjective or common noun. We did not extract pairs consisting of two proper nouns because they introduced too much noise (e.g. personal names or geographical entities). We use the patterns shown in Figure 3.3 to capture the co-occurrence relation.

The noun-noun pattern for compounds is straightforward. The adjective-noun patterns are slightly more complex because they are designed to match adjacent as well as more distant adjective-noun modification. In addition, they allow proper nouns as modifiers to account for combinations such as *Wellington boots*.

Figure 3.4 shows the noun-noun pair *soccer team* and two adjective-noun pairs, *serious fever* and *fatal fever* that were extracted. Note that in this

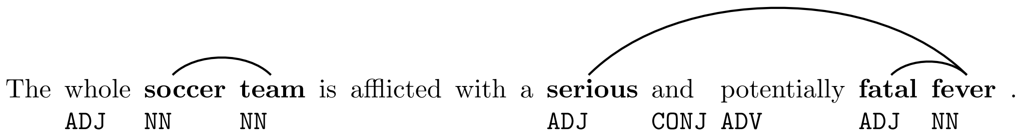

The whole **soccer team** is afflicted with a **serious** and potentially **fatal fever** .
ADJ NN NN ADJ CONJ ADV ADJ NN

Figure 3.4: The bold phrases are three examples of adjective-noun-noun and noun-noun pairs that match the co-occurrence relation. Only relevant POS tags are shown.

example, the false positive match *whole soccer* would have been extracted.

We extracted 2,014,116 word pairs (types). In order to remove noise, we applied a frequency filter of $f \geq 3$. For the remaining 391,454 pairs, we calculated association scores for the measures t-score, log-likelihood, chi-square and frequency. In addition, we computed the conditional probabilities $p(w_1|w_2)$ and $p(w_2|w_1)$ for each word pair.

3.3.2 Asymmetric Measures

We need measures that can capture asymmetry in lexical association. The association measures defined in Section 2.1 are symmetric in the sense that they do not capture the left-predictiveness or right-predictiveness that Kjellmer observed in many word combinations. All measures are invariant under transposition of the contingency table, i.e. the association score remains the same if the rows and columns are exchanged. In preliminary experiments we have shown the principal feasibility of corpus-based asymmetric association measures (Michelbacher et al., 2007). To capture paradigmatic asymmetric relations of the *apple-fruit* kind, we defined asymmetric rank measures based on the chi-square association measure and conditional probabilities. We gathered asymmetric association data from the BNC and evaluated the results against data computed from the USF Free Association Norms. The measures were able to predict asymmetry in associations; however, the error rate was relatively high.

We now generalize the notion of a rank measure to arbitrary symmet-

w_1	w_2	t-score		w_1	w_2	$R_t^{\rightarrow}(w_1, w_2)$
rich	**man**	**16.563**		**rich**	**man**	**1**
rich	peasant	12.919		rich	peasant	2
rich	country	12.756		rich	country	3
rich	people	8.386		rich	people	4
rich	variety	7.423	→	rich	variety	5
rich	source	7.861		rich	source	6
rich	color	7.568		rich	color	7
rich	soil	6.018		rich	soil	8
rich	nation	6.766		rich	nation	9
rich	world	5.714		rich	world	10

Table 3.2: Determining the forward rank of *rich man*

ric association measures and evaluate the ability of these rank measures to capture the asymmetry of syntagmatic associations.

In order to transform a standard symmetric association measure into a rank measure that computes separate scores for the left- and right-predictiveness of a word pair, we implement the following procedure. For example, for a left-to-right rank measure based on t-score:

1. Compute symmetric t-scores for all word pairs (w_1, w_2).
2. For each word w_1, create an *association list* of all components w_2 that co-occur with w_1 in the corpus and sort the list by association strength in descending order.
3. Starting at the top, replace the association scores by ranks $1, 2, 3, \ldots$[6]

Right-to-left rank measures are computed accordingly, exchanging w_1 and w_2 in the ranking procedure.

Table 3.2 shows the ten nouns (w_2) that are most strongly associated with the adjective *rich* (w_1), together with the association scores computed by the

[6] Ties are handled as in a typical “sports” ranking: if n consecutive items have the same score, they are all assigned the lowest free rank r; the next item will be assigned rank $r + n$.

w_1	w_2	t-score		w_1	w_2	$R_t^{\leftarrow}(w_1, w_2)$
young	man	62.492		young	man	1
old	man	51.602		old	man	2
tall	man	19.270		tall	man	3
dead	man	18.661		dead	man	4
rich	**man**	**16.563**	→	**rich**	**man**	**5**
poor	man	15.986		poor	man	6
white	man	14.279		white	man	7
married	man	14.620		married	man	8
gay	man	14.487		gay	man	9
big	man	14.456		big	man	10

Table 3.3: Determining the backward rank of *rich man*

t-score measure (t). We write the left-to-right rank measure based on the t-score measure as $R_t^{\rightarrow}(w_1, w_2)$ and call it the *forward rank* of (w_1, w_2). Note that a *small* forward rank indicates a *high* degree of right-predictiveness. For example, $R_t^{\rightarrow}(\text{rich}, \text{man}) = 1$ means that *man* is the noun most strongly predicted by the adjective *rich* according to the t-score measure.

Analogously, we denote the *backward rank* of a word pair (w_1, w_2) according to the t-score measure by $R_t^{\leftarrow}(w_1, w_2)$. As can be seen from Table 3.3, the backward rank of *rich man* is $R_t^{\leftarrow}(\text{rich}, \text{man}) = 5$. In this case, the forward rank (1) is lower than the backward rank (5), indicating higher right-predictiveness than left-predictiveness.

Note that the association score of the pair *rich man* is 16.536 in both association lists. This score was computed from a single contingency table of observed frequencies, which is all the information that a standard association measure has access to. By contrast, the corresponding left-to-right rank measure $R_t^{\rightarrow}$ looks at the distribution of the association scores for all word pairs $(w_1, \cdot)$; and the right-to-left measure $R_t^{\leftarrow}$ looks at the distribution for all word pairs $(\cdot, w_2)$. This way, we can calculate different degrees of right- and left-predictiveness.

Rank measures are a general and flexible tool for capturing asymmetry effects in word combinations. They can be applied to any symmetric asso-

abbr.	base AM
R_f	based on frequency
R_{G^2}	based on log-likelihood (G^2)
R_t	based on t-score
R_{χ^2}	based on chi-square (χ^2)

Table 3.4: Abbreviations of rank measures

w_1	w_2	$R_f^{\rightarrow}$	$R_f^{\leftarrow}$	$R_{G^2}^{\rightarrow}$	$R_{G^2}^{\leftarrow}$	$R_t^{\rightarrow}$	$R_t^{\leftarrow}$	$R_{\chi^2}^{\rightarrow}$	$R_{\chi^2}^{\leftarrow}$
heavy	*smoker*	17	1	9	1	15	1	5	4
bonsai	*tree*	1	64	1	37	1	53	1	25

Table 3.5: Different association measures give rise to different rank measures.

ciation measure and transform it into an asymmetric measure of right- and left-predictiveness. Each association measure gives rise to a different asymmetric rank measure. The four different rank measures that we used in our experiments are defined in Table 3.4. Table 3.5 illustrates the differences between rank measures by showing left-to-right and right-to-left rank scores for the word pairs *heavy smoker* and *bonsai tree*, according to four different rank measures based on the standard association measures introduced in Section 2.1.

According to the first three rank measures, *heavy smoker* is a strongly left-predictive combination. The backward rank is 1 in all three cases whereas the forward rank is considerably higher. The rank measure based on χ^2 does not agree with the other measures, suggesting an almost symmetric pair with equal right- and left-predictiveness (although the backward rank is still slightly lower). The pair *bonsai tree* is strongly right-predictive according to all four measures. The forward and backward ranks are in accordance with the assessment of Kjellmer who used *bonsai tree* as an example for a clearly right-predictive combination.

The ranks do not take the frequency of the words into account and are therefore independent of association strength. For our purpose, this is not a problem. First, we want to examine asymmetry for each word pair individually without comparing ranks of different pairs. Second, in an elicitation experiment, low-frequency words will still trigger responses and the best responses will receive low ranks. The ranks tell us how good the associations are *relative to the stimulus.*

In accordance with Michelbacher et al. (2007), we also measure right- and left-predictiveness with conditional probabilities.

$$P^{\rightarrow}(w_2|w_1) = \frac{P(w_1, w_2)}{P(w_1)} \qquad P^{\leftarrow}(w_1|w_2) = \frac{P(w_1, w_2)}{P(w_2)}$$

We added arrows to emphasize right-predictiveness ($P^{\rightarrow}$) and left-predictiveness ($P^{\leftarrow}$). For example, $P^{\leftarrow}(w_1|w_2)$ denotes the probability that w_1 appears as the first component in a pair when $(\_, w_2)$ is already given. The probabilities are maximum-likelihood estimates.

Note that because

$$\frac{P(w_2, w_1)}{P(w_1)} = \frac{\frac{O_{11}}{N}}{\frac{O_{11}+O_{12}}{N}} = \frac{O_{11}}{O_{11} + O_{12}},$$

the rank measure based on conditional probabilities is identical to R_f. It is therefore not included separately in our evaluation. $P^{\rightarrow}(w_2|w_1)$ and $P^{\rightarrow}(w_1|w_2)$ are effectively forward and backward 2-gram language models.

3.3.3 Analysis of the Distribution of Ranks

Corpus-based measures of asymmetry are only interesting if such asymmetry is a frequent phenomenon. As we have argued earlier, we expect that syntagmatic associations are often asymmetric and can only be characterized adequately by a measure that allows for large differences in ranks. In order to explore this property of rank measures, we cross-tabulated the forward and backward ranks for the 391,454 word pairs with $f \geq 3$ extracted from the BNC . Rank values were collected into logarithmically scaled bins (ranks 1–2, 3–5, 6–10, 11–20, 21–35, 36–60, 61–100, 101–160, 161–250, 251–500, 501+), such that all bins contain a similar number of items.

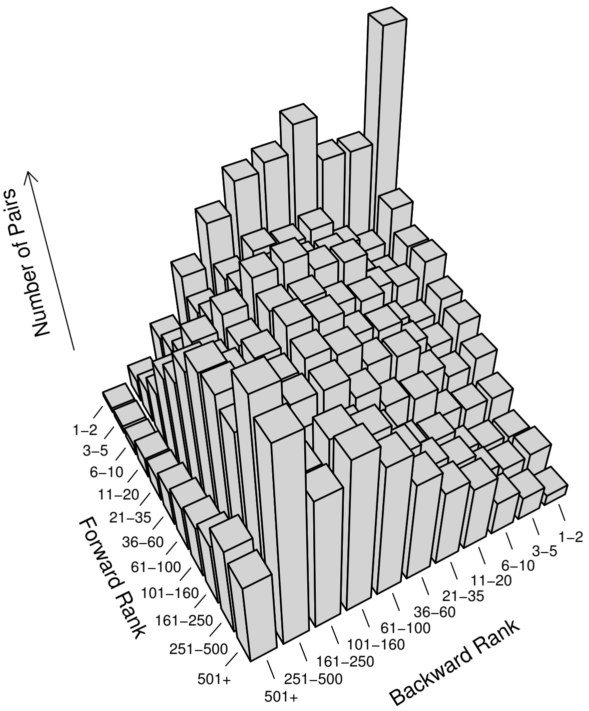


Figure 3.5: Cross-tabulation of forward and backward ranks for the log-likelihood measure.

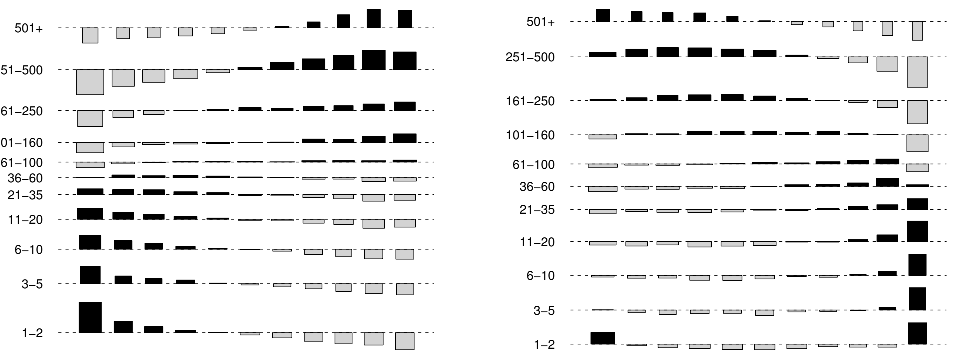


Figure 3.6: Association plots of forward and backward ranks for the log-likelihood-based rank measure R_{G^2} (left panel) and the frequency-based rank measure R_f (right panel).

Figure 3.5 shows a bar plot of the cross-tabulation of forward and backward ranks obtained for R_{G^2}, the rank measure based on log-likelihood. Bars along the main diagonal of the histogram – running from bottom to top in the printout – correspond to *symmetric* word pairs with nearly equal forward and backward ranks. The greater the distance of a bar from this main diagonal, the more asymmetric the corresponding word pairs are.

It is obvious that there is a considerable number of asymmetric word pairs with low forward and high backward rank (bars along the back left of the plot), and also vice versa (bars along the back right). On the other hand, very low forward ranks (ranks 1–2) correlate strongly with very low backward ranks, and very high forward ranks correlate with very high backward ranks (tall bars at both ends of the main diagonal). This is hardly surprising, since forward and backward ranks are based on the same symmetric association score: a highly associated word pair is more likely to achieve a low rank both in the "forward" and the "backward" list. Likewise, a word pair where w_1 and w_2 are close to statistical independence is more likely to be assigned high ranks in both lists.

Interestingly, the plot shows many word pairs with forward ranks 1 or 2, but much higher backward rank (roughly between 10 and 100, along the back left side of the histogram). According to R_{G^2}, these word pairs are strongly right-predictive. In comparison, the number of strongly left-predictive pairs is much smaller – there are no equally high bars along the back right side of the chart (corresponding to backward ranks of 1 or 2 and forward ranks between 10 and 100). This suggests that right-predictiveness is more common in English than left-predictiveness, at least for adjective-noun and noun-noun combinations. This observation is supported by our elicitation experiments, in which more word pairs were found to be right-predictive than left-predictive (see Section 3.5). The prevalence of right-predictive combinations is probably related to the fact that the preceding word is an important factor when deciding which word to produce next. This causal relationship along the time axis promotes the formation of right-predictive combinations. There exists no equally strong mechanism for producing left-predictive combinations.

The association plot in the left panel of Figure 3.6 shows more clearly to what extent forward and backward ranks are correlated. Black bars above the midlines indicate that a given combination of forward and backward rank appears for more word pairs than expected if the rankings were independent (i.e., a positive correlation between forward and backward rank). Grey bars below the lines indicate a smaller number of word pairs than expected (i.e., a negative correlation). It is obvious from the plot that very low forward ranks correlate strongly with very low backward ranks, and similarly for very high ranks. Again, this shows that very strongly and very weakly associated word pairs tend to be symmetric according to the rank measure. By contrast, the almost vanishing bars near the center of the plot show that forward and backward ranks are practically independent in a middle range (roughly ranks 10–100). Here, the rank measure is able to make a distinction between symmetric and asymmetric pairs.

A second important question is whether different symmetric association measures lead to different rank distributions. The right panel of Figure 3.6 shows an association plot for R_f (the rank measure based on co-occurrence frequency). The distribution of ranks is strikingly different from that of R_{G^2}, with forward and backward ranks almost independent for ranks below about 250. There is a considerable number of highly asymmetric word pairs, characterized by a very high rank (above 500) in one direction and a low rank (below 35) in the other direction (black bars along the top and right edges of the plot). This observation may be surprising at first, but it is easily explained for the left-predictive case by combinations of a high-frequency word w_2 (e.g. *disease*) with a low-frequency word w_1 (e.g. *adiposogenital*) that almost always occurs with w_2 (and analogously for the right-predictive case).

Association plots for the other two measures are qualitatively similar to the log-likelihood (G^2) pattern, with a somewhat stronger correlation for chi-square (χ^2) and a slightly larger region of near-independence for t-score (t). This is perhaps not surprising since all three measures are based on statistical hypothesis tests. The observed differences between the rank distributions agree with the known tendencies of χ^2 to overestimate and of t to underestimate the significance of association (Evert, 2004, 111).

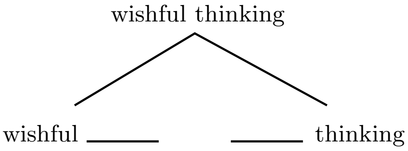


Figure 3.7: Forward and backward stimuli for the pair *wishful thinking*

3.4 Elicitation Experiment

This section describes our experimental design and the results of the elicitation experiment.

Free association experiments have frequently been conducted to gather data about spontaneous human associations. In these experiments, a stimulus is presented and the subject is asked to produce one or more related words, e.g. those words that first come to mind when thinking about the stimulus.

In this type of experiment, there are no restrictions on what type of response the subject can give (cf. Table 3.1). When humans associate freely, they produce mostly paradigmatic combinations. While there are some syntagmatic associations in the norms produced from such experiments – e.g. *blue* → *sky* or *big* → *deal* – they are always right-predictive, making these norms unsuitable for our purpose.

Nevertheless, as noted in Section 3.2.3, word association norms do contain a portion of syntagmatic responses. Furthermore, it has been shown that grammatical stimulus-response pairs can be collected systematically in elicitation experiments when subjects are explicitly asked to produce them (McGee, 2009). With these findings in mind, we decided to base our experimental design on classical free associations experiments but with a restriction to syntagmatic responses.

We instructed subjects to produce responses that result in a well-formed phrase when combined with the stimulus.[7] This way, stimulus and response form a combination that could occur in production. The key problem is how to present stimuli in a way that elicits the desired data without biasing the subjects' responses.

The experimental design we decided on splits each word pair (w_1, w_2) into two separate stimuli: a *forward stimulus* $[w_1 \_]$ and a *backward stimulus* $[\_ \, w_2]$. That is, either the first or the second component of the pair is replaced by the blank _ to indicate to participants that a word has been removed and needs to be provided. This design allows for testing both directions of association, from w_1 to w_2 and from w_2 to w_1. An example is shown in Figure 3.7.

Subjects were instructed to fill the blank in a way that created a well-formed phrase. We imposed no other restrictions on admissible responses to avoid any type of bias. In particular, no context was provided that might have disambiguated ambiguous stimuli or suggested a response from a particular domain.

Because of this unrestricted nature of the experiment, subjects often produced part-of-speech combinations that were not compatible with our data. Such unusable responses included determiners, pronouns and cases where subjects interpreted a stimulus word as a verb or adverb rather than as an adjective or noun. For example, [*cut* _] was often extended to *cut down* or *cut off* instead of a noun-noun or adjective-noun phrase such as *cut glass*. We discarded part-of-speech mismatches in order to be able to perform a clean analysis of adjective-noun and noun-noun associations.

3.4.1 Pair Selection

We used a hybrid selection method to sample stimuli, adapting the methodology of Krenn and Evert (2005). We started with a pool P of candidates and took a random sample M from P. We then created a subset Q from M by removing extraction noise and domain-specific terminology. Finally, we took a further random sample S from Q to get the desired number of stimuli.

This procedure was applied to three different pools:

- P_1: all pair types (w_1, w_2)

[7] The complete instructions are shown in Appendix B.

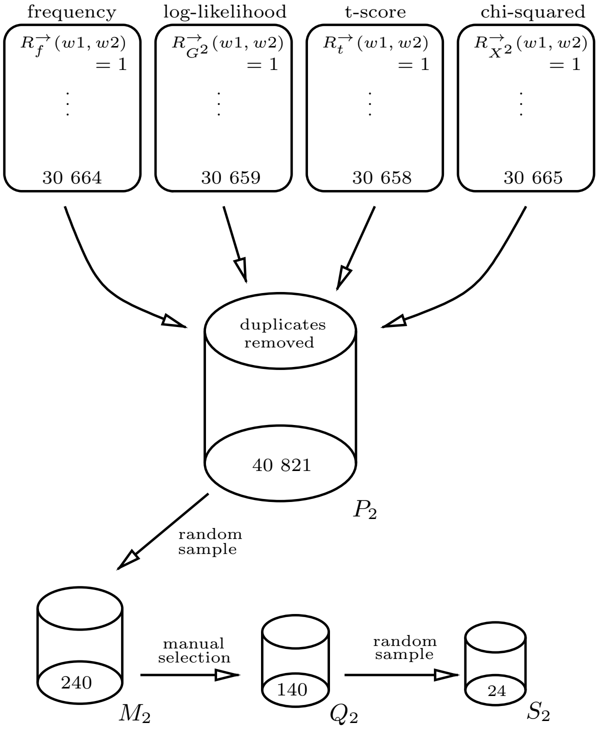


Figure 3.8: Sampling procedure for stimuli (illustrated for a sample S_2 of right-predictive word pairs from the pool P_2)

- P_2: pair types with strong right-predictiveness (according to at least one of the association measures)
- P_3: pair types with strong left-predictiveness (according to at least one of the association measures)

The process is illustrated in Figure 3.8 for strongly right-predictive word pairs (i.e., candidate sets $P_2/M_2/Q_2/S_2$).

The first pool, P_1, contains all 2,014,116 pair types that we extracted from the BNC.[8] The pools P_2 and P_3 are motivated by two constraints that any experiment designed to elicit syntagmatic responses must satisfy. First, we can only present a limited number of stimuli to each subject. This means the

[8] Note that no frequency threshold is applied at this stage, resulting in a very large number of pair types.

overall number of stimuli must be relatively small. Second, we must ensure that the elicited data are useful for our evaluation. Since a random sample would mostly contain weakly associated pairs, it was necessary to bias the selection of stimuli.

To this end, we created a set P_2 of strongly right-predictive pairs and a set P_3 of strongly left-predictive pairs, where strong predictiveness was defined as $R^{\rightarrow}(w_1, w_2) = 1$ (for P_2) and $R^{\leftarrow}(w_1, w_2) = 1$ (for P_3) for at least one measure. For example, for 30,664 pairs $R_f^{\rightarrow}$ is 1. We chose this criterion to obtain good candidates for asymmetric word combinations.

To create P_2, the rank criterion $R^{\rightarrow}(w_1, w_2) = 1$ was applied to all pairs and for all four association measures. The four resulting sets (represented by rectangles in Figure 3.8) were merged. Multiple occurrences of the same pair were removed, resulting in a pool P_2 of 40,821 candidates. The same procedure was carried out to obtain P_3 with a size of 26,600 candidates.

In the next step, random samples of fixed sizes were drawn from each of the P_i. The three resulting samples were Q_1 (336 pairs), Q_2 (240 pairs) and Q_3 (240 pairs). The sample sizes were chosen to be low enough to allow for manual review of all pairs. Pairs with frequency $f \leq 5$ were removed. We then reviewed each of the remaining pairs and removed extraction noise, rare technical terms and rare proper nouns. Specifically, domain-specific terminology from fields like mathematics, biology, computing, and medicine were removed. Examples include *ileocolonic resection*, *configurational entropy*, *Unix file* and *non-zero element* – terms which are unlikely to be familiar to the general population.

The resulting pools Q_1, Q_2 and Q_3 were the basis for three final random samples: S_1 (16 pairs from Q_1), S_2 (24 pairs from Q_2) and S_3 (24 pairs from Q_3). These 64 pairs were then replaced by their most frequent surface realizations in the BNC, in order to ensure that subjects would not be distracted by the use of uncommon base forms from the automatic lemmatization. For example, *wellington boot* was turned into *wellington boots* and *christmas decoration* into *Christmas decorations*.

3.4.2 Conducting the Experiment

Test subjects were randomly split into two groups, group I and group II. When group I was presented a pair with the first component missing, group II saw the same pair with the second component missing and vice versa. This procedure ensured that subjects were not biased by a previous stimulus (e.g. seeing [_ *tree*] after [*bonsai* _]). Stimuli of types [w_1 _] and [_ w_2] were split equally between the two groups.

Subjects were given detailed instructions to ensure they would not mistake the experiment for a free association task. They were encouraged to take some time to think of the stimulus word in different contexts and scenarios. They were also permitted to give multiple answers, or no answer at all.

The experiment was carried out online at the Portal for Psychological Experiments on Language.[9] The subjects were informed that only native speakers of English were allowed to participate. 168 subjects took part in the experiment, 74 for group I and 94 for group II. The discrepancy between the two groups is due to the fact that some subjects did not complete the experiment. We collected a total of 43,101 responses. This means that on average, a subject provided 4 responses per stimulus.

We only included data from completed experiments in our analysis. We removed 3 pairs – *common destiny*, *independent charts*, and *old self* – because they were never successfully elicited in either direction. For example, *common* was never elicited as w_1 for the stimulus [_ *destiny*] and *destiny* was never elicited as w_2 for the stimulus [*common* _]. The analysis described in the next section was performed for the remaining 61 pairs.

We lemmatized the subjects' input for our analysis. Spelling variants were unified to British English to facilitate the comparison with the corpus data. Manual spelling correction and normalization was applied when necessary, e.g. *Xmas* was normalized to *Christmas*.

For each subject and stimulus, we kept track of the order in which responses were given. We assume that the order of elicitation directly corresponds to association strength in that the first answer given has the highest

[9] `http://language-experiments.org/`.

association to the stimulus word and so on.

In elicitation experiments like free association norms or our experiment, agreement is usually not reported. Standard agreement measures such as the ones discussed in Artstein and Poesio (2008) are not applicable here because there is no predefined set of possible answers.

3.5 Experimental Results and Analysis

In this section, we define the direction scores used to evaluate subjects' responses and perform both a qualitative and a quantitative evaluation of the experimental results.

3.5.1 Direction Scores

We scored the subjects' responses using a mean reciprocal rank measure (cf. Voorhees, 1999). Two *direction scores* were defined – a *forward score* $f(w_1, w_2)$ and a *backward score* $b(w_1, w_2)$, as given by the following equations:

$$f(w_1, w_2) = \frac{1}{C([w_1\ \_])} \sum_{i=1}^{C([w_1\ \_])} \frac{1}{r_i(w_2)}$$

$$b(w_1, w_2) = \frac{1}{C([\_\ w_2])} \sum_{i=1}^{C([\_\ w_2])} \frac{1}{r_i(w_1)}$$

Here, $C([w_1\ \_])$ is the total number of subjects that were presented with stimulus $[w_1\ \_]$ and $r_i(w_2)$ is the rank of w_2 in the list of responses to $[w_1\ \_]$ given by subject i. $C([\_\ w_2])$ is the number of subjects presented with stimulus $[\_\ w_2]$ and $r_i(w_1)$ the rank of w_1 in the list of responses to $[\_\ w_2]$ by subject i. If a subject did not produce the response in question, we assigned rank $r = 1000$. The highest possible direction score in this scheme is 1.0.

3.5.2 Qualitative Evaluation

The composition of responses given in our study differs considerably from previous association norms. As an example, Table 3.6 shows the 10 highest-

	syntagmatic	paradigmatic		syntagmatic	paradigmatic	
stimulus	[*white* _]	*white*		[*water* _]	*water*	
No.		EAT	USF		EAT	USF
1	wash	black	black	bottle	wet	drink
2	Christmas	red	pure	works	drink	cool
3	house	snow	clean	fall	tap	wet
4	out	sheet	snow	fountain	sea	swim
5	board	ice	light	slide	cold	thirsty
7	wedding	beach	color	cooler	h2o	faucet
6	water	nothing	paper	jug	hot	pool
8	dress	blank	red	pipe	rain	thirst
9	man	block	—	park	river	ice
10	noise	blue	—	balloon	thirst	cold

Table 3.6: Comparison of our results with free association norms

scoring responses for *white* and *water* in our syntagmatic experiment and in two free association experiments.[10] For this comparison we did not filter out responses like *white out* that do not constitute adjective-noun or noun-noun combinations. For the stimulus *white*, half of the responses in the EAT norms are of a paradigmatic nature except for *sheet*, *beach*, *block*, *snow* and *ice*. The same holds for the USF norms; syntagmatic responses are *snow*, *light*, *color* and *paper*. For the stimulus *water*, there are no syntagmatic responses at all in either of the data sets.

The responses in the new experiment, however, exclusively consist of syntagmatic associations, that is, they all produce well-formed phrases when the response is inserted into the empty slot of the stimulus.

A qualitative analysis of the 61 pairs revealed four major groups. Group A contains all pairs where the rank measures conform with human responses in that they agree on which direction of association is stronger. The bulk of the pairs (48) belong to group A. Group B is a small group consisting of 4 cases where corpus data and human elicitations contradict each other. We also found borderline cases where the rank measures provide evidence for

[10] The USF data set lists only eight responses for the stimulus *white*.

both right-predictiveness and left-predictiveness, but could not be aligned with human elicitations (group C, 4 elements). We regard pairs where the rank measures suggest very strong association (rank ≤ 2) in both directions as a special case and put these pairs in a separate group of high mutual predictiveness (group D, 5 elements).

Table 3.7 shows word pairs from the four groups with detailed information on corpus ranks and scores from the elicitation experiment.

For most pairs in groups A and B, the four measures agree on the direction of predictiveness. There were 15 pairs where the four measures did not agree, marked with '*' in Table 3.7. In most cases, it is the χ^2 measure that disagrees with the other measures.

For about 80% of the pairs – those in group A – the statistical measures indicate the correct direction of association. This demonstrates that the rank measures are able to model human behavior in the elicitation experiment for most pairs. Group B, which contains pairs where the rank measures failed to make correct predictions is reassuringly small, with only 4 pairs. There are various possible explanations for the failure of the rank measures in these cases. For example, the pair *missile silos* exhibits almost equally strong predictiveness in both directions according to the subjects. This view is only partly reflected by the rank measures. The χ^2 measure comes close with a low forward and backward rank. However, the other measures only have a low backward rank, but not a low forward rank for this pair. They rank other words (e.g. *crisis*, *launcher* or *technology*) more highly. This discrepancy between human judgements and corpus data could be due to the fact that missile silos were a more dominant topic during the cold war – at the time when the BNC data were collected – and that subjects today are less familiar with them.

Group D contains word pairs where the rank measures indicate the strongest possible predictiveness in both directions, regardless of which status the human data suggest. The five phrases in this category are MWUs with strong lexical association in both directions.

We do not give a deeper discussion of the five pairs in D and the four pairs in B but we suspect that corpus data fail to provide a good prediction of human behavior in these cases because of differences between spoken and written language. For example, *wishful thinking* is not very left-predictive according to the human subjects – subjects gave responses like *quick* ($b = 0.175$), *good* (0.165), *clear* (0.114) and *critical* (0.108) more often than *wishful* (0.006). But in the BNC, *wishful* is by far the most common adjective preceding *thinking* (147 instances vs. 65 for *new thinking* and 43 for *critical thinking*). Other reasons for the discrepancy between corpus-derived and human associations for groups B and D could be part-of-speech ambiguity (*thinking* is predominantly a present participle, not a noun) and dialectal differences – *bloody hell* and *unleaded petrol* are British English expressions that American English speakers may not be familiar with.

One important difference between the rank measures and raw conditional probabilities can be found in all pairs of group D except for *bloody hell.* We will illustrate the phenomenon for *wishful thinking.* Human judgement for this pair is overwhelmingly right-predictive ($f = .952$, $b = .006$). The word *wishful* only occurs with two different nouns in the corpus and almost all its occurrences are with *thinking* which in turn occurs with about 100 other adjectives. This is naturally reflected in the conditional probabilities: $P^{\rightarrow}(\text{thinking}|\text{wishful}) = .924$ and $P^{\leftarrow}(\text{wishful}|\text{thinking}) = .089$. However, the association score of the two words in the combination *wishful thinking* is high enough to outrank all other adjectives that appear with *thinking* resulting in rank 1 in both directions. This can simply be interpreted as the rank measures suggesting likely completions to a stimulus (based on the distribution in the corpus) whereas conditional probabilities are suited to measure absolute association strength. The other two pairs where conditional probabilities perform better than the rank measures are *South East* and *laboratory experiments* from group C. Here, the rank measures are ambivalent but the conditional probabilities capture the correct direction of predictiveness.

For the pair *aching void*, conditional probability makes the wrong prediction and the rank measures are correct. Conditional probability suggests near-symmetry ($P^{\rightarrow} = .038$, $P^{\leftarrow} = .037$) whereas the ranks (except for R_{χ^2}) conform with the subjects' judgements of left-predictiveness.

f	b	(w_1, w_2)	$R_f^{\rightarrow}$	$R_f^{\leftarrow}$	$R_{G^2}^{\rightarrow}$	$R_{G^2}^{\leftarrow}$	$R_t^{\rightarrow}$	$R_t^{\leftarrow}$	$R_{\chi^2}^{\rightarrow}$	$R_{\chi^2}^{\leftarrow}$
***group A**: rank measures and direction scores conform*										
.589	.254	Academy Award	1	9	1	2	1	7	1	2
.236	.001	alarming rate	1	56	1	44	1	45	2	55
.332	.001	ancestral home	1	25	1	13	1	19	1	11
.698	.043	bated breath	1	5	1	2	1	5	1	2
.555	.160	cable television	2	7	1	4	2	5	1	2
.229	.033	choux pastry	1	5	1	5	1	5	1	4
.012	.008	cut glass	1	75	1	46	1	58	1	40
.020	.001	disclosure letter	1	9	1	3	1	6	1	4
.676	.001	felled tree	1	54	1	33	1	45	1	17
.101	.001	gleaming teeth	3	42	4	36	3	35	6	42
.068	.002	hunched shoulders	1	16	1	7	1	14	1	2
.332	.238	irritable bowel	1	4	1	2	1	3	1	2
.087	.001	old-fashioned way	1	98	1	60	1	62	4	70
.204	.010	radiant smile	1	48	1	32	1	47	3	20
.166	.006	rightful place	1	26	1	6	1	15	2	4
.177	.055	rising tide	3	4	3	4	3	4	1	3
.150	.049	rope ladder	1	4	1	4	1	4	2	4
.728	.064	sedimentary rocks	1	6	1	4	1	5	3	5
.024	.001	shrewd idea	3	109	6	49	3	68	10	41
.002	.001	stunning effect	2	135	9	129	2	99	15	190
.171	.001	thick-set man	1	519	1	169	1	318	1	86
.220	.009	vivid memories	1	5	1	4	1	4	2	5
.064	.006	well-worn path	1	71	1	34	1	58	1	22
.475	.065	*wellington boots	1	5	1	3	1	5	1	1
.012	.012	*impending retirement	9	18	8	14	9	18	9	9
.243	.014	*blackout curtains	1	9	1	4	1	8	1	1
.060	.056	*speech recognition	1	2	1	1	1	2	2	4
.001	.009	annual rent	29	2	20	1	28	2	15	7
.086	.119	Christian religion	12	4	12	1	12	2	18	5
.026	.820	Christmas decorations	11	1	8	1	11	1	9	3
.001	.145	civil libertarians	47	1	26	1	38	1	10	1
.001	.010	female preferences	63	34	92	44	60	34	95	48
.001	.002	great delicacy	473	1	206	1	260	1	199	3
.001	.065	hard frost	39	1	21	1	35	1	10	1
.017	.696	high fidelity	352	2	164	1	215	2	137	3

The table is continued on the next page.

f	b	(w_1, w_2)	$R_f^{\rightarrow}$	$R_f^{\leftarrow}$	$R_{G^2}^{\rightarrow}$	$R_{G^2}^{\leftarrow}$	$R_t^{\rightarrow}$	$R_t^{\leftarrow}$	$R_{\chi^2}^{\rightarrow}$	$R_{\chi^2}^{\leftarrow}$
.001	.031	legal wrangling	151	1	58	1	110	1	25	1
.001	.012	major shake-up	248	1	74	2	136	1	42	4
.008	.132	smoked mackerel	5	1	3	1	5	1	3	1
.001	.066	social solidarity	161	1	94	1	118	1	74	4
.001	.003	southern bypass	21	1	15	1	20	1	10	1
.026	.203	water sports	21	2	25	2	20	1	47	18
.004	.042	welcome diversion	17	3	15	1	16	3	8	3
.003	.042	*bond issuance	10	1	7	1	10	1	2	2
.038	.632	*food poisoning	10	1	3	1	10	1	2	2
.054	.450	*deadly nightshade	7	1	3	1	7	1	1	1
.006	.010	*aching void	5	3	3	1	5	3	1	2
.047	.389	*white collar	11	1	6	1	9	1	5	6
.274	.840	treasure trove	4	1	1	1	4	1	1	1
***group B**: rank measures and direction scores do not conform*										
.095	.116	healthy food	6	19	6	20	5	15	13	53
.156	.154	missile silos	16	1	8	1	16	1	2	1
.001	.006	seasoned campaigners	1	9	1	6	1	9	1	6
.292	.530	*precious metals	1	2	1	2	1	2	1	1
***group C**: rank measures ambivalent*										
.541	.462	*epileptic seizure	2	3	2	1	2	3	2	1
.076	.033	*dedicated follower	7	3	2	3	4	3	3	6
.434	.023	*laboratory experiments	2	1	1	1	2	1	1	1
.068	.183	*South East	1	2	3	2	1	2	5	2
***group D**: high mutual predictiveness*										
.296	.133	bloody hell	1	1	1	1	1	1	1	1
.127	.283	*special needs	1	1	1	1	1	1	1	2
.681	.279	toxic waste	1	1	1	1	1	1	1	1
.261	.158	unleaded petrol	1	1	1	1	1	1	1	1
.952	.006	wishful thinking	1	1	1	1	1	1	1	1

Table 3.7: Forward and backward scores and rank measures for the word pairs used in the elicitation experiment; '*' indicates disagreement of rank-measures on direction of predictiveness.

3.5.3 Asymmetry as an MWU Property

With the available dataset of elicited forward and backward scores for 61 word phrases, we want to investigate asymmetry as a property of MWUs. We define a phrase as an MWU based on look-up in a lexical resource and a world-knowledge resource (see below for details about the database look-up). For this evaluation, we consider strong asymmetry which we define as either forward or backward score being > 0.5. Table 3.8 shows the relation between the asymmetry property and MWU status (if a phrase was found to be an MWU, it has MWU status TRUE). We found that there are 32 MWUs and 29 non-MWUs in the data. Roughly 40% (13 out of 32) MWUs have strong asymmetry and 19 do not. None of the non-MWUs have strong asymmetry and hence there are 29 phrases that are non-MWUs and do not have the asymmetry property. To summarize, strong asymmetry is a phenomenon that can be found in some MWUs but MWUs do not necessarily exhibit this property. Furthermore, none of the phrases that were classified as non-MWUs had strong asymmetry.

An MWU classifier with strong asymmetry as its only feature (the model that predicts MWU status TRUE if a candidate has strong asymmetry and FALSE otherwise) outperforms a baseline majority classifier (the model that predicts MWU status TRUE for every candidate) on our data. The baseline classifier has 52% accuracy (32 out of 61 phrases are MWUs) whereas the asymmetry-based model has 69% accuracy (13 phrases have strong asymmetry and are MWUs; 29 do not have asymmetry and are non-MWUs). Thus, it can be argued that strong asymmetry is a useful indicator for a phrase being an MWU.

We base our analysis on the assumption that collocations and MWUs represent a “conventional way of saying things” (Manning and Schütze, 1999, 151). For this purpose, we need to decide if a phrase is an MWU. We consult two freely accessible knowledge sources, Wikipedia as a resource for world knowledge and Wiktionary as a lexical resource. Our decision criterion is that a phrase is an MWU if it has an entry in either of the two databases. Being listed in these resources means that a phrase is worth recording be-

		strong asymmetry	
		TRUE	FALSE
MWU status	TRUE	13	19
	FALSE	0	29

Table 3.8: Confusion matrix of asymmetry feature and MWU status

cause it refers to a real-world entity, e.g. *Academy Award* or *cable television*, or because it is a habitual combination used to express an important recurring concept or sentiment, e.g. *wishful thinking*. It can be argued that combinations that have an entry in these resources do in fact represent a conventional way of saying things and can thus be classified as MWUs.[11]

3.5.4 Quantitative Evaluation

We have introduced three different approaches to predictiveness and asymmetric association: (i) direction scores f and b computed from the elicitation experiment; (ii) rank measures $R^{\rightarrow}$ and $R^{\leftarrow}$ and (iii) conditional probabilities $P^{\rightarrow}$ and $P^{\leftarrow}$. The latter two are based on corpus data. Scores, ranks and conditional probabilities are capable of capturing asymmetries between the two components of a word pair. In this section, we perform a quantitative

[11] Some of the phrases redirect to articles with a different name on Wikipedia (e.g. *food poisoning* redirects to *foodborne illness*). We count a redirect as a match. We did not count a match when the database entry consisted only of proper nouns (e.g. *Rising Tide* is the name of a novel) because or strategy for pair extraction (Section 3.3.1), we did not allow all-proper-noun combinations. Database look-ups were carried out on November 22nd, 2012 on `http://www.en.wikipedia.org` and `http://www.en.wikitionary.org`.

evaluation of how well the predictions made by the corpus-based measures agree with the human scores.

Our test case is the distinction between right-predictive and left-predictive word pairs. We cast the task as a classification problem of right-predictive pairs. We predict the class directly from the scores or respective measures. The output variable Y takes on the values TRUE if the prediction is right-predictive and FALSE otherwise.[12]

$$Y = \begin{cases} \text{TRUE} & \text{if } f \geq b \text{ (right-predictive)} \\ \text{FALSE} & \text{if } f < b \text{ (left-predictive)} \end{cases}$$

Analogously, we transformed the conditional probabilities into a corresponding input variable:[13]

$$X = \begin{cases} \text{TRUE} & \text{if } P^{\rightarrow} \geq P^{\leftarrow} \text{ (right-predictive)} \\ \text{FALSE} & \text{if } P^{\rightarrow} < P^{\leftarrow} \text{ (left-predictive)} \end{cases}$$

We applied a similar procedure to each corpus-based rank measure. For instance, the input variable for the t-score measure t is given by:

$$X_t = \begin{cases} \text{TRUE} & \text{if } R_t^{\rightarrow} \leq R_t^{\leftarrow} \text{ (right-predictive)} \\ \text{FALSE} & \text{if } R_t^{\rightarrow} > R_t^{\leftarrow} \text{ (left-predictive)} \end{cases}$$

Recall that a lower rank indicates higher association. Therefore, a pair with $R^{\rightarrow} > R^{\leftarrow}$ is *left*-predictive and is assigned an input value of X = FALSE. In the case of equal ranks, we also assigned X = TRUE (right-predictive) because the human subjects – as well as our corpus data, see Section 3.3.3 – showed a preference for right-predictiveness – the elicitation experiment yielded 34 word pairs with $f > b$, compared to 27 with $f < b$. The first four rows of Table 3.9 show the accuracy of predictions made by the four rank measures.

The data set contains 34 RP and 27 LP pairs. Therefore, a baseline

[12] The case $f = b$ did not occur in the human data.

[13] Again, the case $P^{\rightarrow} = P^{\leftarrow}$ did not occur.

rank measure	accuracy	95% confidence interval
R_f	88.5%	77.8% ... 95.3%
R_{G^2}	**90.2%**	79.8% ... 96.3%
R_t	88.5%	77.8% ... 95.3%
R_{χ^2}	82.0%	70.0% ... 90.6%
cond. prob.	**90.2%**	79.8% ... 96.3%
baseline	62.3%	49.0% ... 74.4%

Table 3.9: Accuracy of predictions made by the corpus measures

classifier that assigns every pair to category RP (i.e., $X = \textsc{true}$) achieves an accuracy of 55.7%. In our evaluation, we use a more optimistic cross-validation baseline where the most frequent category is chosen separately for each of the six data folds.[14] The resulting baseline accuracy of 62.3%, calculated over all 61 items, is reported in the last row of Table 3.9.

Because of the small sample size used for the evaluation, statistical significance testing is essential. As an indication of the amount of random variation, we calculated binomial 95% confidence intervals for the proportion of correct predictions, shown in the rightmost column of Table 3.9.

All rank measures perform well, even compared to the optimistic baseline. The best result is achieved by log-likelihood (R_{G^2}) with an accuracy of 90.2%. The binomial confidence interval indicates that the R_{G^2} rank measure will achieve a prediction accuracy of at least 79.8% on larger data sets. Frequency (R_f) and t-score (R_t) are tied in second place, with a score of 88.5%. This is no coincidence: the two measures happen to make identical predictions for all items in our data set (i.e., $X_f = X_t$), although they are not equivalent in general.[15] Chi-square (R_{χ^2}) performs considerably worse than the other rank

[14] The baseline is optimistic because the most frequent category is determined from the test fold in each case, rather than from the training folds. For instance, if the first fold contained 7 RP pairs and 3 LP pairs, the optimistic baseline classifier would assign all pairs in this fold to category RP. If the third fold contained 4 RP and 6 LP pairs, the optimistic baseline would assign all pairs in this fold to category LP.

[15] Note that the difference between these measures and log-likelihood corresponds to a single word pair: R_{G^2} makes 55 correct predictions vs. 54 for f and t. Our experiment therefore provides no reliable evidence that any of the three measures is better than the

measures, but is still much better than the baseline, with a 95% confidence interval ranging from 70% to about 90% accuracy. Conditional probabilities perform as well as the best rank measure (but with different predictions).

We used an exact version of McNemar's test (Hollander and Wolfe, 1999, 468–470) to assess the significance of result differences. This test considers only items for which the two models to be compared made different predictions. Due to the small sample size, there are no significant differences between any of the models. In particular, we were not able to show that R_{G^2} is significantly better than R_{χ^2} (exact McNemar, $p = .063$), even though Table 3.9 shows a clear difference. However, all models except for χ^2 are significantly better than the optimistic baseline (with p-values ranging from $p = .001$ for R_{G^2} to $p = .017$ for R_{χ^2}).

A simple ranking by co-occurrence frequency (R_f) once again performs astonishingly well, reaching the same accuracy as t-score (R_t). Both measures only take the first cell of the contingency table into account, but t-score additionally considers the difference between observed and expected frequencies. It is interesting to note that R_{G^2} is the best of the five measures and R_{χ^2} the worst, even though they are both independence tests using information from the full contingency table. A possible explanation is the tendency of R_{χ^2} to overestimate significance in highly skewed contingency tables (see Dunning, 1993; Evert, 2004).

3.6 Summary

In this chapter we highlighted one important property of syntagmatic word combinations: asymmetric association between their components. This property has long been neglected due to a lack of appropriate techniques for corpus data. Our rank-based asymmetric association measures provide, for the first time, a suitable empirical operationalization of asymmetry in syntagmatic word combinations. In addition, future theoretical discussions can draw on the results of our syntagmatic association experiment as a complementary

other two.

form of evidence.

We have discussed two important distinctions in this chapter, the distinction between syntagmatic and paradigmatic relations and the distinction between symmetric and asymmetric relations. Asymmetry in paradigmatic relations (e.g. asymmetric similarity) has received attention in the past in psychological and corpus-based studies and it has been shown that asymmetric similarity measures can be of use for a number of applications.

Previous research was often based on free association norms and rating experiments which capture mostly paradigmatic relations and do not allow for systematic investigation of asymmetry. We have investigated asymmetry in syntagmatic relations. We designed a novel experiment setup to collect human data on syntagmatic combinations. In our study, we compared syntagmatic combinations in corpora and in human-subject experiments and demonstrated that corpus-derived rank measures and conditional probabilities can predict the asymmetry of human syntagmatic associations with high accuracy. We found that conditional probabilities are suited to measure absolute association strength, whereas rank measures are a good indicator for which responses could be the best completion to a given stimulus. We have shown asymmetry occurs in human production and that strong asymmetry is characteristic of MWUs. We found that right-predictive asymmetry is more prevalent than left-predictive asymmetry.

Chapter 4

Multi-Word Tokenization

4.1 Introduction

The research presented in this chapter is joint work with Alok Kothari, Christina Lioma and Martin Forst (Michelbacher et al., 2011b). We present an implementation of MWT and evaluate its performance against a gold standard of MWUs. Additionally, we measure its impact on an information retrieval system. At the heart of our implementation is the notion of the semantic head. The semantic head addresses the property of non-compositionality in MWUs and captures the non-compositional core of a phrase. Our hypothesis is that an MWU is a phrase that is identical to its semantic head; in contrast, if a phrase is a compositional combination then it is not an MWU and the semantic head is identical to the syntactic head. In our implementation, the MWU decision step is realized as semantic head recognition. We propose a supervised cascaded classification approach to semantic head recognition. This approach can process phrases of arbitrary length and leverages different feature types, namely statistical association measures and features of contextual similarity. The latter feature type is targeted at non-compositionality detection. We show that context features significantly enhance a baseline semantic head recognizer. However, we also identify a the challenges of using contextual similarity in high-confidence semantic head recognition.

We run experiments on a collection of documents from the physics domain. We focus on the physics domain because domain-specific terminology is often encoded in MWUs. In the face of resource scarcity in specialized domains, MWT for such domains is of particular importance; for many domains, comprehensive and up-to-date lists of domain-specific terminology are not available.

In this study, we focus on the property of non-compositionality to recognize MWUs. In the context of information retrieval this is the most important MWU property because satisfying an information need is primarily a semantic question and less a question of, for example, subtle morphosyntactic variations which fall under non-modifiability.[1]

The remainder of this chapter is structured as follows. Section 4.2 details semantic heads and describes our implementation of MWT based on semantic head recognition. In Section 4.3, we describe the evaluation setup consisting of (i) the creation of the gold standard of MWUs (Section 4.3.1) and (ii) the incorporation of MWU information into information retrieval (Section 4.3.2). In Section 4.4 we present and discuss the results of the two experiments. In Section 4.5 we give a summary of this chapter.

4.2 An Implementation of MWT

In this section, we describe an implementation of MWT. The main part of MWT is the MWU decision step. In our implementation, the decision MWU vs. non-MWU is made by recognizing the semantic head in MWU candidates. For this purpose, we propose a cascaded classification approach to semantic head recognition.

The cascaded model allows us to process MWU candidates of arbitrary length. For classification, we use a number of previously proposed features for recognizing non-compositionality based on association measures. In addition, we compare features that address non-compositionality with measures of contextual similarity.

[1] For MWT in general this restriction does not hold.

We first introduce the concept of semantic head and its relation to MWT. Second, we describe the cascaded classification approach. Third, we define and discuss the features that we use in classification.

4.2.1 Semantic Heads

The importance of syntactic heads for many NLP tasks is generally accepted. For example, in coreference resolution, identity of syntactic heads is predictive of coreference; in parse disambiguation, the syntactic head of a noun phrase is a powerful feature for resolving attachment ambiguities. However, in all of these cases, the syntactic head is only an approximation of the information that is really needed; the underlying assumption made when using the syntactic head as a substitute for the entire phrase is that the syntactic head is representative of the phrase. This is not the case when the phrase is non-compositional.

We define the *semantic head* of an NP as the largest non-compositional part of the phrase that contains the syntactic head. For example, *dog* is the semantic head of *new dog* in (4.1) and *hot dog* is the semantic head of *tasty hot dog* in (4.2). In the first case, syntactic and semantic heads coincide.

Semantic heads would serve most NLP tasks better than syntactic heads. For example, a coreference resolution system is misled if it looks at syntactic heads to determine possible coreference of *the new dog . . . a tasty hot dog*. This is not the case for a system that makes the decision based on the semantic heads *hot dog* of *a tasty hot dog* and *dog* of *the new dog*.

(4.1) I took the new dog to the vet.

(4.2) Then I had a tasty hot dog.

The semantic head is either a single noun or a non-compositional noun phrase. In the latter case, the modifier(s) introduce(s) a non-compositional, unpredictable shift of meaning; *hot* shifts the meaning of *dog* from live animal to food. In contrast, the compositional meaning shift caused by *small* in *small dog* is transparent. The semantic head always contains the syntactic head; for compositional phrases, syntactic head and semantic head are identical.

The connection between semantic heads and MWT is straightforward: The goal of MWT is to process MWUs as a whole. In non-compositional phrases the semantic head constitutes an MWU. This means that by recognizing the semantic head in a candidate, we have all the information necessary to process an MWU as a single token.

4.2.2 The Term *Semantic Head* in the Literature

In the literature, the term semantic head is sometimes used to refer to that part of a noun phrase that carries its (main) meaning. We discuss two concrete works that use the term and that are related to MWT.

Korkontzelos and Manandhar (2009) use the term semantic head in a study on detecting compositionality in MWUs. They use it to refer to the head noun in NPs. Their compositionality detection strategy is similar to ours in that it uses distributional semantics to measure semantic distance between the semantic head and the whole phrase. Their method to identify semantic heads relies on an arbitrary similarity threshold rather than statistical classification. Furthermore, they do not mention semantic heads that span multiple words.

Fillmore et al. (2002) draw the distinction between the syntactic and semantic head of a phrase in the context of predicates and the semantic roles of their arguments. They identify "case[s] in which there is a discrepancy between the syntactic 'head' of the phrase and its semantic 'head.'" This phenomenon occurs with what the authors call transparent nouns. In the noun phrase in (4.3), the syntactic head *kind* is a transparent noun whereas the semantic head *proposal*, the part of the phrase that carries the meaning, is embedded in prepositional phrase:

(4.3) *I object to that kind of proposal.*

Fillmore et al. (2002) suggest a list of common transparent constructions, e.g.

Parts	*part of the room*
Measures	*liter of wine*
Aggregates	*herd of wildebeest*
Types	*kind of fish*
Unitizers	*bout of the flu*
Evaluatives	*her jerk of a husband*

The above uses of the term semantic head are similar to ours although in both cases, the term semantic head is used without proper introduction. Korkontzelos and Manandhar (2009) have the same goal, namely MWU identification, but their scope is narrower (no multi-word semantic heads) and their model is less powerful (arbitrary threshold instead of a learning algorithm and only one comparison method). Fillmore et al. (2002) use semantic heads in a different scenario, namely semantic roles and their goals are not directly related to MWT. They do, however, touch on an important aspect of the idea of semantic heads that our model does not currently cover: the case where the syntactic and semantic head do not share any words. For our purposes of MWT in this thesis, the focus is not on this aspect of semantic heads but on investigating the feasibility of a particular classification approach to MWT. To address this problem in a future implementation of MWT, we suggest a stop list of transparent nouns to "forward" semantic head recognition into the prepositional phrase.

For example, an MWT system that is aware of transparent nouns, e.g. by including a stop list of common nouns of measurement, would label *wine* as the semantic head in *liter of wine*. For most NLP applications, we are more interested in the semantic content of phrases with transparent nouns (*wine*) rather than the meaning of the transparent noun itself (*liter*).[2]

[2] In another scenario, an information extraction system might be designed to discover the quantities of consumed substances and an appropriate MWT system should label *liter* as the semantic head. For cases like these, the system could include a white list that contains nouns of measurement.

(1)		*neutron*	**star**	
(2)	*unusual*	black	**hole**	
(3)	*bright*	optical	**afterglow**	
(4)		*small*	**moment**	of inertia

Figure 4.1: Example phrases with modifiers. Peripheral elements are set in italics, syntactic heads in bold.

4.2.3 Cascaded Model for MWU Decision

Semantic head recognition implements the MWU decision step that is at the core of MWT. To recognize the semantic head of a phrase, we use a cascaded classification approach. We need a cascade because we want to recognize the semantic head in noun phrases of arbitrary length.

The starting point is a phrase of length n: $p = w_1 \dots w_n$. We distinguish between the syntactic head of a phrase and the remaining words, the modifiers. Figure 4.1 shows phrases of varying syntactic complexity. The phrases are taken from the iSearch corpus (see Section 4.3). The syntactic head is marked in bold. The model accommodates prenominal modifiers as in examples (1) through (3) and post-nominal modifiers like PPs in example (4).

Among the modifiers, there is a distinguished element, the *peripheral element* u (italicized in the examples). The remaining words are called the *rest* v. We can now represent any phrase p as $p = uv$.[3] The element u is always the outermost modifier. *of*-PPs are treated as a single modifier and they take precedence over prenominal modification because this analysis is dominant in our gold standard data. This means that in the phrase *small moment of inertia*, *small* (and not *of inertia*) is the peripheral element u.

Cascaded classification then operates as shown in Figure 4.2. In each iteration, the classifier decides whether the relation between the current peripheral element u and the rest v is compositional (C) or non-compositional (NC).[4] If the relation is NC, processing stops and uv is returned as the seman-

[3] We use the abstract representation $p = uv$ even though u can appear after v in the surface form of p.

[4] This “outside-in” approach was chosen to reflect our view of MWU decision which is

```
function $recognize\_semantic\_head(p)$
  $u \leftarrow peripheral(p)$
  $v \leftarrow rest(p)$
  while $decision(u, v) \neq NC$ do
    $u \leftarrow peripheral(v)$
    if $u = \emptyset$ then
      return $v$
    $v \leftarrow rest(v)$
  return $uv$
```

Figure 4.2: Cascaded classification of p

tic head of p. If the relation is compositional, u is discarded and classification continues with v as the new input phrase, which again is represented in the form $u'v'$. In case there is no more peripheral element u, i.e. the new v is a single word, it is returned as the semantic head of p.

Table 4.1 shows two examples. For the fully compositional phrase *bright optical afterglow*, the process runs all the way down to the syntactic head *afterglow* which is also the semantic head. In the second case, the process stops earlier, in step 2, because the classifier finds that the relation between *moment* and *of inertia* is NC. This means that the semantic head of *small moment of inertia* is *moment of inertia*. Cascaded classification provides a framework for recognizing semantic heads that allows us to treat noun phrases of arbitrary length.

4.2.4 Feature Definitions

MWT is a classification task with the MWU decision step at its core. For this kind of problem, linear regression is not the optimal approach because it is not a classification algorithm. The classification method of logistic regression is an appropriate tool for MWU decision because the output variable is categorical (C vs. NC). For our experiments, we use the implementation of

to identify MWUs starting with the candidate rather than an "inside-out" approach where we would add modifiers to syntactic heads until we reach an MWU.

step	u	v	decision
1	*bright*	optical **afterglow**	C
2	*optical*	**afterglow**	C
3	∅	**afterglow**	
1	*small*	**moment** of inertia	C
2	*of inertia*	**moment**	NC

Table 4.1: Cascaded decision processes

a logistic regression classifier provided by the Stanford classifier.[5]

The cascaded classifier introduced above has to decide if a meaning shift occurs when removing a peripheral element. In this section we will show how the classifier's features are designed to achieve this goal. There are two types of features; AM-based and based on contextual similarity. AM features are the standard approach for MWU recognition and we expect those features to provide good basic performance. Similarity-based features are designed to recognize meaning shifts on the semantic level leveraging distributional semantics.

Features Based on Association measures

Collocation extraction is not traditionally seen in connection with tokenization. In many cases, it is carried out with lexicography in mind, i.e. assisting lexicographers in compiling and extending dictionaries (Choueka, 1988; Church and Hanks, 1990; Smadja, 1993; Schone and Jurafsky, 2001). Consequently, the idea of fully automated extraction, which is necessary in the context of tokenization, is not prevalent. The standard use case is to compile lists of collocation candidates sorted by association strength and pass these lists on to experts who do manual classification.

Our interest in MWT evolved from the study of MWUs and classic col-

[5] Version 2.0, available from `http://nlp.stanford.edu/software/classifier.shtml`.

feature name	association measure
am_t	student's t-score
am_z	z-score
am_{χ^2}	chi-square
am_{PMI}	pointwise mutual information
am_D	Dice coefficient
am_f	frequency
am_{G^2}	log-likelihood
am_{scp}	symmetric conditional probability

Table 4.2: Features based on association measures

location extraction. Statistical association measures are frequently used for MWU detection and collocation extraction. We use association measures as features for automatic candidate classification. This approach has been used in recent studies in collocation extraction with encouraging results (Pecina, 2008; Ramisch et al., 2010).

Features based on association measures are defined as the respective association scores. We use all association measures from Schone and Jurafsky (2001) that can be derived from a phrase's contingency table. These measures are t-score, z-score, χ^2, pointwise mutual information (PMI), Dice coefficient, frequency, log-likelihood (G^2) and symmetric conditional probability.

The measures are designed to deal with two random variables U and V that traditionally represent single words. In our model, we use U to represent peripheral elements u and V for rests v. See Table 4.2 for a summary of the features we consider for MWU decision.

Features Based on Contextual Similarity

In recent years, a number of studies have investigated the relationship between distributional semantics and non-compositionality. These studies compute the similarity between words and phrases represented as semantic vectors in a word space model. The underlying idea is similar to the one proposed in Lin (1999). The gist of Lin's idea is that the meaning of a non-compositional phrase somehow deviates from what one would expect given the semantic vectors of parts of the phrase. This is exactly the meaning shift that we would like to detect.

The standard measure to compare semantic vectors is cosine similarity. The questions that arise are (i) which vectors to compare, (ii) how to combine the vectors of the parts and (iii) how to translate similarity values into non-compositionality. There are no generally accepted answers to these questions.

Regarding (i), Schone and Jurafsky (2001) compare the semantic vector of a phrase p and the vectors of its component words in two ways: one includes the contexts of p in the construction of the semantic vectors of the parts and one does not. Regarding (ii), they suggest weighted or unweighted sums of the semantic vectors of the parts. Baldwin et al. (2003) investigate semantic decomposability of noun-noun compounds and verb constructions. They address (i) by comparing the semantic vectors of phrases with the vectors of their parts *individually* to detect meaning changes; e.g. they compare *vice president* to *vice* and *president*. With respect to (iii), the above-mentioned studies use ad hoc thresholds to separate compositional and non-compositional phrases but do not offer a principled decision criterion.[6] In contrast, we train a statistical classifier to learn a decision criterion.

We use three methods of comparing semantic vectors (Table 4.3): sj1 and sj2, both introduced by Schone and Jurafsky (2001). Additionally, we propose a new comparison, which we call alt. Method alt compares the semantic vector of a phrase with its alternative vector.

We build the *alternative vector* as follows. For a phrase $p = uv$ with peripheral element u and rest v, we call the phrase $p' = u'v$ an *alternative phrase* if the rest v is the same and $u' \neq u$. e.g. *giant star* is an alternative phrase of *neutron star* and *isolated neutron star* is an alternative of *young neutron star*. The alternative vector of p is then the semantic vector that is computed from the contexts of all of p's alternative phrases. The alternative vector is a representation of the contexts of v except for those modified by u.

[6] Lin (1999) uses a well-defined criterion but his approach is not based on vector similarity.

feature name	vector comparison (example)
sim_{sj1}	$s(w(\textit{black hole}),\ w(\textit{black})\ +\ w(\textit{hole}))$
sim_{sj2}	$s(w(\textit{black hole}),\ w^*_{\textit{black hole}}(\textit{black})\ +\ w^*_{\textit{black hole}}(\textit{hole}))$
sim_{alt}	$s(w(\textit{black hole}), \sum_u w(u\ \textit{hole}));\ u\ \neq\ \textit{black}$

Figure 4.3: Features based on vector similarity. Here, s represents vector similarity, w a semantic vector and w^*_p the semantic vector of a part of a phrase p that does not include those occurrences that were part of p itself. In this example, $p = \textit{black hole}$.

This technique bears resemblance to the substitution approach of Lin (1999). The difference is that he relies on a similarity thesaurus for substitution and monitors the change in mutual information for each substitution individually whereas we substitute with general alternative modifiers and combine the alternative contexts into one vector for comparison.

We use the cosine similarities sim_{sj1}, sim_{sj2} and sim_{alt} as features for the classifier. We refer to these features as context features because (unlike AM features) they take into account the similarity of the contexts in which words and phrases occur. Our intuition is that cosine similarity should be small if a phrase is non-compositional and large if it is compositional. In other words, if the contexts of the candidate phrase are too dissimilar to the contexts of the sum of its parts or to the alternative phrases, then we suspect non-compositionality.

Previous work has compared the semantic vector of a phrase with the vectors of its components. Our approach is more "head-centric" and only compares phrases in the same syntactic configuration. Our question is: Is the typical context of the head *hole* if it occurs with a modifier that is not *black* different from when it occurs with the modifier *black*?

To create semantic vectors, we used a bag-of-words model with a co-occurrence window of 10 words in each direction. We only kept the content words in the window which we defined as words that are tagged as either noun, verb, adjective or adverb. To add information about the variability of syntactic contexts in which phrases occur, we add the words immediately before and after the phrase with positional markers (-1 and $+1$, respectively) to the vector. These words were not subject to the content-word filter. The dimensionality of the vectors is then $3V$ where V is the size of the vocabulary: V dimensions each for bag-of-words, left and right syntactic contexts. We did not include vectors for the stop word *of* for sj1 and sj2.

4.3 Evaluation Setup

The evaluation comprises two experiments, an intrinsic and an extrinsic evaluation of our MWT implementation. The intrinsic part evaluates the task of semantic head recognition and focuses on the impact of different features of contextual similarity on MWU decision performance. The extrinsic part evaluates MWT in the context of an application, in this case, information retrieval. In this section, we describe the experimental setups for both parts of the evaluation. We present and discuss the results of both parts in Section 4.4.

For both experiments we used the iSearch collection (Lykke et al., 2010). This collection is a suitable resource for our purposes because it provides (i) a domain-specific resource with a large repository of textual data which we need for the computation of corpus statistics and semantic vectors; and (ii) a resource specifically designed for information retrieval research. The collection is composed of documents and book records from the physics domain mainly from high energy physics, condensed matter physics and astrophysics. There are around 140,000 full articles and 290,000 abstracts in the collection. We use the text contained in these documents as our corpus. The text in the collection amounts to about 1 billion words. Our preprocessing steps were SWT, part-of-speech annotation and lemmatization.

```
[pos = "DT"]?                                    #optional determiner
[pos = "CD|JJ.?"]*                               # adjectival modifier
[pos = "N(P|N).?"]                               # nominal modifier
[pos = "NN.?"]                                   # head noun
([pos = "IN" & word = "of"] [pos = "NN"])?       # of-PP (optional)
```

Table 4.3: Part-of-speech pattern for candidate extraction

4.3.1 Semantic Head Recognition

This section describes the experiment for the evaluation of the MWT implementation. The objective is semantic head recognition, i.e. recognizing the semantic head of MWU candidates. We created a gold standard of MWU candidates for the experiment. The candidates where extracted from the corpus. We then asked human annotators to identify the semantic head of each candidate. Based on the judgements, we separated the candidates into MWUs and non-MWUs. The group of non-MWUs are the candidates whose semantic head and syntactic head are the same.

In what follows, we describe how candidates were extracted from the corpus, how we collected non-compositionality judgements from human experts and the concrete evaluation tasks.

Candidate Phrases

Initially, we extracted all noun phrases from the corpus that consist of a syntactic head with up to four modifiers. The prenominal modifiers can be nouns, proper nouns, adjectives or cardinal numbers. Table 4.3 shows the part-of-speech pattern we used. Almost all domain-specific terminology in the corpus is captured by this pattern.

The baseline accuracy of a classifier that always chooses compositionality is very high (> 90%) for phrases of the type *[noun] of the/a [noun] (sg.)* (e.g. *rest of the paper*) and *[noun] of [noun] (pl.)* (e.g. *series of papers*). We therefore restrict post-nominal modifiers to prepositional phrases with the word *of* followed by a non-modified, indefinite, singular noun, e.g. *speed of* *light* or *moment of inertia*.

Out of all phrases extracted with part-of-speech patterns, we keep only the ones that appear more often than 50 times because it is hard to compute reliable features for less frequent phrases. All experiments were carried out with lemmatized word forms. Finally, we drew a random sample of 1650 candidate phrases that form the basis of the gold standard.

Human Judgements

Since the domain of the corpus is physics, highly specialized vocabulary had to be judged. We use three domain experts as raters (one engineering and two physics graduate students). The gold standard was created based on semantic head annotations for 1650 candidates.[7] The candidates consisted of a syntactic head with at least one and at most four modifiers (three prenominal modifiers and one postnominal PP with *of*).

See Figure 4.4 for possible candidates (light gray) and potential semantic heads (dark gray). Given a phrase with several modifiers, the annotators have a varying number of choices for the semantic head depending on the length of the phrase. Depending on the modifiers, a phrase can have up to 8 different possible semantic heads. Of all 1650 phrases, we discarded the ones that did not get valid responses from all raters. A response is valid if it constitutes a potential semantic head. From this set we only kept the phrases where at least two annotators agreed on the semantic head.[8] The resulting gold standard data has 1560 elements.

To calculate annotator agreement, we used the Fleiss' κ measure (sometimes called multi-κ) which returns a single agreement number for multiple annotators working on the same task. The value for our data set is $\kappa = 0.58$ which is classified as *moderate* (verging on *substantial*) according to a commonly used scale (Artstein and Poesio, 2008).

[7] We gave the domain experts a thorough introduction to semantic heads before the annotation. See Appendix B for details.

[8] For example, the candidate *tasty hot dog* has three different possible semantic heads which means that each annotator could pick a different semantic head.

Evaluation Modes

The goal of this experiment is to evaluate the performance of semantic head recognition in MWU candidates. Here, we test the cascaded classification approach introduced in Section 4.2.3. We define three evaluation modes: *dec-1st*, *dec-all*, and *semh*. Mode *dec-1st* only evaluates the first decision for each phrase. In mode *dec-all*, we evaluate all decisions that were made in the course of recognizing the semantic head. This mode emphasizes the correct recognition of semantic heads in phrases where multiple correct decisions in a row are necessary. We define the confidence value (class probability) for multi-decision classification as the product of the confidence values of all intermediate decisions. The mode *semh* evaluates how many semantic heads were recognized correctly. This mode directly evaluates the task of semantic head recognition.

We compare different models, a basic model that uses association measures as features and several models that additionally use combinations of context features. We randomly split the gold standard data set into a training set of 1300 items and a test set of 260 items.

We give technical details of the classification and describe the different models. All feature values are binned into 5 bins. We applied a log transformation to the four AMs with large values: am_f, am_{G^2}, am_{χ^2} and am_z. For our application there is little difference between statistical significance at $p < .001$ and $p < .00001$. The log transformation reduces the large gap in magnitude between high significance and very high significance. If co-

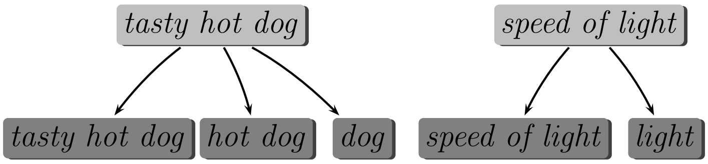


Figure 4.4: Potential semantic heads for the phrases *tasty hot dog* and *speed of light*

occurrence of u and v in uv is below chance, then we set the association scores to 0 since this is an indication of compositionality (even if it is highly significant).

Since AMs have been shown to be correlated (e.g. Pecina (2010)), we first perform feature selection on the AM features. We tested accuracy of all $2^r - 1$ non-empty combinations of the $r = 8$ AM features on the task of deciding whether the first decision during the classification of a phrase was C or NC. We then selected those AM features that were part of at least one top 10 result in each fold. These features were am_t, am_f and am_{scp} and they form the base-AM model. We trained models using the context features sim_{sj1} sim_{sj2} and sim_{alt}. Each of them uses the features of base-AM and one of the 7 possible combinations of context features. Below, we give a summary of the models:

base-AM: The basic model that uses the association measures am_t, am_f and am_{scp} as features (determined by feature selection).

base-AM + context features: The models that use the same features as base-AM and all combinations of the three context features sim_{sj1} sim_{sj2} and sim_{alt} (7 models).

4.3.2 Application to Information Retrieval

The field of information retrieval (IR) deals with the task of satisfying an information need by retrieving relevant information from a collection of information sources. For an in-depth introduction to information retrieval, see Manning et al. (2008).

A typical example of information retrieval is a web search engine. Users formulate their information need as query terms which they enter into a search field. The retrieval system then returns information sources from its collection (e.g. documents or images) ranked by their relevance to the query. In our experiments with the iSearch collection, information sources are text documents.

Evaluation of IR systems requires a data set that contains use cases of information need paired with relevance assessments. For example, query terms and documents that should be returned by the system for a particular query. The iSearch collection provides such data. We measure retrieval performance in terms of recall (REC), precision at 20 (P20) and mean average precision (MAP). REC measures the proportion of relevant documents retrieved to all relevant documents. P20 measures the proportion of relevant documents among the top 20 documents deemed most relevant by the system. Mean average precision (MAP) is the mean of precision at k for $k \in 1 \ldots N$ where N is the number of retrieved documents.

Typically, IR systems do not process non-compositional MWUs as one semantic entity, potentially missing out on important information encoded in non-compositional combinations. In this experiment, we illustrate one way of adjusting the retrieval process so that MWUs are processed as semantic entities that may enhance retrieval performance. The underlying hypothesis is that, given a query that contains an MWU, boosting the retrieval weight of documents that contain this phrase will improve overall retrieval performance.

We introduced MWT as a method that marks MWUs in the token stream for further processing. This is one possible way of incorporating MWUs into the pipeline. An alternative way is to use the identified MWUs directly if a particular application is better suited for this approach. In IR, there are several places to integrate MWUs. Applying MWT before indexing is one possibility. That way, MWU information would be hard-coded into the system. Another way is adjusting the weights of documents containing MWUs that appear in the query. This approach moves the concrete handling of MWU information to retrieval time. Here, we chose the approach of weight adjustment over index manipulation because adjusting weights at retrieval time is more efficient than re-indexing with different MWT models.

Experimental Setup

We boost retrieval weights using Indri's[9] combination of the language modeling and inference network approaches (Metzler and Croft, 2004), which allows assigning different degrees of belief to different parts of the query. This belief can be drawn from any suitable external evidence of relevance. In our case, this source of evidence is the knowledge that certain query terms constitute an MWU. Under this approach, and using the *#weight* and *#combine* operators for combining beliefs, the relevance of a document D to a query Q is computed as the probability that D generates Q, $P(Q|D)$:

$$P(Q|D) = \prod_{t \in Q} P(t|D)^{\frac{w_t}{W}} \qquad (W = \sum_{t \in Q} w_t) \tag{4.4}$$

where t is a term and w_t is the belief weight assigned to t. The higher w_t is, the higher the rank of documents containing t. In this experiment, we distinguish between two types of query terms: terms occurring in MWUs (Q_{nc}), and the remaining query terms (Q_c). Terms $t \in Q_{nc}$ receive belief weight w_{nc} and terms $t \in Q_c$ belief weight w_c, ($w_{nc} + w_c = 1$ and $w_{nc}, w_c \in [0, 1]$). To boost the ranking of documents containing MWUs, we increase w_{nc} at the expense of w_c. We estimate $P(t|D)$ in Equation 4.4 using Dirichlet smoothing (Zhai and Lafferty, 2002).

We use Indri for indexing and retrieval without removing stopwords or stemming. This choice is motivated by two reasons: (i) We do not have a domain-specific stopword list or stemmer. (ii) Baseline performance is higher when keeping stopwords and without stemming, rather than without stopwords and with stemming.

The collection includes a set of 65 queries with relevance assessments created by physicists. To match documents to queries without any MWU treatment (baseline run), we use the Kullback-Leibler language model with Dirichlet smoothing (KL-Dir) (Zhai and Lafferty, 2002). We then identified MWUs in the queries with the base-AM model (see Section 4.4.1 for why we used this model).

[9] `http://www.lemurproject.org/`

Our approach for boosting the weight of these MWUs uses the same retrieval model enhanced with belief weights as described in Equation 4.4 (real NC run). In addition, we include five runs that boost the weight of pseudo MWUs that were created randomly from the query text (pseudo MWU runs). These pseudo MWUs have exactly the same length as the observed MWUs for each query. For each evaluation measure, we separately tune the following parameters and report the best performance: (i) the smoothing parameter μ of the KL-Dir retrieval model (following Zhai and Lafferty (2002), we tested $\mu \in \{100, 500, 800, 1000, 2000, 3000, 4000, 5000, 8000, 10000\}$); (ii) the belief weights $w_{nc}, w_c \in \{0.1, \ldots, 0.9\}$ in steps of 0.1 while preserving $w_{nc} + w_c = 1$ at all times.

4.4 Results and Discussion

4.4.1 Semantic Head Recognition

Table 4.4 shows 8 × 3 runs, corresponding to the three modes (see Section 4.3.1) tested on the base-AM model and the seven context-feature models. The baseline for mode *dec-1st* is .554 since 55.4% of the first decisions are C. There is no obvious baseline for *dec-all* because the number of decisions depends on the classifier – a classifier whose first decision on a four-word phrase is NC makes one decision, another one may make three. The baseline for *semh* is the tokenizer that always returns the syntactic head; this baseline is .488.

For all modes, the context-feature model that uses sim_{alt} achieves the best result; the accuracies are .692, .703 and .680, respectively. The improvements over the baselines (for *dec-1st* and *semh*) are statistically significant at $p <$ .01 (binomial test, $n = 260$). For *semh*, accuracy for base-AM model is .603; this is significantly better than the .488 baseline ($p < .01$). Accuracy for the base-AM model is significantly lower than the best context-feature model (.680) at $p < .01$ and significantly lower than the worst context-feature model (.653) at $p < .1$. However, the differences between the context-feature models are not significant.

mode	base	cont. feat.	base-AM	additional context feature subsets						
		sim_{alt}	-	●	●	●	●	-	-	-
		sim_{sj1}	-	-	●	–	●	●	-	●
		sim_{sj2}	-	-	-	●	●	-	●	●
dec-1st	.554		.604	.692	.669	.685	.677	.654	.654	.662
dec-all	-		.615	.703	.681	.696	.688	.666	.669	.675
semh	.488		.603	**.680**	.657	.673	.665	.653	.653	.661

Table 4.4: Accuracy for base-AM and context-feature models. A '●' indicates the use of the corresponding context feature (cont. feat.).

type	freq	definition
r_{semh}	92	sem. head correct ($\neq$ synt. head)
r_{synth}	85	sem. head correct ($=$ synt. head)
r^{+}	48	sem. head too long
r^{-}	35	sem. head too short
all	260	

Table 4.5: Distribution of result types

When the semantic head recognizer processes a phrase, there are four possible results. Result r_{semh}: the semantic head is correctly recognized and it is distinct from the syntactic head. Result r_{synth}: the semantic head is correctly recognized and it is identical to the syntactic head. Result r^{+}: the semantic head is not correctly recognized because the cascade was stopped too early, i.e., a compositional modifier that should have been removed was kept. Result r^{-}: the semantic head is not correctly recognized because the cascade was stopped too late, i.e., a modifier causing a non-compositional meaning shift was removed. Table 4.5 shows the distribution of result types. It shows that r^{+} is the more common error: the classifier more often regards compositional relations as non-compositional than vice versa.

Table 4.6 shows the top 20 classifications where the semantic head was not the same as the syntactic head sorted by confidence in descending order. In the third column "phrase ... " we list the candidates with semantic heads in bold. The columns to the right show the predicted semantic head and the feature values. All five errors in the list are of type r^+.

Two r^+ phrases are *schematic view* and *many others*. The two phrases are clearly compositional and the classifier failed even though the context feature points in the direction of compositionality with a value greater than .5. It can be argued that *many others* is a trivial example that does not require complex machinery to be identified as compositional, e.g. by using a stop list. We included it in the analysis since we want to be able to process arbitrary phrases without additional hand-crafted resources.

Another incorrect classification occurs with the phrase *massive star birth*[10] for which *star birth* was annotated as the semantic head. Here we have a case where the peripheral element *massive* does not modify the syntactic head *birth* but *massive star* is itself a complex modifier. In the test set, 5% of the phrases exhibit structural ambiguities of this type. Our system cannot currently deal with this phenomenon.

The remaining r^+ phrases are *peculiar velocity* and *local group*. However, Wikipedia lists both phrases with an individual entry defining the former as *the true velocity of an object, relative to a rest frame*[11] and the latter as *the group of galaxies that includes Earth's galaxy, the Milky Way*[12]. Both definitions provide evidence for non-compositionality since the velocity is not peculiar (as in strange) and the scope of *local* is not clear without further knowledge. Arguably, in these cases our method chose a justifiable semantic head, but the raters disagreed.[13]

Against the background of MWT, it is acceptable to sacrifice recall and only make high-confidence decisions on semantic heads. A tokenizer that reliably detects a subset of MWUs is better than one that recognizes none. However, our attempts to use the sim_{alt} recognizer (bold in Table 4.4) in this way were not successful. Precision is .68 for confidence > .7 and does

[10] i.e. the birth of a massive star, a certain type of star with very high mass

[11] `http://en.wikipedia.org/wiki/Peculiar_velocity` (July 10th, 2012)

[12] `http://en.wikipedia.org/wiki/Local_group` (October 6th, 2012)

[13] Further evidence that *local group* is non-compositional is the fact that one of the domain experts annotated the phrase as non-compositional but was overruled by the other two.

c.	type	phrase (semantic head in bold)	predicted semantic head	am_t	am_f	am_{cp}	sim_{alt}
.99	r_{semh}	**ellipsoidal figure of equilibrium**	ellipsoidal figure of equilibrium	18.03	325	6.23e-01	.219
.99	r_{semh}	**point spread function**	point spread function	95.03	9056	2.33e-01	.529
.99	r^+	massive **star birth**	massive star birth	19.99	402	4.81e-03	.134
.98	r_{semh}	**high angular resolution imaging**	high angular resolution imaging	13.07	179	1.27e-03	.173
.98	r_{semh}	**integral field spectrograph**	integral field spectrograph	24.20	586	4.12e-02	.279
.98	r^+	local **group**	local group	153.54	24759	8.73e-03	.650
.98	r_{semh}	**neutral kaon system**	neutral kaon system	1.38	108	4.17e-03	.171
.97	r_{semh}	**IRAF task**	IRAF task	49.07	2411	2.96e-02	.517
.92	r_{semh}	**easy axis**	easy axis	44.66	2019	2.79e-03	.599
.89	r^+	schematic **view**	schematic view	40.56	1651	8.06e-03	.612
.87	r_{semh}	**differential resistance**	differential resistance	31.71	1034	6.38e-04	.548
.86	r_{semh}	**TiO band**	TiO band	36.84	1372	2.21e-03	.581
.86	r^+	many **others**	many others	97.76	9806	6.54e-03	.708
.86	r_{semh}	**VLBA observation**	VLBA observation	43.95	2004	9.35e-04	.648
.85	r^+	peculiar **velocity**	peculiar velocity	167.63	28689	2.37e-02	.800
.84	r_{semh}	**computation time**	computation time	43.80	1967	1.35e-03	.657
.83	r_{semh}	**Land factor**	Land factor	21.15	453	6.30e-04	.360
.83	r_{semh}	**interference filter**	interference filter	31.44	1002	1.27e-03	.574
.83	r_{semh}	**line formation calculations**	line formation calculations	14.20	203	1.96e-03	.381
.82	r_{semh}	**Wess-Zumino-Witten term**	Wess-Zumino-Witten term	9.60	94	8.12e-05	.291

Table 4.6: The 20 most confident classifications where the prediction is semantic head $\neq$ syntactic head. “c.” = confidence

not exceed .77 for higher confidence values.

To understand this effect, we analyzed the distribution of sim_{alt} scores. Surprisingly, moderate similarity between .4 and .6 is a more reliable indicator for NC than low similarity $< .3$. Our intuition for using distributional semantics in Section 4.2.4 was that low similarity indicates non-compositionality. This does not seem to hold for the lowest similarity values possibly because they are often extreme cases in terms of distribution and frequency and then give rise to unreliable decisions. This means that the context features enhance the overall performance of the classifier, but they are unreliable and do not support the high-confidence decisions we need for MWT.

For comparison, the base-AM model achieves 90% precision at 14% recall with confidence $> .7$ – although it has lower overall accuracy than the sim_{alt} recognizer. We decided to use the AM-only recognizer for the IR experiment because it has more predictable performance.

In summary, the results show that, for the recognition of semantic heads, basic AMs offer a significant improvement over the baseline. We have shown that some wrong decisions are defensible even though the gold standard data suggests otherwise. Context features further increase performance significantly, but surprisingly, they are not of clear benefit for a high-confidence classifier that is targeted towards recognizing a smaller subset of semantic heads with high confidence.

4.4.2 Information Retrieval

Table 4.7 displays retrieval performance of our approach against the baseline and five runs with pseudo MWU. We see a 9.61% improvement in recall over the baseline. MAP and P20 also show improvements. Our approach is better than any of the 5 random runs on all three metrics – the probability of getting such a good result by chance is $\frac{1}{2^5} < .05$, and thus the improvements are statistically significant. On doing a query-wise analysis of AP scores, we find that large improvements over the baseline occur when an MWU aligns with what the user is looking for. The system seems to retrieve more relevant

run	MAP	REC	P20
baseline	0.0663	0.2675	0.1385
real MWU	**0.0718**	**0.2932**	**0.1538**
pseudo MWU_1	0.0664	0.2738	0.1385
pseudo MWU_2	0.0658	0.2717	0.1462
pseudo MWU_3	0.0671	0.2699	0.1477
pseudo MWU_4	0.0681	0.2804	0.1462
pseudo MWU_5	0.0670	0.2720	0.1423

Table 4.7: IR performance without considering MWUs (*baseline*), versus boosting real and pseudo MWUs (*real MWU, pseudo MWU_i*). All metrics are averages over all queries.

documents in that case. e.g. the improvement in MAP is 0.0977 for a query where the information need contains *"articles ... on making tunable vertical cavity surface emitting laser diodes"* and *laser diodes* was one of the MWUs recognized by our system.

On the other hand, a decrease in MAP occurs when phrases unrelated to the information need receive a higher weight. In a query where the user is looking for *"protein-protein interaction, the surface charge distribution of these proteins and how this has been investigated with Electrostatic Force Microscopy"* our system falsely recognized *Force Microscopy* as an MWU (see problem with complex modifiers in Section 4.3.1). Boosting this phrase did not reflect the core information need which is specified as *"The proteins of interest are the Avidin-Biotin and IgG-anti-IgG systems."*

To summarize, we have shown that our recognition system can find MWUs in queries and that boosting the importance of documents containing these MWUs results in an overall increase of performance on all evaluation metrics. Furthermore, our approach offers significant improvements because it yields better results than boosting random phrases.

4.5 Summary

In this chapter, we have presented and evaluated an implementation of MWT. Central to our implementation is the idea of the semantic head which captures the core of non-compositional MWUs. We have cast MWT as the problem of semantic head recognition and created a cascaded classification approach for this purpose. We experimented with different feature types namely association measures and features of contextual similarity. We carried out an intrinsic evaluation of semantic head recognition and an extrinsic evaluation where we integrated semantic head recognition into an information retrieval system.

In the task of semantic head recognition, the models using context features outperformed a basic AM model which in turn outperformed a baseline recognizer. However, the context-feature models gave rise to unpredictable decisions and exhibited less precision than the basic AM model for high-confidence decisions.

We integrated MWU information into an information retrieval experiment by way of recognizing MWUs in queries and boosting the importance of documents containing these MWUs. The impact of boosting MWUs is positive, increasing retrieval performance on all metrics and beating a number of random baseline boosting approaches.

The MWT implementation we have presented in this thesis focuses on non-compositionality and the semantic head. It does not explicitly consider other properties that could also be helpful for identifying MWUs such as non-substitutability, non-modifiability or asymmetry. For this experiment, we chose this approach to explore the impact of different measures of contextual similarity on non-compositionality detection. We see this approach and the insights into non-compositionality detection we gained in the experiment as a valuable building block for future MWT implementations that incorporate more features addressing different properties of MWUs.

Chapter 5

Conclusions and Future Research

This chapter summarizes the main results of this thesis. Additionally, we give starting points for future research.

5.1 Contributions

Communicating with computers in unrestricted natural language is still an unfulfilled goal. In particular, handling of collocations in NLP is not a solved problem. In this thesis, we presented research aimed at improving automatic handling of collocations. Focusing on MWUs, our basic assumption was that automatic processing has to approach the problem by addressing MWUs' idiosyncratic properties. The main idea presented in this thesis is multi-word tokenization (MWT) as part of preprocessing for NLP. MWT is a supervised classification approach to recognizing MWUs whose features are targeted at idiosyncratic properties of MWUs.

In many NLP tasks, the current standard approach of single-word tokenization (which ignores MWUs) causes errors that propagate to higher-level NLP tasks and reduce the overall quality of NLP applications. With MWT, these errors could be avoided.

The contributions in this thesis approach MWT from two angles, covering one theoretical and one practical aspect.

5.1.1 Asymmetric Association

The theoretical contribution is the exploration of asymmetry in syntagmatic word combinations and MWUs in particular. In NLP and psycholinguistics, research into word associations and elicitation experiments are traditionally focused on paradigmatic associations. Creating a novel experiment setup, we collected human syntagmatic word associations. Based on the collected data, we found asymmetry in human syntagmatic production and strong asymmetry to be indicative of MWUs. Additionally, we created corpus-based measures of asymmetry based on association measures. Classic association measures are a mainstay of collocation research but they cannot capture asymmetry in word combinations. We showed that with the new measures we created, it is possible to predict the asymmetry found in human production based on corpus-based data.

5.1.2 Multi-Word Tokenization

The practical part deals with an implementation of MWT. For this implementation, we cast the MWU decision problem as semantic head recognition. The semantic head of a candidate phrase is its non-compositional core and recognizing the semantic head is equivalent to recognizing the MWU. For the semantic head recognizer we experimented with two kinds of features: association scores aimed at general lexical association serving as a baseline and measures of contextual similarity specifically designed for non-compositionality detection. We have shown that models using context features significantly improve upon the baseline model. In an experiment that incorporated an MWU recognizer into information retrieval system, we were able to show that using MWU information at query time improves overall retrieval performance.

5.2 Future Avenues of Research

5.2.1 Asymmetric Association

We suggest two directions for work on asymmetric association measures. First, we present possible improvements to the measures of syntagmatic asymmetry we introduced in Section 3.3.2. Second, we propose large-scale application of a paradigmatic measure of asymmetry for the purpose of hypernymy mining.

Improved asymmetry prediction We present two improvements for asymmetry prediction from corpus data. First, we suggest using ensemble learning by taking advantage of multiple asymmetry models with different underlying characteristics. Second, we suggest regression models for the prediction of absolute forward and backward scores, emulating asymmetry scores based on human judgements.

We have seen that different association measures give rise to different rank measures. All rank measures included in our study are based on association measures derived from statistical significance tests, which are known to correlate strongly with co-occurrence frequency. Hence, some of the asymmetric measures make similar decisions. Rankings obtained from measures of effect size such as PMI, for example, may provide entirely new perspectives on the right- and left-predictiveness of syntagmatic combinations. With different rank measures reflecting different characteristics of the underlying association measures, asymmetry prediction could be further improved and made more reliable. Combining multiple perspectives for asymmetry prediction can be achieved with ensemble methods (Friedman et al., 2001, Ch. 16). With ensemble methods, classification decisions of multiple models are pooled and a final decision is made, e.g. by majority vote.

Our analysis of forward and backward scores computed from human judgements has shown that strong asymmetry (defined as either forward or backward score $> .5$) is an indicator for a phrase being an MWU. In Section 5.2.2, we propose asymmetry scores as features for an improved MWT system. In order to predict absolute association strength from corpus data, a regression model is required. The prediction of direction of asymmetry we used for evaluation in Section 3.5.4 was based directly on ranks.

Unsupervised hypernymy mining In the context of MWUs, we have focused on syntagmatic asymmetric association. We discussed how Weeds (2002) linked asymmetric paradigmatic association to the hypernymy relation. (Hearst, 1992) proposed an automatic approach to hypernymy mining. This approach relies on lexico-syntactic patterns for extraction. In a test study, Weeds showed that the hypernymy relation between two words can be automatically discovered using the asymmetric skew divergence measure. Carried out on a larger scale, this approach could be used for automatic hypernymy mining. The extracted relationships can be used to create or enrich lexical resources such as ontologies and conceptual hierarchies. The approach based on asymmetric measures could be run in an unsupervised fashion, which means that it does not require any lexico-syntactic patterns to discover new relationships.

5.2.2 Multi-Word Tokenization

In this section, we discuss improvements that could be incorporated into a future implementation of MWT. We discuss context-dependent classification and new feature types. Finally, we propose to integrate all improvements by modeling MWT as a sequence labeling problem.

Context dependence One important aspect of MWT that we have not addressed is context dependence. Traditionally, within the scope of collocation extraction, the focus is on creating a list of collocations or collocation candidates that would typically be passed on to a lexicographer or domain expert for further inspection. In this scenario there is a *global* decision about the status of a phrase – it either is an MWU or it is not. This approach is called *type-based* classification. In a different approach, the decision whether or not a candidate should be labeled as an MWU is *local* and has to be made for every instance of the candidate with potentially different outcomes. This approach is called *token-based* classification.

The MWT implementation presented in Chapter 4 is type-based. For future work on MWT, we propose a token-based approach. For MWT as part of an automated preprocessing pipeline, a token-based approach would be more appropriate because there are phrases that have to be labeled as MWUs in some contexts but not in others. The distinction is often motivated by literal and non-literal uses of the same phrase. Consider the following uses of *red tape* in the BNC.

(5.1) There are regulations, laws and *red tape.*

(5.2) Hugh Dalton was about to open a factory by snipping the *red tape* with a pair of ornamental scissors.

In (5.1), we want *red tape* to be tokenized as a unit because it is used non-literally, i.e. in this context the phrase exhibits non-compositionality. The use in (5.2), however, is literal and we want to tokenize *red* and *tape* separately.

There are previous works that deal with token-based collocation classification. These studies are concerned with MWEs that have literal and non-literal uses such as *get the sack* or *play with fire.* Some of the features these approaches used are based on distributional semantics (Katz and Giesbrecht, 2006; Cook et al., 2007) and lexical cohesion (Sporleder and Li, 2009). The features based on distributional semantics are comparable to the context features we used in Section 4.2. Sporleder and Li (2009) created a cohesion graph by representing the MWE and its context as nodes and connecting them with weighted edges based on the semantic relatedness between the nodes. They measure the drop of overall similarity in the cohesion graph when removing the MWE nodes. For example, if *play with fire* appears in the context of a barbecue with the words *coal*, *grill* and *smoke*, removing *play* and *fire* from the graph results in less overall similarity, or cohesion, in the graph. If the phrase appears in the context of, say, politics, removing the same nodes will have less impact on overall similarity.

See below how we propose to use these methods of token-based MWE classification for context-dependent MWT.

MWT as sequence labeling The best-performing implementations for common NLP tasks such as named-entity recognition[1], POS tagging (Brill, 2000) and NP chunking[2] rely on learning algorithms summarized under the term of sequence labeling. Commonly applied in a supervised setting, sequence-labeling methods are algorithms that assign class labels to a sequence of outputs. Such algorithms are suited for NLP problems because of the sequential nature of language and the power of contexts for ambiguity resolution. In POS tagging, for example, tags (labels) have to be assigned to a sequence of words (outputs) and the label for the current word often depends on the previous word (e.g. *bank* is a noun in *the bank* and a verb in *to bank*).

The basic workings of sequence labeling are as follows: we observe words and try to discover the most likely sequence of labels that generated the output. Sequence labeling models rely on the Markov property, which means that the current label prediction depends only on (the properties of) the current and the immediately preceding labels and words.

What makes sequence labeling methods attractive for NLP is that with state-of-the-art algorithms we can encode any linguistic knowledge about the current word as features for the classification decision and that context (i.e. the previous word and label) is taken into account because of the Markov property. Furthermore, in this kind of model, classification decisions are not made in isolation for each word but alternative label sequences and their probabilities are stored until the last word has been processed. The most probable label sequence is then determined in the decoding step using dynamic programming.

The best known sequence-labeling method used in NLP is the Hidden Markov model which has been popularized by its usefulness for POS tagging (DeRose, 1988). Other well-known methods are Maximum-Entropy Markov-Models which were used for POS tagging (Ratnaparkhi, 1996) and the current state-of-the-art sequence labeling method conditional random fields (CRFs, Lafferty et al., 2001) which were used, for example, for NER (Finkel et al., 2005) and NP chunking (Sha and Pereira, 2003).

[1] See the CoNLL shared task on NER (Tjong Kim Sang and De Meulder, 2003).

[2] See the CoNLL shared task on NP chunking (Tjong Kim Sang and Buchholz, 2000).

MWT is related to these NLP tasks but differs in certain aspects: it differs from NER in that it does not try recognize named entities (e.g. places, people, institutions) but generic noun phrases that are not named-entities. Chunking finds constituents whereas MWT tries to find MWUs within an NP constituent. For example, chunking marks *the best hot dog ever* as an NP but MWT identifies the sub NP *hot dog* as a unit. The relation between MWT and POS tagging is two-fold. On the one hand, MWT can use POS information to identify candidates and for feature engineering. On the other hand, POS tagging can be performed *after* MWT assigning POS tags to MWUs. The choice depends on the available tools and requirements of a particular application.

For future work, we propose an implementation of MWT as a sequence labeling task.[3] This way, we can integrate token-based classification and any number of new types of features.

The general idea is that the sequence classifier labels a sequence of words (e.g. a sentence) into three classes of words following the IOB labeling scheme (Jurafsky and Martin, 2008, p.487): MWU-I, MWU-O and MWU-B. The three labels represent words that are inside MWUs, outside of MWUs or mark the beginning of an MWU, respectively.

Below, we present the features we envision for the sequence labeling approach to MWT. For each feature explanation, recall that we are looking at the current and the previous word and that we have access to other linguistic information such as the words in the current sentence or document. We include features based on the features that were used in our MWT implementation presented in Chapter 4.

word features: These are features that encode information about the current word, e.g. its most likely part of speech, prefixes, suffixes, word shape, etc.

[3] Constant and Sigogne (2011) present similar ideas for French in the context of POS tagging. Their approach extends a POS tagger with an MWU recognition module which uses external resources such as MWU dictionaries. In contrast, we suggest an approach that focuses on engineering features that address MWU properties for recognition instead of relying on dictionaries.

label feature: This feature encodes whether or not the previous word was classified as belonging to an MWU.

contextual similarity: These features address non-compositionality and non-modifiability. For example we try to detect a meaning shift when removing words or replacing words with similar words.

lexical association: These features encode general lexical association by means of association scores between the current and the previous word.

asymmetric association: These features encode asymmetric association by means of forward and backward scores for the current and the previous words.

lexical fixedness: These features measure lexical fixedness to address non-modifiability. Fazly et al. (2009) make use of an MWE's *canonical form*. An expression appearing in its canonical form is assumed to be idiomatic (Riehemann, 2001, p. 34) and more frequent than other forms. They identify the canonical form of an MWE with measures of syntactic and lexical fixedness. For example, a lexical feature is *type of determiner*: with an indefinite determiner (*he got a sack*), the idiomatic phrase *to get the sack* becomes literal. Like MWEs, MWUs, can have canonical forms. For example, the non-literal meaning of *red tape* is always singular and does not appear with an indefinite article.

lexical cohesion: This feature address token-based classification. It accesses words in the context of the current word, e.g. the current sentence or document.

Cohesion-based features from MWE classification could be adapted for token-based MWU classification. For example, the words *snipping* and *scissors* occurring in (5.2), are clues for a literal use of *red tape*.

heuristics: These are various features based on heuristics or formatting clues such as quotation marks or italics. Additionally, we include orthographic clues, e.g. the frequent spellings *website* and *web-site* are an indicator that *web site* is an MWU.

Appendix A

Comparison of Distributional Models

A.1 Introduction

In this chapter, we present a comparative study of two models of distributional semantics, namely graph-based similarity using the SimRank similarity measure and cosine similarity in a word space model. The goal of this comparison is to decide which model to use for non-compositionality detection in MWT. For MWU decision, it is important to detect meaning shifts between words and phrases. For this purpose, we need a reliable tool for meaning representation. We compare the similarity measures by evaluating them as measures of semantic relatedness. We assess a measure's performance by predicting for a test word the related words and manually sorting the predictions into a set of predefined categories of semantic relations. We explore several types of filters and weighting strategies that impact similarity computation. In particular, the questions that interest us are: (i) which model performs better? (ii) what influence do weight adjustments and filters have? (iii) what are the advantages and disadvantages of each model?

The chapter is structured as follows. Section A.2 introduces the SimRank computation we used in our experiments. The word space model is as defined in Section 2.2.2. In Section A.3, we describe the data, filters and weighting strategies we used as well as the concrete experiments. In Section A.4, we present and discuss the results. In Section A.5 we examine higher-order similarity, a feature of SimRank, and its impact on similarity computation.

A.2 SimRank Modification

Recall that the similarity S_{ij} between two nodes i and j with neighbors $N(i)$ and $N(j)$ is given through the SimRank measure by:

$$S_{ij} = \frac{c}{|N(i)|\ |N(j)|} \sum_{k \in N(i), l \in N(j)} S_{kl}$$

Preliminary experiments have shown that SimRank favors nodes with few neighbors which means that rare (low-degree) words are often assigned the highest similarity values for a given test word. To counter this effect, Laws et al. (2010) introduced a modified normalization step, square root normalization (SQN). In the original SimRank equation, the number of a node's neighbors is used directly for scaling. The modified scaling aims at lessening the influence of low-degree nodes. By incorporating the square root of the number of neighbors, the punishment of high-degree is dampened. This is reflected in the modified definition of SimRank:

$$S_{ij} = \frac{c}{f(|N(i)|)f(|N(j)|)} \sum_{k \in N(i), l \in N(j)} S_{kl}$$

with $f(n) = \sqrt{n} * \sqrt{\max_k(|N(k)|)}$. The goal of the modified weighting scheme is to favor words with more neighbors (i.e. frequent words). For rare words, which have a small number of neighbors, f grows quickly, while returning values close to the linear term for frequent words. This guarantees that rare words have less influence on final similarity scores.

A.3 Evaluation Setup

The evaluation method is to let both models find semantically related words to a given test word. The discovered words will be manually classified into a set of categories that represent certain semantic relations. The idea behind this method is that a measure that finds words that are related or similar to a test word consistently, is a reliable and robust measure to serve as a basis for non-compositionality detection.

We explored the effects of different filters and weights on the quality of word similarities. The following sections describe these filters and weights and give an overview of the concrete experiment runs that we carried out. Additionally, we briefly describe our data and the test set.

A.3.1 Data and Evaluation

We compare two distributional models which means that meaning will be computed from the contexts of words in a corpus. For both models, we define context as the grammatical relation of nouns appearing as direct objects of verbs. The verb-object data were extracted from an English Wikipedia dump of October 2008. We parsed all articles with BitPar (Schmid, 2004) which resulted in 11.8M verb-object pairs. We used a lemmatizer to improve the quality of the extracted relations.

We use a test set of 65 nouns taken from a 100-word test set previously used for bilingual lexicon extraction (Rapp, 1999). We removed adjectives and verbs from the set. In each experiment, we generated a list of the ten most similar words (target words) for each test word. We carried out manual evaluation. For every test word, the author decided if there is a semantic relation between the test word and each corresponding target word.

Target words were manually classified into one of the following classes of semantic relations to the test word: hypernym (R), hyponym (H), co-hyponym (C), synonym (S) or other (O) where class (O) is the default class for any related target word that is not covered by any of the other classes (e.g. *hand–finger* or *man–manhood*). The remaining target words are labeled unrelated (U). This classification follows Michelbacher et al. (2010).

A.3.2 Weights and Filters

When considering co-occurrence data, the most basic information we have is the frequency with which a pair occurred. It can be used for edge weights (graph) and weighted dimensions (word space) directly. A common approach, however, is to use statistical measures to compute an *association score* between the components of a word pair (see Section 2.1). The assumption is that association measures such as the log-likelihood measure (Dunning, 1993) will assist in weeding out insignificant co-occurrences. Additionally, we examine the effect of using the logarithms of edge weights instead of the actual weight. This step is carried out in order to weaken the impact of very frequent combinations.

Another intuitive filter method is to remove nodes with a degree $d < n$, assuming that a node that has few neighbors is less important than one with many neighbors. Considering the aforementioned tendency of SimRank to favor low-degree nodes, degree-based filtering seems to be an attractive approach. To summarize, we considered the following filters and weight adjustments:

- log-likelihood scores instead of frequency as weights
- logarithm of weights instead of plain weights
- degree-based node filter

A.3.3 Experiments

Table A.1 shows an overview of the experiments that we carried out. The *name* column contains a short identifier that summarizes an experiment's properties and which will be used to refer to the corresponding experiment throughout the text. The *similarity* column shows which similarity method was used. The *degree* column shows which nodes were removed because they had too few neighbors or "–" if none were removed. Further, the *weight* column contains the weighting strategy that was applied: frequency (f), log-

name	similarity	degree	weight	SQN
sr-sqn-d02	SimRank	$d < 2$	$log(G^2)$	●
sr-sqn-d05	SimRank	$d < 5$	$log(G^2)$	●
sr-sqn-d10	SimRank	$d < 10$	$log(G^2)$	●
sr-sqn-d15	SimRank	$d < 15$	$log(G^2)$	●
sr-sqn-d20	SimRank	$d < 20$	$log(G^2)$	●
cos-f	cosine	–	f	–
cos-logf	cosine	–	$log(f)$	–
cos-log-lik	cosine	–	G^2	–
cos-log-log	cosine	–	$log(G^2)$	–
cos-log-log-d20	cosine	$d < 20$	$log(G^2)$	–
sr-d15-f	SimRank	$d < 15$	f	–
sr-d15-logf	SimRank	$d < 15$	$log(f)$	–
sr-d15-log-lik	SimRank	$d < 15$	G^2	–
sr-d15-log-log	SimRank	$d < 15$	$log(G^2)$	–

Table A.1: Experiment names and descriptions

likelihood (G^2) or a logarithm thereof.[1] Finally, a bullet symbol in the *SQN* column means that square root normalization was turned on in graph-based experiments.

We carried out five main experiments for each similarity method. The first batch tests degree filters of 2, 5, 10, 15 and 20 in conjunction with SQN using SimRank. Here, we try to study the effect that the degree filter has on quality leaving all other settings untouched. The next five experiments deal with cosine similarity. The first four focus on different weighting strategies. For comparison, we ran a fifth cosine experiment with a degree filter of $d < 20$.[2] Finally, the last group of SimRank experiments examines weighting strategies in a graph-based setting without SQN. Here, we chose a degree filter of $d < 15$, the value that proved best in the first group of experiments.

[1] Pairs with G^2 scores ≤ 0 were discarded.

[2] The degree filter was defined with nodes and links in mind but the concept can be

name	(S)	(R)	(H)	(C)	(O)	total	(U)	n/a	total perc.
sr-d02sqn-log-log	9	4	71	95	49	228	422	0	35.07%
sr-d05sqn-log-log	9	6	75	137	85	312	338	0	48.0%
sr-d10sqn-log-log	10	8	71	155	94	338	312	2	52.0%
sr-d15sqn-log-log	11	12	74	156	107	360	290	2	55.38%
sr-d20sqn-log-log	12	16	65	145	107	345	305	5	53.07%
cos-f	6	17	47	95	75	240	410	0	36.92%
cos-logf	6	27	24	148	162	367	283	0	56.46%
cos-log-lik	4	14	48	80	47	193	457	0	29.69%
cos-log-log	9	22	63	175	136	405	245	0	62.30%
cos-log-log-d20	11	30	55	156	139	391	259	5	60.15%
sr-d15-f	3	2	9	18	24	56	594	2	8.61%
sr-d15-logf	9	4	29	79	44	165	485	2	25.38%
sr-d15-log-lik	4	6	12	34	47	103	547	2	15.84%
sr-d15-log-log	9	8	46	120	71	254	396	2	39.07%

Table A.2: Results with different weighting schemes and filters

A.4 Results and Discussion

See Figure A.1 for a graphical representation of the results which facilitates interpretation. Table A.2 show the results in tabular, more detailed form. The column *n/a* represents the number of test words that were missing in the corresponding experiment. Missing test words occur due to node filtering. The column *total perc.* specifies the total percentage of related words among the target words (i.e. target words that were not classified as (U)).

The degree filter gradually improves performance from *sr-d02sqn-log-log* over *sr-d05sqn-log-log* and *sr-d10sqn-log-log* to *sr-d15sqn-log-log* (first four bars on the left). See Table A.3 for an example. Here, most of the target words are co-hyponyms of *thief* with the common hypernym *criminal*. The performance increase from *sr-d02sqn-log-log* to *sr-d15sqn-log-log* is statisti-

transferred to the word space model by simply regarding the number of unique verbs that the noun appeared with as links.

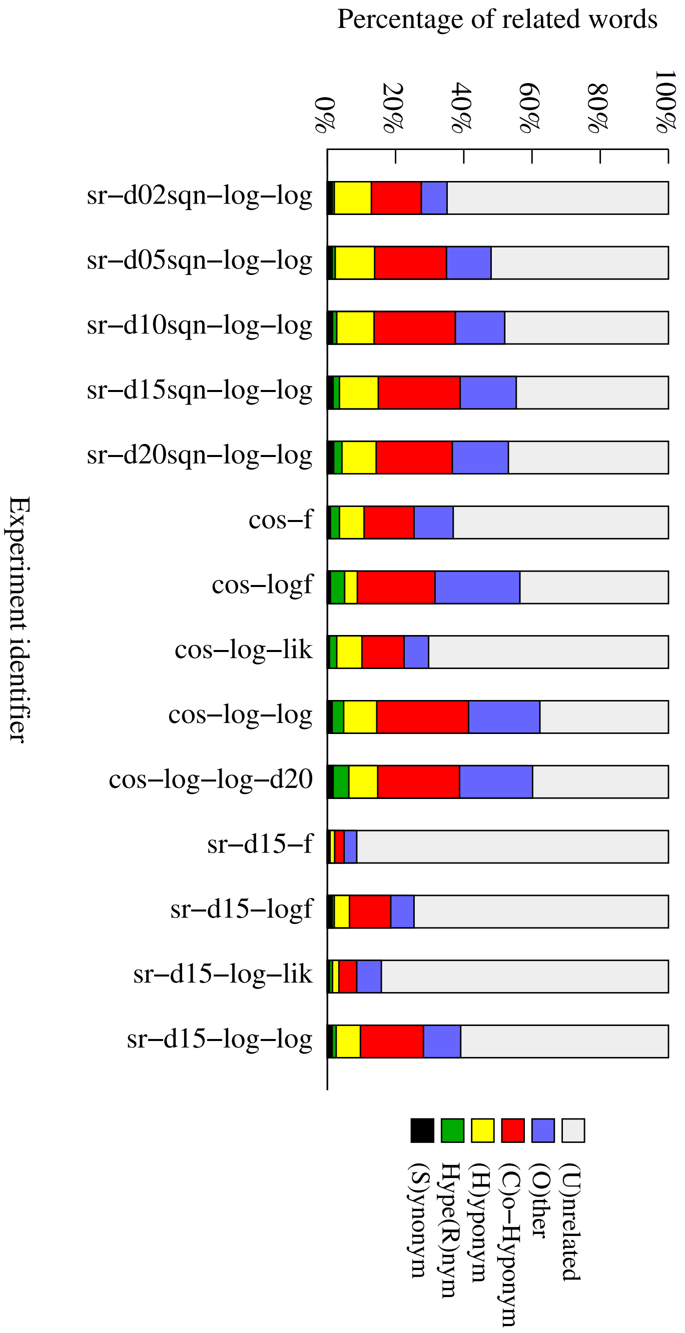


Figure A.1: Evaluation results according to each experiment

test word	**thief**				
filter	$d < 2$	$d < 5$	$d < 10$	$d < 15$	$d < 20$
	robber^C	robber^C	robber^C	robber^C	robber^C
	thrall	kidnapper^C	kidnapper^C	kidnapper^C	criminal^R
	kidnapper^C	burglar^C	burglar^C	burglar^C	murderer^C
	burglar^C	speeder	murderer^C	murderer^C	killer^C
	illusionist	murderer^C	crook^C	crook^C	kidnapper ^{C}
	beautician	sorcerer^C	smuggler^C	smuggler^C	terrorist^C
	savior	killer^C	sorcerer^C	sorcerer^C	burglar^C
	vandal^C	crook^C	killer^C	killer^C	villain^R
	dervish	smuggler^C	culprit^R	culprit^R	pirate^C
	hurdler	rustler^H	gunman^C	gunman^C	assassin^C
# related	4	8	9	9	9

Table A.3: Effect of the degree filter for test word *thief* (semantic relations to the target words displayed with superscripts)

cally significant[3] ($\alpha = .01$). At degree filter $d < 20$ (*sr-d02sqn-log-log*), loss of coverage is starting to take effect manifested in five missing test words. We explain the performance drop (non-significant, $\alpha = .05$) from *sr-d15sqn-log-log* to *sr-d20sqn-log-log*, the most aggressive degree filter we used, by this loss of coverage.

With regard to the weighting strategies, there are three findings. First, the application of the logarithm to the weights improves performance significantly (from *cos-f* to *cos-logf*, from *cos-log-lik* to *cos-log-log*, from *sr-d15-f* to *sr-d15-logf*, from *sr-d15-log-lik* to *sr-d15-log-log*, $\alpha = .01$). Second, the combination of log-likelihood weights and logarithmic dampening yields the best results (*cos-log-log* and *sr-d15-log-log*). Third, logarithmic frequency weights perform significantly better than plain log-likelihood scores (*cos-logf* vs *cos-log-lik* and *sr-d15-logf* vs *sr-d15-log-lik*, $\alpha = .01$).

The most important finding, however, is that cosine similarity outperforms graph-based similarity for all different weighting strategies. The best

[3] All significance test in this section are binomial tests.

overall result, *cos-log-log* with 62.3% related target words over the whole test set, yields significantly better results than *sr-d15sqn-log-log*, the best graph-based result with 55.38% ($\alpha = .01$). Note that *cos-log-log* does not use any word filter. We ran an experiment, *cos-log-log-d20*, which uses a degree filter $d < 20$ and cosine similarity. There is a performance drop of 2.15 percentage points (non-significant $alpha = .05$) which we, again, attribute to the loss of coverage that comes with this filter. It seems that the computation of cosine similarity is not influenced by the degree filter.

We also found that SQN has a positive effect on SimRank performance. The improvement of *sr-d15sqn-log-log* (55.38% related word) which uses SQN – over *sr-d15-log-log* (39.07%) which does not – is significant ($\alpha = .01$).

Another effect of the degree filter is that with more aggressive filtering, related words become more concrete. This can be explained by the fact that with less filtering, more low-degree words are available which are in turn favored by SimRank. Consider Table A.4. Here, the annotator chose to only annotate *scorpion* as a co-hyponym of *spider* (both are arachnids) and *insect* as related in another way. The rest of the target words are almost all animals or classes of animals. With a low filter ($d < 2$), the target words are in fact almost all animals, but rather specific ones (*shrew*, *warbler*, *partridge* and *kingfisher*) or animals of microscopic size (*amphipod* and *rotifer*). Moving on to stricter filters, the target words are more familiar and feasible. With $d < 20$, results include clear animal classes like (*rodent*, *reptile*, *insect*, *invertebrate*) as well as familiar animals (*frog*, *shark*, etc.).

It shows that with increasing degree filter, the target words become more familiar (i.e. more frequent). In contrast, *cos-log-log* achieves the same without the use of any filtering which suggests that the computation of cosine similarity is robust against low-degree (i.e. rare) words.

For some test words, none of the experiments produced any feasible results, e.g. *bath* shown in Table A.5. We only show *sr-d15sqn-log-log* and *cos-log-log* to illustrate the problem. It is striking, that none of the target words is even remotely related to *bath.* On closer examination, however, *bath* and many of the target words do have something in common: they appear as direct objects of the verb *to take.* Among these there are light verb construc-

		spider			
$d < 2$	$d < 5$	$d < 10$	$d < 15$	$d < 20$	*cos-log-log*
shrew	shrew	rodent	rodent	rodent	rodent
warbler	earthworm	crustacean	lizard	lizard	lizard
earthworm	crustacean	squirrel	crabs	reptile	crustacean
rotifer	mackerel	lizard	reptile	invertebrate	shark
diatom	rodent	scorpion^C	crab	frog	reptile
equestrian	jaguar	crabs	invertebrate	moth	crabs
gazelle	otters	squid	toad	shark	shrew
partridge	squirrel	amphibian	caterpillar	crocodile	scorpion^C
kingfisher	antelope	rattlesnake	eel	boar	frog
amphipod	lizard	fern	frog	insect^O	squirrel
0	0	1C	0	1O	1C

Table A.4: Effect of the degree filter for test word *spider* (semantic relations to the target words displayed with superscripts)

tions (*to take a stroll*) or idioms (e.g. *to take umbrage* or *to take a shine*). In cases like this, the abundance of constructions involving *take* as a light verb, obscure other, possibly useful verb-object relations, e.g. *to take a shower*.

See Table A.6 for the distribution of classes among the related words for the *sr-d20-sqn-log-log* run (exemplary choice). It is striking that 42% of the target words are co-hyponyms. We attribute this to the fact that co-hyponymy is not as restrictive a relation as hyponymy, hypernymy and synonymy which only constitute 19%, 5% and 3%, respectively. For the latter three relations, there are simply less potential words in the data. We observed no significant deviance of class distribution between SimRank and cosine similarity or between SimRank with and without SQN.

A.5 Higher Order Similarity

An attractive characteristic of the graph-based model is that it enables higher-order similarity. In this section, we investigate the impact this feature

bath	
sr-d15sqn-log-log	*cos-log-log*
shine	stroll
precedence	umbrage
delight	stead
shelter	Changchun
vacation	intoxicant
precaution	conn
subway	cognizance
helm	pratfall
hermitage	layover
oven	pot-shot

Table A.5: Selected results for test word *bath*

synonym (S)	3%
hypernym (R)	5%
hyponym (H)	19%
co-hyponym (C)	42%
other (O)	31%

Table A.6: Distribution of classes among related words for the *sr-d20-sqn-log-log* run.

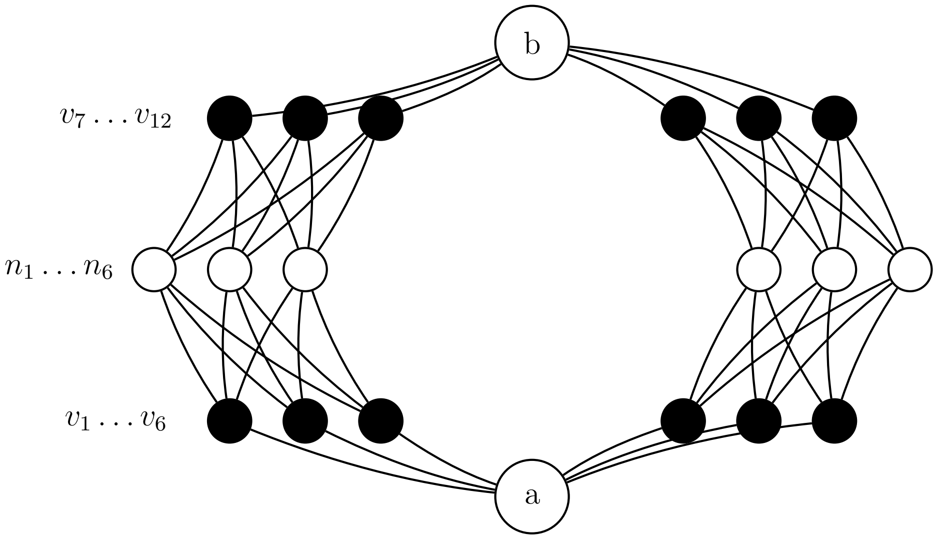


Figure A.2: Hypothetical graph demonstrating higher-order similarity with SimRank

has on similarity computations.

In case of vector similarity measures like cosine similarity, we compute the similarity between two words based on their co-occurrence vectors. If the two words do not share any first-order co-occurrences, i.e. two nouns are never the direct object of the same verb, the cosine similarity of the two words is 0. This is not the case with the SimRank measure. Here, two nodes whose sets of neighbors are disjoint can have a similarity value greater than 0. This is a consequence of SimRank's recursive nature that spreads similarity with each iteration. Figure A.2 depicts a hypothetical example of two nouns, a and b, that are related in that sense that the verbs that occur with a ($v_1 \ldots v_6$) and the verbs that occur with b ($v_7 \ldots v_{12}$) co-occur with a number of different nouns ($n_1 \ldots n_6$) but not with a or b.

Table A.7 shows the progression of the similarity values between a and the rest of the nodes in the graph, where n_i represents $n_1 \ldots n_6$.[4] The similarity

[4] Due to symmetry, the similarity values between $n_1 \ldots n_6$ and a (and b) are the same.

similarity	iteration					
	1	2	3	4	5	6
$S(a, n_i)$	.06	.13	.16	.19	.20	.22
$S(a, b)$	–	.06	.09	.12	.14	.16

Table A.7: Progression of higher order similarity values with SimRank (no SQN)

similarity	iteration					
	1	2	3	4	5	6
$S(a, n_i)$	.01	.003	.002	.001	.001	.001
$S(a, b)$	–	1×10^4	5.8×10^{-5}	4.9×10^{-5}	4.7×10^{-5}	4.7×10^{-5}

Table A.8: Progression of higher order similarity values with SimRank and SQN

between a and n_i starts growing from the very first iteration whereas the one between a and b is still 0 at this point. Over time, the higher order similarity climbs up to almost three quarters of the magnitude of the similarity based on direct co-occurrence.

A.5.1 The effect of SQN on higher order similarity

The application of SQN has an effect on the magnitude of the similarity scores and consequently on higher order similarity. Table A.8 shows the similarities for the hypothetical graph with SQN turned on. Due to its modified denominator, the scaling factor becomes smaller. This leads to smaller similarity values in general and higher order similarities becoming vanishingly small.

Table A.9 shows the test word *eagle* together with the top 10 target words after the first two and last two iterations of the experiment *sr-d15-log-log*, an experiment without SQN. It can be seen that here higher-order similarities have no influence whatsoever. The top ten target words after 6 iterations

eagle			
iteration 1	*iteration 2*	*iteration 5*	*iteration 6*
crescent	crescent	crescent	crescent
lily	penguin	penguin	penguin
penguin	lily	lily	lily
falcon	falcon	falcon	falcon
owl	owl	owl	owl
hawk	hawk	hawk	hawk
leopard	leopard	leopard	leopard
ornamentation	mermaid	mermaid	mermaid
mermaid	ornamentation	dove	dove
dove	dove	ornamentation	ornamentation

Table A.9: Progression of rank order with *sr-d15-log-log* (no SQN)

are exactly the same top ten words that were found after the first iteration. The only difference is a difference in the ordering of the words. We found the same to be true for the rest of the test set. We conclude that theoretically, SimRank does have the possibility to draw on higher order relationships but this feature seems to have no impact on real world calculations.

A.6 Summary

We presented a study comparing the semantic relatedness of nouns on the grounds of two different underlying similarity measures, the graph-based SimRank and the vector-based cosine similarity. To pick up on the questions posed in the introduction:

Which model performs better? The best overall performance was achieved with cosine similarity with a score of 62.30% related target words over 55.38% for the best SimRank result.

Do filters and weighting influence the quality of the results? The degree filter turned out to be a decisive improvement for the graph-based model up to the point where loss of coverage sets in. In terms of weighting strategies, the key to a huge performance increase in both models is the combination of log-likelihood weights and logarithmic weight dampening. Furthermore, square root normalization, a modification of degree-based scaling for SimRank, proved to be essential in making SimRank results feasible at all.

What are the advantages and disadvantages of each model? Cosine similarity has the advantage that it seems less susceptible to the influence of rare words. The aforementioned top score is reached without any frequency or degree filter. SimRank, on the other hand, struggles with low-degree nodes, over-emphasizing their importance when run with a small degree filter. SimRank offers the possibility to incorporate higher order relationships. While this feature proved effective in a hypothetical setting, it had no impact in a real world experiment. Furthermore, the influence of higher order neighbors is virtually neutralized by square root normalization.

In the light of the results of the comparison, we decided to favor the word space model over the graph-based implementation of distributional semantics.

Appendix B

Experiment Instructions

In the next two sections, we include the instructions that were given to (i) the subjects participating in the asymmetric association study and (ii) the three domain experts that marked semantic heads in phrases. The first section contains the instructions for an online experiment, where spontaneous association was required. These instructions are rather high-level because no deeper linguistic understanding of asymmetry was required to complete the experiment. The instructions for the domain experts, on the other hand, are longer and more detailed. In that scenario, there was more work for individual annotators. The subjects were domain experts but not familiar with NLP and the phenomenon of non-compositionality. We included an introduction to semantic heads and a number of examples because our goal was to ensure that the domain experts understood the task and the linguistic background thoroughly. The example phrases and contexts are taken from the iSearch corpus. Note that our annotation procedure collected graded non-compositionality judgements. We collected judgements this way for possible future experiments on compositionality grading but did not use the graded data for the experiment in this thesis.[1] For MWT, a binary decision between MWU or non-MWU has to be made. For an introduction to the task of compositionality grading, see Biemann and Giesbrecht (2011)

[1] Grade 1 was chosen to represent compositionality (C) and higher grades were chosen for non-compositionality (NC).

B.1 Asymmetric Association in Multi-Word Units

Please read the instructions carefully. Since this is an experiment about English it is vital that you only continue if you are a **native speaker of English**. Thank you.

What you have to do for this experiment is to type in words. You'll be presented a word pair where either the first or the second word has been blanked out. Your task is to fill in the blanks with as many words as you can think of.

It is important that you only give words that would appear **right after** or **right before** the displayed word in normal speech (depending on the position of the blank line). The experiment is **not** a *Free Association* experiment. In a *Free Association* experiment you are presented a cue word and the goal is to give words that come to your mind after seeing the cue word. For example *boy* → *girl* or *food* → *drink*. This is not what this experiment is about. There may be overlaps between the two kinds of experiments but here the goal is to give answers so that the blank line and the word that is already there make up a fixed expression, for example, *boy* → *scout* or *food* → *court*.

Try to think of as many words as possible for each answer. It often helps to imagine the words in several different contexts or to say it out loud. Let your mind wander – as long as you are sure that the answers you give are actually being used by English native speakers. Give all the words that you can think of but try not to spend too much time on a single question. If you can't think of anything, press the 'next' button.

B.2 Semantic Head Recognition

In natural language processing (NLP), **noun phrases** play a central role. Noun phrases are groups of words like **the dog**, **dogs**, **the small dog**, **that hot dog** or even **the very old dog that never barks**. Names, e.g. **Albert Einstein** are also noun phrases. For successful NLP it is important to know which words of a noun phrase represent the main meaning of the phrase. For more information on noun phrases, see `http://en.wikipedia.org/wiki/Noun_phrase`.

Problem definition

The problem with noun phrases and meaning is how to tell which parts of the noun phrase are important for determining its meaning. For example:

(B.1) My neighbors have a **brown dog**.

(B.2) I want a **new dog**.

(B.3) He invented the **hot dog**.

Here, all noun phrases have the same **syntactic head**: *dog.* The syntactic head of a noun phrase is its most important word, all other words in the phrase (brown, new, hot) **modify** the head. These **modifiers** give more information about the syntactic head.

In sentences (B.1) and (B.2), the modifiers introduce a specification of the head that does not affect the meaning of the whole noun phrase very much. A *brown dog* is still a dog and a *new dog* is still a dog as well. That means if we leave out the modifiers, the meaning of the phrase and the sentence do not change very much.

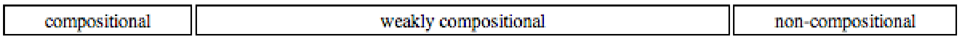
In example (B.3), on the other hand, the modifier *hot* completely changes the meaning of the whole phrase in an unexpected way. A hot dog is a kind of food and not an animal. If we left the modifier out in this case, the meaning of the phrase and the sentence change in an undesired way.

The meaning of *hot dog* is not a simple combination of the meaning of the individual words *hot* and *dog.* The phrase is called **non-compositional** because the meaning of the whole phrase cannot be *composed* from its parts. The opposite is a **compositional** phrase where the meaning of the whole phrase is the combination of its parts, e.g. *small dog.* In compositional phrases, the modifier usually *specifies* the syntactic head but it does not change its meaning.

In the non-compositional case of *hot dog*, we say that the modifier *hot* is **crucial** *in this context*, because leaving it out changes the meaning of the phrase dramatically so the sentence makes no sense as in (B.4).

(B.4) He invented the ~~hot~~ **dog**.

There are also cases in which phrases are **weakly compositional**. These phrases are located in the middle of a spectrum between compositional and non-compositional phrases:



Example:

(B.5) He fell off a **high wire**.

High wire in (B.5) is an example for a weakly compositional phrase. Both *high* and *wire* contribute their original meaning to the meaning of the whole phrase. However, a *high wire* is a special kind of wire that acrobats balance on and not just any string of metal that is high up in the air.

Often it helps to translate the phrase in question into another language. If **a literal, one-to-one translation is not possible** without sounding unnatural in the target language, then the phrase is weakly compositional.

(B.6) Er ist von einem **hohen Draht** gefallen.

(B.7) Er ist von einem **Hochseil** gefallen.

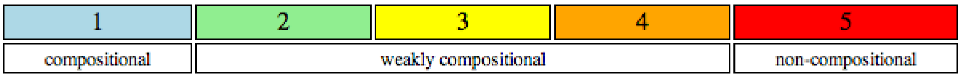
In sentence (B.6) *high wire* was translated literally into German and the sentence sounds quite unnatural. With a bit of imagination, however, a listener might still figure out what the intended meaning is.

In (B.7) the correct German translation *Hochseil* was used and the sentence sounds fine. In the correct translation *Draht* (*wire*) became *Seil (rope)*. This is **an unpredictable choice of translation** and **has to be learned by the translator** at some point in their life.

Compositionality score

We have seen that there are weakly compositional phrases whose degree of compositionality lies in the middle of the compositionality spectrum. For our model we want to quantify the degree.

For this purpose, we introduce a numeric scale that corresponds to the degrees of the spectrum:



1	2	3	4	5
compositional	weakly compositional			non-compositional

- score(*new dog*) = 1 (fully compositional)
- score(*hot dog*) = 5 (fully non-compositional)
- score(*high wire*) = 3 (both meanings flow in but additional knowledge is required to know what exactly a *high wire* is.)

When giving a score, consider the following questions:

- how much additional knowledge is required to understand the meaning?
- how natural would a literal translation sound to a native speaker?

N.B. The score always refers to the degree of compositionality between the syntactic head and the **last modifier** that was tested (see *estimated flux density* below).

Definition of semantic head

Summary: The **semantic head** of a noun phrase is the syntactic head and all modifiers that are crucial in the context. It is the part of the phrase that carries its **main meaning**. For example:

- SemHead(*brown dog*) = *dog*

- SemHead(*new dog*) = *dog*
- SemHead(*hot dog*) = *hot dog*
- SemHead(*high wire*) = *high wire*

From now on, when we say *crucial*, we mean *crucial in the context.*

How to find the semantic head?

Step 1: Find the **syntactic head**

This is the rightmost word of the phrase or the last word before *of.*

little black ***dog***

several thousand ***orders*** *of magnitude*

cosmic microwave background ***radiation***

Step 2: Find the **modifiers**

These are all the others words that are not the syntactic head.

little$_1$ ***black***$_2$ *dog*

several$_1$ ***thousand***$_2$ *orders* ***of magnitude***$_1$

cosmic$_1$ ***microwave***$_2$ ***background***$_3$ *radiation*

Step 3: Does the meaning of the phrase or sentence change considerably when each modifier is left out?

When there is more than one modifier, start with the outermost one (index 1). If it is not crucial, continue with modifier 2 and so on. In the case of modifiers appearing to the left and to the right of the syntactic head, you have to test them both separately (i.e. first look at *orders of magnitude* and then at *several orders*).

When you find a modifier that is crucial, this modifier and the ones with higher indices belong to the semantic head. For example, if *cosmic* was crucial then *microwave* and *background* would be too.

Step 4: Is a modifier crucial?

ask the following questions:

substep: Does the modifier turn the syntactic head into a more specific kind in a transparent way? → strong indicator that it is **not** a crucial modifier, indicates score 1

substep: Does the modifier change the meaning of the syntactic head in an unexpected way? → strong indicator that it **is** a crucial modifier, indicates score 5

substep: Is additional knowledge required to understand a modification? Would a literal translation be possible and how natural would it sound to a native speaker? (remember *high wire*) → strong indicator for a **weakly compositional** phrase, indicates medium score 2, 3 or 4

another hint: Try a Physics/Mathematics dictionary and get an idea about the phrase in question and to make a better decision.

Step 5: Are there modifiers that are crucial for the meaning of the phrase?

substep: no? → The semantic head is the same as the syntactic head.

substep: yes? → The semantic head is the syntactic head + the crucial modifiers.

Examples

Here are some examples phrases with examples usages, comments, correct answers and scores.

<u>most galaxies</u>

- In **most galaxies**, neither atomic hydrogen nor molecular gas will obey the equation.

- One of the promising ways to investigate galaxy formation is to study the ubiquitous globular star clusters that surround **most galaxies**.

comment: *most* does not change the meaning of galaxies, it is not crucial.
answer: SemHead(*most galaxies*) = *galaxies*
score: 1

black hole

- It is shown that instability of stringy matter near the horizon of a **black hole** (the spreading effect) can be characterized by the Lyapunov exponents.
- We shall thus consider a **black hole** of mass m_1 M and a much smaller object of mass m_2.

comment: A *black hole* is an astronomic phenomenon but a *hole* is a opening e.g. in the floor or the street.
answer: SemHead(*black hole*) = *black hole*
score: 5

degree of freedom

- This model of a black hole has one thermodynamic **degree of freedom**
- By taking into account all **degrees of freedom** of electromagnetic fields and explicitly limiting the bandwidth of the pulses, our result overcomes all the shortcomings.

comment: *degree of freedom* is a special term in Physics. The meaning of the whole phrase cannot be seen from the meaning of the components *degree* and *freedom.* However, the original meaning of *freedom* is still in the phrase because the *degree of freedom* describes how many parameters or dimensions can be changed are in a system. So we give a score of 3.
answer: SemHead(*degree of freedom*) = *degree of freedom*
score: 3

estimated flux density

- The error in the **estimated flux density**, both due to calibration and systematic, is 5.

- This confusion causes an uncertainty of about 0.1 Jy in the **estimated flux density**.

comment: Here we have two modifiers: *estimated* and *flux*. The modifier *estimated* does not change the meaning of *flux density* in an unpredictable way so it is not crucial. Then we look at *flux* as a modifier of *density*. The modifier *flux* does not change the meaning very much. But *density* is a very general term and *flux density* is a well-known special kind of *density* so we decide that in this context, *flux* is a crucial modifier but we give a low score.

answer: SemHead(*estimated flux density*) = *flux density*

score: 2 (between *flux* and *density*)

above assumptions

- Under the **above assumptions**, the characterization of the stellar sources requires only two free parameters.

- The **above assumptions** should be valid for the core of a forming cluster or sub-cluster

comment: above is not crucial here because the modification is compositional.

answer: SemHead(*above assumptions*) = assumptions

score: 1

average surface brightness

- The galaxy contribution per pixel is computed as the azimuthal **average surface brightness** at the distance of the knot from the core of the galaxy.

- In calculating this mean , the contribution of each disk is its **average surface brightness** out to one exponential scale length.

comment: the modifier *average* does not affect the meaning of *surface brightness* strongly and is therefore not crucial. This leaves us with *surface brightness* which is a term used in astronomy to describe the brightness of large objects like galaxies. It is still a kind of brightness but it is not obvious that we are dealing with large astronomical objects. This justifies a score of 3.

answer: SemHead(*average surface brightness*) = *surface brightness*

score: 3 (between surface and brightness)

red dwarf

- Our Sun is termed a yellow dwarf and there are many stars cooler than the Sun called **red dwarfs**.
- **Red dwarfs** are too luminous, or they would have been detected directly in the Hubble Deep Field (HDF).

comment: A *red dwarf* is a kind of star but a *dwarf* is a small person.

answer: SemHead(*red dwarf*) = *red dwarf*

score: 5

near future

- In the **near future** we plan further calculations using cylindrical geometry and finer resolution to study the mixing better.
- None of these galaxies is in danger of running out of gas in the very **near future**.

comment: The modification by *near* is completely compositional and the modifier is not crucial.

answer: SemHead(*near future*) = *future*

score: 1

Bibliography

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4).

Aston, G. and Burnard, L. (1998). The BNC handbook. http://www.natcorp.ox.ac.uk/.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the 2003 Workshop on Multiword Expressions*, pages 89–96, Sapporo, Japan. Association for Computational Linguistics.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Biemann, C. and Giesbrecht, E. (2011). Distributional semantics and compositionality 2011: Shared task description and results. In *ACL 2011 Workshop on Distributional Semantics and Compositionality*, pages 21–28.

Brill, E. (2000). Part-of-speech tagging. *Handbook of Natural Language Processing*, pages 403–414.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Bücker, R. (1999). *Statistik f"ur Wirtschaftswissenschaftler*. Oldenbourg, 4 edition.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Choueka, Y. (1988). Looking for needles in a haystack. In *Proceedings of RIAO'88*, pages 609–623.

Church, K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on Applied natural language processing*, ANLC '88, pages 136–143, Stroudsburg, PA, USA. Association for Computational Linguistics.

Church, K. and Gale, W. (1991). Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1989). Parsing, word associations and typical predicate-argument relations. In *Proceedings of the workshop on Speech and Natural Language*, pages 75–81. Association for Computational Linguistics.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Clark, H. H. (1971). Word associations and linguistic theory. In Lyons, J., editor, *New Horizon in Linguistics*, pages 271–286. Penguin, London.

Clark, S. and Pulman, S. (2007). Combining symbolic and distributional models of meaning. In *Proceedings of the AAAI Spring Symposium on Quantum Interaction*, pages 52–55.

Clarke, D. (2012). A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.

Coecke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the 2007 on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Dagan, I., Lee, L., and Pereira, F. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1):43–69.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In Klavans, J. L. and Resnik, P., editors, *The Balancing Act*, pages 49–66. MIT Press, Cambridge, MA.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):pp. 297–302.

Dridan, R. and Oepen, S. (2012). Tokenization: Returning to a long solved problem — a survey, contrastive experiment, recommendations, and toolkit —. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–382, Jeju Island, Korea. Association for Computational Linguistics.

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

Edmonds, P. (1997). Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 507–509. Association for Computational Linguistics.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii. Association for Computational Linguistics.

Erk, K. and Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97, Uppsala, Sweden. Association for Computational Linguistics.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

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

Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communication*. MIT Press, New York.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Feng, Y. and Lapata, M. (2010). Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99, Los Angeles, California. Association for Computational Linguistics.

Fillmore, C., Baker, C., and Sato, H. (2002). Seeing arguments through transparent structures. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 787–91.

Finkel, J. R., Grenager, T., and Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer Series in Statistics.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence, IJCAI*, pages 1606–1611.

Garrette, D., Erk, K., and Mooney, R. (2011). Integrating logical representations with probabilistic information using markov logic. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 105–114. Association for Computational Linguistics.

Giuliano, V. E. (1965). The interpretation of word associations. In Stevens, E., Giuliano, V., and Heilprin, L., editors, *Proceedings of the symposium on statistical association methods for mechanized documentation*, Washington, D.C.

Green, S., de Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 725–735, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Grefenstette, E. and Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Gries, S. T., Hampe, B., and Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, 16(4):635–676.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *2nd International Joint Conference on Natural Language Processing, IJCNLP*, pages 767–778.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Harris, Z. (1970). Distributional structure. In *Papers in structural and transformational Linguistics*, pages 775–794. Dordrecht.

Hartung, M. and Frank, A. (2011). Exploring supervised lda models for assigning attributes to adjective-noun phrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages

540–551, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19:103–120.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. Wiley, Hoboken, 2nd edition.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin and Review*, 10(4):785–813.

Jackendoff, R. S. (1997). *The Architecture of the Language Faculty*. MIT Press.

Jeh, G. and Widom, J. (2002). Simrank: A measure of structural-context similarity. In *KDD '02*, pages 538–543.

Jelinek, F. (2009). The dawn of statistical asr and mt. *Computational Linguistics*, 35(4):483–494.

Jenkins, J. J. (1970). The 1952 Minnesota word association norms. In Postman, L. and Keppel, G., editors, *Norms of word association*, pages 1–38. Academic Press, New York.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing*. Prentice Hall, 2nd edition.

Justeson, J. S. and Katz, S. (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(01):9–27.

Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the 2006 Workshop on Multiword Expressions*, pages 12–19, Sydney, Australia. Association for Computational Linguistics.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Kent, G. H. and Rosanoff, A. J. (1910). A study of association in insanity. *American Journal of Insanity*, 67(1):317–390.

Kiss, G. R., Armstrong, C., Milroy, R., and Piper, J. (1973). An associative thesaurus of English and its computer analysis. In Aitken, A. J., Bailey, R. W., and Hamilton-Smith, N., editors, *The Computer and Literary studies*. University Press, Edinburgh.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Kjellmer, G. (1991). A mint of phrases. In Aijmer, K. and Altenberg, B., editors, *English Corpus Linguistics*. Longman, London.

Korkontzelos, I. and Manandhar, S. (2009). Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 65–68. Association for Computational Linguistics.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. (2010). Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359.

Krenn, B. and Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *ACL Workshop on Collocations*, pages 39–46.

Krenn, B. and Evert, S. (2005). Separating the wheat from the chaff: Corpus-driven evaluation of statistical association measures for collocation extraction. In Fisseni, B., Schmitz, H.-C., Schröder, B., and Wagner, P., editors, *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*, volume 8 of *Computer Studies in Language and Speech*, pages 104–117. Lang, Peter, Frankfurt am Main.

Kulkarni, N. and Finlayson, M. (2011). jmwe: A java toolkit for detecting multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 122–124, Portland, Oregon, USA. Association for Computational Linguistics.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Landauer, T. and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lapata, M., Keller, F., and McDonald, S. (2001). Evaluating smoothing algorithms against plausibility judgements. In *39th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 346–353.

Laws, F., Michelbacher, L., Dorow, B., Scheible, C., Heid, U., and Schütze, H. (2010). A linguistically grounded graph model for bilingual lexicon extraction. In *Proceedings of the 23nd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China. Coling 2010 Organizing Committee.

Lee, L. (1999). Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 25–32.

Lee, L. (2001). On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*, pages 65–72.

Lee, L. (2004). I'm sorry dave, i'm afraid i can't do that": Linguistics, statistics, and natural language processing circa 2001. *Computer Science: Reflections on the Field, Reflections from the Field.*

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 768–774, Montreal, Quebec, Canada. Association for Computational Linguistics.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, Maryland, USA. Association for Computational Linguistics.

Lykke, M., Larsen, B., Lund, H., and Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, pages 627–630.

Manning, C. and Klein, D. (2003). Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials-Volume 5*, pages 8–8. Association for Computational Linguistics.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, MA.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

McGee, I. (2009). Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores. *Corpus Linguistics and Linguistic Theory*, 5(1):79–103.

McRae, K., Cree, G., Seidenberg, M., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Melamed, I. D. (1997). Automatic discovery of non-compositional compounds in parallel data. In *EMNLP*.

Metzler, D. and Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40(5):735–750.

Michelbacher, L., Evert, S., and Schütze, H. (2007). Asymmetric association measures. In *International Conference on Recent Advances in Natural Language Processing, RANLP*.

Michelbacher, L., Evert, S., and Schütze, H. (2011a). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, 7(2):245—-276.

Michelbacher, L., Han, Q., and Schütze, H. (2013). Unsupervised feature adaptation for cross-domain nlp with an application to compositionality grading. In *CICLing (1)*, pages 1–12.

Michelbacher, L., Kothari, A., Forst, M., Lioma, C., and Schütze, H. (2011b). A cascaded classification approach to semantic head recognition. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 793–803, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Michelbacher, L., Laws, F., Dorow, B., Heid, U., and Schütze, H. (2010). Building a cross-lingual relatedness thesaurus using a graph similarity measure. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8).

Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory*, 5(2):175–200.

Montague, R. (1973). The proper treatment of quantification in ordinary english. *Approaches to natural language*, 49:221–242.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://web.usf.edu/FreeAssociation/.

Nicholson, J. and Baldwin, T. (2008). Interpreting compound nominalisations. In *LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 43–45.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Palermo, D. S. and Jenkins, J. J. (1964). *Word association norms: Grade school through college.* Universtity of Minnesota Press, Minneapolis.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the 2008 Workshop on Multiword Expressions*, page 54.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):138–158.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Porter, M. F. (1980). An algorithm for suffix stripping. In van Rijsbergen, C. K., Robertson, S., and Porter, M. F., editors, *New models in probabilistic information retrieval.* MCB UP Ltd.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010). mwetoolkit: a framework for multiword expression identification. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *COLING 1999.*

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

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA.

Reddy, S., McCarthy, D., Manandhar, S., and Gella, S. (2011). Exemplar-based word-space model for compositionality detection: Shared task system description. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 54–60, Portland, Oregon, USA. Association for Computational Linguistics.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual*

Conference of the North American Chapter of the Association for Computational Linguistics, pages 109–117, Los Angeles, California. Association for Computational Linguistics.

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61(1):127–159.

Riehemann, S. (2001). *A Constructional Approach to Idioms and Word Formation.* PhD thesis, Stanford University.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Mexico City.

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD thesis, Department of Linguistics, Stockholm University.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Commununications of the ACM*, 18:613–620.

Saussure, F. (1966). *Course in general linguistics.* McGraw-Hill, New York.

Schmid, H. (2004). Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *COLING '04*, page 162.

Schone, P. and Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108, Pittsburgh, Pennsylvania, USA. Association for Computational Linguistics.

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing, Supercomputing'92*, pages 787–796. IEEE.

Schütze, H. (1993). Word space. In *Advances in Neural Information Processing Systems 5*.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 134–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Shannon, C. E. and Weaver, W. (1949). The mathematical theory of information.

Silva, J. and Lopes, G. P. (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Sixth Meeting on Mathematics of Language*, pages 369–381.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford Universtity Press, Oxford.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177.

Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22:1–38.

Socher, R., Huang, E., Pennington, J., Ng, A., and Manning, C. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in Neural Information Processing Systems*, 24:801–809.

Spence, D. P. and Owens, K. C. (1990). Lexical Co-Occurrence and Association Strength. *Journal of Psycholinguistic Research*, 19(5):317–330.

Sporleder, C. and Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Strube, M. and Ponzetto, S. (2006). Wikirelate! Computing semantic relatedness using wikipedia. In *Twenty-First National Conference on Artificial Intelligence, AAAI*, pages 1419–1424.

Tanaka, T. and Matsuo, Y. (1999). Extraction of translation equivalents from non-parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 109–19.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press.

Tjong Kim Sang, E. and Buchholz, S. (2000). Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132. Association for Computational Linguistics.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Turney, P. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. In De Raedt, L. and Flach, P., editors, *Machine Learning: ECML 2001*, volume 2167, pages 491–502. Springer Berlin / Heidelberg.

Turney, P. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Tversky, A. N. (1977). Features of similarity. *Psychological review*, 84(4):327–352.

Voorhees, E. M. (1999). The TREC-8 question answering track report. In *8th Text Retrieval Conference, TREC*, pages 77–82.

Washtell, J. and Markert, K. (2009). A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations. In *Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 628–637.

Weeds, J. (2002). Asymmetry in similarity between words. In *Annual CLUK Colloquium*, pages 1–3, Leeds, UK.

Wermter, J. and Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge). In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, COLING/ACL*.

Wettler, M. and Rapp, R. (1993). Computation of word associations based on the co-occurrences of words in large corpora. In *1st Workshop on Very Large Corpora*, pages 84–93.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, 4(2):253–290.

Zhai, C. and Lafferty, J. D. (2002). Two-stage language models for information retrieval. In *Proceedings of the 25th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pages 49–56. ACM.