University of Zurich[UZH]

Master's thesis
for the degree of
Master of Arts
presented to the Faculty of Arts and Social Sciences
of the University of Zurich

The "Raison d'Être" of Word Embeddings in Identifying Multiword Expressions in Bilingual Corpora

Author: Phillip Ströbel
Student ID: 10-740-967

Examiner: Prof. Dr. Martin Volk
Institute of Computational Linguistics
Submission date: 05.01.2017
Corrected version of 10.01.2017

Abstract

Multiword expressions do not only pose problems in linguistics, but also in natural language processing. Linguists, for instance, besides being concerned with finding a proper definition of this phenomenon, tackle multiword expressions from different angels, e.g., semantics, phraseology, or syntax. The computational linguist, at the same time, tries to get to grips with multiword expressions in parsing or statistical machine translation. It is especially the translation of multiword expressions to which we dedicate this thesis. Since statistical machine translation relies on word alignments and phrase tables, and are not always able to produce satisfactory results, we are on the lookout for other methods. A promising, rather novel technique are bilingual word embeddings. After having determined the best bilingual word embedding model from a total of 48 candidates trained on different parameters for a German-French parallel corpus, we compare its performance against phrase tables and a multilingual concordance system which relies on word alignments. The findings are sobering, since bilingual word embeddings cannot yet compete with traditional methods. Depending on the translation direction and the method it is compared to, the bilingual word embedding model provides correct translations in only 15.04% to 34%. Nevertheless, we find that bilingual word embeddings are indeed able to capture similarities and relationships that go undetected by conventional methods, which is why we argue for further development of hybrid systems.

Remark: This version contains spelling and formatting corrections as proposed by Prof. Dr. Martin Volk and is thus not identical to the thesis handed over to the ZB.

Acknowledgement

First of all, I would like to express my gratitude to Prof. Dr. Martin Volk, who can be held responsible for me taking on the M.A. studies in Multilingual Text Analysis. He also provided me with inspiration for the topic of this very thesis and gave me enough room and freedom to put my ideas into practice. His favourite words during a lecture – "This would be a great topic for a master thesis!" – could also be heard in connection to multiword expressions and parallel corpora.

Next, my thanks go Johannes Graën. His expertise in database design and structure helped me numerous times when I did not know how to go further. I highly appreciated his many and diverse suggestions and ideas he gave me in our discussions. Thanks to him, I finally made peace with *PostgreSQL* and *make*.

Last but not least, I thank all the rest of the staff at the Institute of Computational Linguistics. Whomever I had addressed during the production of this thesis never hesitated to be of the best help he or she could provide.

Contents

List of Figures

List of Tables

List of Acronyms

HTML	HyperText Markup Language
MWE	Multiword expression
NLP	Natural Language Processing
PDF	Portable Document Format
STTS	Stuttgart-Tübingen tagset
XML	eXtensible Markup Language

1 Introduction

1.1 Motivation

Multiword expressions (MWEs) are ubiquitous in natural language. Jackendoff [1997] claims that our mental lexicon of MWEs is as extensive as the one that is made up of single words. If this assumption is true, it not only changes the way we look at speech production and understanding in linguistics, but also how we will have to tackle this phenomenon in natural language processing (NLP). Multiword expressions as such have always been of interest to the computational linguist, and research on this topic is still gaining in popularity. The focus lies mostly in the recognition and processing of MWEs, as well as on translation and semantics. In computational linguistic research concerning MWEs, researchers are almost exclusively concerned with the processing of such units, while the origins and peculiarities of this phenomenon are often neglected. In this respect, much work would profit from a thorough understanding of MWEs. The difficulty of MWEs is that they presents themselves to us in many forms, that is **idioms**, **fixed expressions**, **lexical verb constructions**, **light verb constructions**, and many more. And often, researchers have given these occurrences different names, among which we find **collocations**, **phrasemes**, **phraseologisms**, or, as we call them throughout this thesis, **multiword expressions**.

Quite ironically, the phenomenon under scrutiny in this thesis is itself a MWE. For a starter, let us take a closer look at this term. We notice that it consists of two words (two nouns, to be exact, one in singular, one in plural), as we illustrate in (1.1).

(1.1) multiword + expressions

Naturally, these words can also occur independently of each other, but if they occur together, their meaning changes. Although we do not observe a change of meaning in the single words which constitute the MWE, we could argue that together, they form a new semantic unit. In this case, we would say that the term from example (1.1) means 'an expression that consists of several words'. This is an example in which we are able to derive the meaning from the constituents of the multiword unit, even if it were not present in our mental lexicon. As we will see when we move to the different types of MWEs, this is not always the case. Although even beginners of English are familiar with the concepts of 'dog' and 'day', they would most probably be unable to guess the meaning of a 'dog day', which means 'a period of inactivity', to take an example from Baldwin and Kim [2010]. The derivation of this meaning is already difficult, if not impossible, without the proper background knowledge (or without having simply learnt the expression by heart). Note that we are still dealing with noun-noun compounds, to which also the term 'multiword

expression' belongs. Various other combinations are possible, like verb-noun constructs, as in 'to pay attention'.

The 'dog day'-example illustrates one of many challenges MWEs pose, namely their lexical status. This peculiarity of MWEs seems especially worrying, mainly for one reason: if even human beings have difficulties in identifying and understanding MWEs, how are machines going to perform in this task? How can they be identified? And, even more importantly when we enter the realm of natural language understanding, how can we derive meaning from MWEs? These questions are pressing issues in NLP and, therefore, MWEs have been a research focus for quite some time. Still, much remains to be done, as the incorrect treatment of this phenomenon causes problems in lexical semantics, part-of-speech tagging, parsing, statistical machine translation, and other subfields of NLP.

In order to understand how MWEs are lexicalised and how their meaning comes about, we need to get deeper insights into their patterns. We also need to understand how they are used in natural language. This is why first of all, we have to broaden our linguistic knowledge. Therefore, we shall outline the concept of a MWE in more detail, as it lies at the intersections of diverse linguistic fields such as morphology, syntax, semantics, phraseology, corpus linguistics, and lexicography and lexicology. In some of these fields, the definition of what constitutes a MWE and the underlying terminology differs. Throughout this thesis, we will stick to the term MWE, though, as it best describes the focus of the study. The motivation for choosing this term follows in Chapter 2. Disregarding terminological issues, it still remains true that only with a thorough understanding, or at least a convincing and probable theory of how MWEs work in natural language, we will be able to tackle this problem from the point of view of a computational linguist more successfully.

Traditional approaches towards the identification and extraction of MWEs are of statistic nature and do not consider any linguistic background knowledge. Such association measures, which allow us to determine the degree of collocability of a MWE, purely work based on assumptions made on the distribution of words. Effective identification of MWEs thus requires additional information, like part-of-speech tags or parsing dependencies. We can use this information and extract MWE candidates based on patterns. Profound linguistic insights undoubtedly are a necessary precondition for the definition of such patterns. Obviously, the performance of association measures in the identification of MWEs depends not only on the preprocessing steps, but also on the corpus data. Hence, if an identified MWE candidate really constitutes a MWE is often subject to human evaluation.

Although we still struggle with the identification of MWEs in monolingual corpora, there is a considerable amount of research in finding translation candidates for MWEs. The question of how MWEs are translated adds another twist to this phenomenon. Clearly, specific linguistic properties influence the translation. Translating from a compounding language to a non-compounding language or vice-versa poses additional challenges. Can we translate German *Weinglas* 'wine glass' to French *vin verre*? The answer is no, since the correct translation is *verre à vin*. This translation must not be confused with *verre*

de vin, which means 'glass of wine'. Translation issues have mostly been approached by taking into account word alignments and phrase tables. However, they cannot provide information about the semantics of a certain expression.

This is where we have to set out to look for alternatives. Over the last few deacdes, computational linguists have tried various ways to come to terms with the semantics and translations of MWEs. The development of new techniques is of course beneficial to the automatic processing of natural language. With every new algorithm, though, it is necessary to ask oneself what it could be used for. The same is true for word embeddings (Mikolov et al. [2013]), that is to say representations of words in a low-dimensional vector space. Word embeddings have, all of a sudden, become some sort of panacea; everybody uses them for everything. However, the usefulness of word embeddings is not evident in all use cases, and determining whether they really benefit some application is always subject to discussion. First and foremost, word embeddings allow us to calculate the similarity of words, or to find the most related words given a query. Since we are dealing with compositional semantics when tackling MWEs, we assume that word embeddings could be of help in identifying the meaning of certain MWEs, or simply to find synonyms. And since it is possible to project several languages into the same vector space by applying word embeddings, we have to ask ourselves whether it is possible to identify translations of MWEs by using bilingual word embeddings. In this way, we would be able to not only capture meaning relationships, but also translations by using only one model, therefore killing two birds with one stone.

1.2 Research Questions

In light of the above, we pose the following research questions:

1. What is the conception of MWEs in linguistics and how can the recognition and processing of MWEs profit from linguistic knowledge?

2. How can we exploit German-French bilingual word embeddings in order to find and filter translation candidates for MWEs?

3. Can German-French bilingual word embeddings achieve similar performance in the translation of MWEs as techniques using word alignment of phrase table information?

1.3 Thesis Structure

In Chapter 2 we will explore the nature of MWEs. We aim at laying the foundations for the thesis, which heavily depends on a clear-cut definition of MWEs (if this is possible at all).

The different types of MWEs that linguists identified will be the topic of Chapter 3. Each type poses different challenges in NLP. We will describe the various types of MWEs in detail, with a special focus on German and French.

Chapter 5 focuses on conventional methods that have been employed to identify and filter MWEs. We will concentrate on the most used and most prominent statistical measures, while rule-based techniques will only play a minor role.

The challenges MWEs pose are manifold. We will address this issue in Chapter 4, highlighting related fields which would profit from an improved treatment of MWEs.

Chapter 6 introduces word embeddings. After a short excursion to the history of word embeddings (which coincides with the advent of neural networks), we will aim at providing a concise description of the different algorithms. As we work with bilingual corpora, we will also present bilingual word embeddings.

After having laid the fundamentals in Chapters 2 to 6, we arrive at the core of this thesis (Chapter 7). Following a short review on identifying MWEs in bilingual corpora, we will present the data we worked with, as well as our approach towards this problem.

In Chapter 8, we will present the evaluation of the employed methods. On the one hand, we will evaluate the system on its own, meaning how well it performs in retrieving MWEs. On the other hand, we will also compare it to a very similar, already existing system.

The end of this thesis (Chapter 9) provides important conclusions and insights after having worked on this topic. Furthermore, we will give suggestions for future work, hoping to inspire researchers to follow up on this approach in order to come to terms with MWEs.

2 Multiword Expressions

The way researchers look at MWEs differs, and there is not always consesus about what constitutes a MWE. Linguists, for example, are concerned with the questions how MWEs are represented in the mental lexicon and how they can be analysed in syntax, while the computational linguist tries to get to grips with MWEs in, say, statistical machine translation. However, and especially after Sag et al. [2002] have published a paper with the catching title "Multiword Expressions: a Pain in the Neck for NLP", MWEs suddenly gained popularity in research. Curiously, this also shows in the Google Ngram Viewer, as we show in Figure 1.

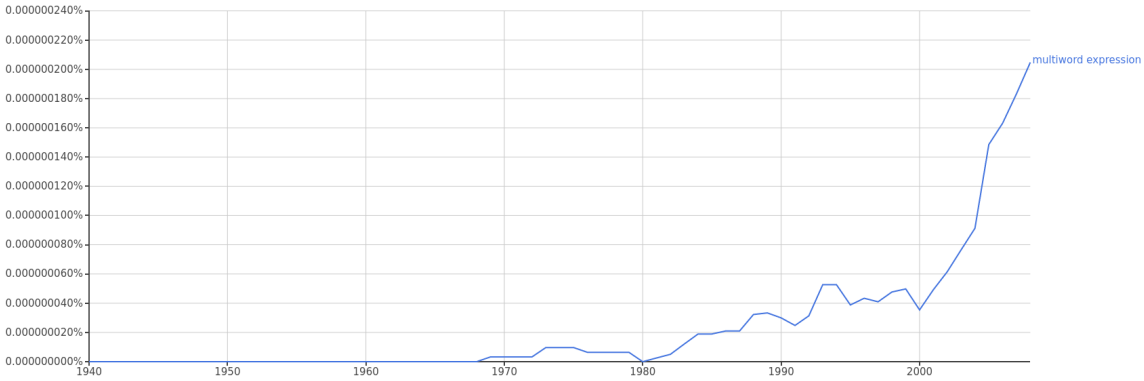


Figure 1: N-gram frequency of the term **multiword expression** taken from the Google Ngram Viewer. We see a steady increase from the 2000s onwards.

As of today, we have a fair amount of literature concerning MWEs at our disposal. And still, the number of publications tackling this topic is increasing. This is a strong indicator that the problems MWEs cause are not yet all solved. Thus, we claim that there is much work do be done in order to come to terms with MWEs. For this reason, and as the phenomenon of MWEs is central to this thesis, we would like to provide an extensive introduction to this topic, starting with a thorough definition of MWEs (Section 2.1). Following that, we will introduce the reader to the different types of MWEs in Chapter 3. We will see that not all MWEs consist of nouns and that there are constructions that have long-distance relationships. As the focus of this thesis lies on German-French bilingual corpora, we will also provide German and French examples to illustrate the various kinds of MWEs. We will also find out where the construction and usages of MWEs in these two languages are similar, yet differences are far more interesting.

2.1 Towards a Definition of the Term *Multiword Expression*

"In the beginning was the Word" - this is how the Gospel of John in the Bible starts (1:1). If we add another word, we might end up with a group of words which we call **phrase** under certain circumstances and if we continue adding words and combining phrases, we might arrive at a hopefully grammatical sentence. But how likely is it that any of these combinations in a sentence is a MWE? According to Jackendoff [1997, p. 97], our lexicon of MWEs is at least as extensive as the one containing only simple words. Yet not every combination of two or more words constitutes a MWE. For this reason, we have to establish a definition of the term. If we take into account that there are other expressions which refer to groups of words, among which we find **multiword unit**, **multiword phrase**, **collocation**, **co-occurrence**, **phraseme**, **phraseological unit**, it is all the more important to precisely define what constitutes a MWE. First and foremost, all these terms, including MWE, refer to groups of words but not all of them denote the same concept. As the differences are sometimes minor, it is not astonishing that there are misconceptions when it comes to the analysis of MWEs in NLP.

As we will see in the following sections, MWEs are complex structures. In order to be able to correctly classify this phenomenon, we have to take into account many different parameters. These parameters have their origins in various areas. We have identified the following fields of studies which are directly (or indirectly) concerned with MWEs:

- **Morphology** As MWEs are made up of words, we have to position this study to some extent in the realm of morphology, which is the study of words and how words are formed, composed, and stored in the mental lexicon.

- **Phraseology** MWEs consist of several words. In some cases, the linking of several words results in phrases, as in *term paper*, which is a noun phrase or *to put a question*, denoting a verb phrase, to borrow concepts from syntax. But there are also other types of phrases, which we will see in Section 4. Because of the phrase-like nature inherent to certain kinds of MWEs, distinct concepts of phraseology come into play.

- **Syntax** Directly following from the phrasal nature is the fact that MWEs are also syntactic structures to some degree, which integrate into sentences. Syntax postulates, however, that only atomic units can fill syntactic slots. In the syntax, we will try to find justifications for seeing some kinds of MWEs as atomic units.

- **Semantics** Another important factor in the analysis of MWEs is semantics, which tell us how words, word groups, and sentences construe meaning. Because the meaning of MWE expression is not always compositional, this task is especially challenging. As concerns MWEs, the semantics of sentences comes into play when they take almost sentence-like form, as in idiomatic expressions like 'A bird in the hand is worth two in the bush'. Since we focus on smaller-scale units, we restrict semantics to lexical and phrasal semantics.

- **Lexicography/lexicology** Capturing MWEs and their meanings is the task of

lexicography and lexicology, which also influence the study of MWEs. We dedicate a few more words to this field in Section 4.1 of Chapter 4.

- **Corpus linguistics** Last but not least, corpus linguistics, which we consider to be a subfield of NLP rather than a pure methodological framework (Aijmer and Altenberg [2014]), also addresses the issue of MWEs, while particular attention is paid to collocations (which we will define in a Section 2.1.6).

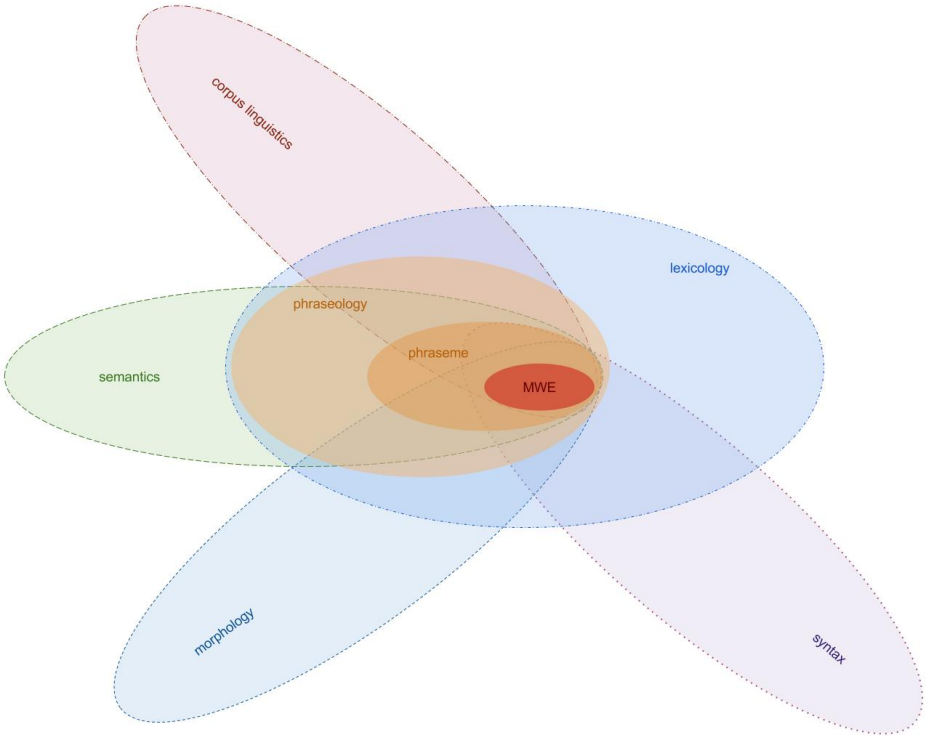


Figure 2: Multiword expressions at the intersection of various research fields.

It becomes evident that we cannot study MWEs in isolation. Rather, we are dealing with an interdisciplinary phenomenon, which we visualise in Figure 2. There we show that all fields interact with each other, which on the one hand allows for this phenomenon to be tackled from many different angles, but on the other hand may turn out as a pitfall, because it is likely that we overlook important concepts in the analysis of MWEs. This statement especially holds when we aim at developing a characterisation of MWEs. In the next sections[1], we will take a closer look at each of the aforementioned fields, with the aim of highlighting how each research area contributes to the phenomenon of MWEs. Finally, we will arrive at a characterisation of the term **multiword expression** that takes as much from linguistics into consideration as possible. The sections on morphology, semantics, syntax, and the mental lexicon are heavily influenced by "The Architecture of the Language Faculty" by Jackendoff [1997]. We mainly employ his terminology, and only

[1]The only exception being lexicography and lexicology, which are topic of discussion in Section 4.1 of Chapter 4.

deviate where necessary. His work, in turn, is mainly based on the **Minimalist Program** as described in Chomsky [1995][2].

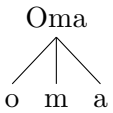
2.1.1 The Role of Morphology

What is a word? What constitutes a word? And much more important: can we (or do we have to) perceive MWEs as one word? At first, these questions seem to be trivial. But since semantics and syntax heavily depend on how we define the notion of a **word**, it is worth to say a few words about morphology. In this section, we will outline the role morphology plays in the analysis of MWEs. Sag et al. [2002, p. 2] give a very broad definition of MWEs in that they describe them as "idiosyncratic interpretations that cross word boundaries (or spaces)". This is vague, especially because neither "idiosyncratic interpretations" nor the notion of a "word" is described in more detail. What we can already tell from the definition, though, is that a MWE consists of two or more **words**, while the fact that they cross word boundaries does not necessarily mean that we must only consider words which occur directly after one another.

We start with an example[3]. Following Sag et al. [2002]'s definition, the German sentence in (2.1) consists of the words (assuming that spaces indeed are word boundaries) *Seine*, *Oma*, *gab*, *gestern*, *völlig*, *unerwartet*, *den*, *Löffel*, and *ab*, nine in total.

(2.1) *Seine* *Oma* *gab* *gestern* *völlig* *unerwartet* *den* *Löffel* *ab.*
His grandmother give.PAST yesterday totally unexpected the spoon away
'His grandmother totally unexpectedly bit the dust yesterday.'

We can attribute a different degree of complexity to each of these words. The word *Oma*, for example, is a simple noun, consisting of nothing more than a single **morpheme**, which is the smallest unit which carries meaning in morphology. We can further split this word into different phonemes. According to (2.2), *Oma* consists of three phonemes, which together make up what Jackendoff [1997] calls the **phonetic form**.

(2.2)


Besides phonetic form, there is also the notion of **phonological structure**. [Jackendoff, 1997, p. 28] states that the "generative system for phonological structure [...] contains such primitives as phonological distinctive features, the notions of syllable, word, and phonological and intonational phrase, the notions of stress, tone and intonation contour". The word *Oma* of course also possesses phonological structure. A more detailed phonological analysis would yield the representation [ˈoːma], which also expresses primary stress

[2]The Minimalist Program is Chomsky's attempt to reduce linguistic analysis to only the most necessary units and functions.

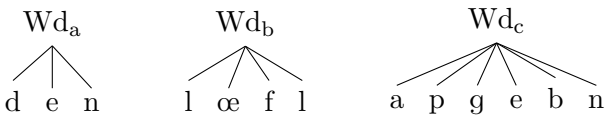
[3]Note that for the sake of simplicity, we will stick to the same example throughout the remainder of this section. We will return and from time to time also elaborate on specific arguments either in footnotes or later when we look at different types of MWEs in Section 3).

and vowel length. Linguists nowadays assume that phonetic form, namely the phonetic representation of a word, is a level of phonological structure. We will again return the concept of phonological structure when we look at how words are represented in the mental lexicon.

In contrast to the simple word *Oma*, the word *unerwartet* is a **complex word**, consisting of four morphemes (*un-*, *er-*, *warten*, *-t*). Booij [2010, p. 1] defines a complex word as "a concatenation of morphemes". As such, the word *unerwartet* is formed from more basic units through the affixation of morphemes which carry meaning.

There is another structure in (2.1), though, which is of interest to us. If we want to understand the meaning of the sentence – note that we already graze semantics – our analysis yields that it basically means that somebody's grandmother died. However, *sterben* 'to die', does not occur in the sentence. So, how do we arrive at this interpretation? On closer inspection of the example sentence, we find the MWE *den Löffel abgeben* in it, which roughly corresponds to 'bite the dust' or 'kick the bucket' in English (which are MWEs too). Clearly, this expression consists of several morphological units and words. Each of these units has itself morphological and phonological structure as we see in (2.3).

(2.3)



Immediately, the question of the morphological complexity of MWEs arises. Booij [2010] links phrase-like units to **Construction Morphology**. Construction Morphology forms the basis both for complex words and word formation. In Construction Morphology, linguists try to find ways to be able to better describe the relationship between morphology, syntax, and the mental lexicon. In order to achieve a better understanding, their focus lies on the semantic properties of complex words. In his work, Booij [2010, p. 8] transfers techniques from the analysis of complex words in Construction Morphology to the analysis of what he calls **word-like phrasal expressions**. He claims that Construction Morphology helps in analysing MWEs, very much in the way that we will attempt in the following sections. Moreover, he explicitly stresses the similarity of such phrasal expressions and complex words, a fact that is somehow denied by Adams [2014].

To illustrate this point, consider French nouns like *sac à dos* 'backpack' or German separable prefix verbs like *auffallen* 'to notice'. The latter is of phrasal nature when used in a sentence, as for example in *Ihm **fällt** ihr hübsches Kleid **auf*** 'He notices her nice dress'. [Booij, 2010, p. 10] argues that we form both the French and the German instance according to phrasal constructional schemas. It is these schemas which are responsible for the inclusion of such words in the mental lexicon. Additionally, these schemas are productive, meaning that if we generalise the French *sac à dos*, we receive something like *N à N*, with which we can build other French nouns like *bagage à main* 'hand lugguage'. Although consisting of more than one word, we would attribute word-like character to this "word", not least because the expression is fully lexicalised and there is no other single word which denotes this concept.

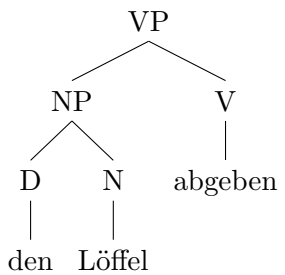
Hence, the crucial point in the analysis is whether to analyse MWEs as a single word

or as a phrase. Very much like light can be described as a wave or a particle at the same time, commonly known as the **wave-particle duality** as postulated by Albert Einstein, a MWE can be seen as (complex) word or phrase. In order to be able to see behind the rationale of this duality, we have to get deeper insights in how MWEs could be represented in the mental lexicon. But first, we will turn our attention to syntax, which allows us to better understand how MWEs can be perceived as atomic units.

2.1.2 The Importance of Syntax

The Minimalist Program assumes that the basis for syntax are units called X^0, which constitute atomic units or, put differently, words. Yet, when we, or rather our **computational system**, assigns a slot to a MWE we want to use in a sentence, can we really perceive such a MWE as an atomic unit? Let us start with a basic syntactic analysis of the expression *den Löffel abgeben*. Such an analysis renders a tree which resembles a syntax tree, as example (2.4) illustrates:

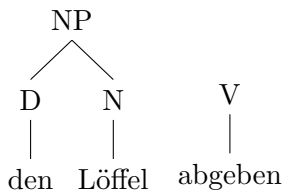
(2.4)



Although we are mainly dealing with morphemes in an analysis of the word like *unerwartet* while words are the focus in (2.4), there are certain analogies to the analysis of complex words. Both, (2.4) and the word *unerwartet* have phonological structure (*unerwartet* [ˈʊnɛɐ̯ˌvaʁtət], *den Löffel abgeben* [deːn ˈlœfˌl ˈapˌgeːbn̩]). Moreover, order is important in either construction. We cannot just move around atomic units as we like in *unerwartet*. This would yield "words" like **erunwartet* or **erwartetun*. The same is true for (2.4), although when this expression is realised as in (2.1), we notice that the bond between the NP and the V is not strong, meaning there is some syntactic variation. We notice that the expression "spans" over several words, just as described in Sag et al. [2002].

Because we observe some degree of freedom in the usage of this MWE, this implies that an idiomatic expression as (2.4) would be better represented as shown in (2.5):

(2.5)



In this case, it is appropriate to split the expression into several other syntactic surface structures so as to cater for the syntactic variability. By assuming such a representation, the realisation of this phrase in, for example, an infinitive construction is possible. If we consider passivisation, however, we find that if we use this expression in passive voice, the semantics of the idiomatic expression is violated. Example (2.6a) shows how the expression

can be used in its infinitive form in the subordinate clause. Note that some morphological variation happened in the verb. The infix *zu* has been added. On the other hand, (2.6b) exemplifies the passivisation of the expression. Although being grammatically sane, the inherent semantics of the expression is lost and the reading has to be taken literally, namely that there really was a spoon which was handed from one person to another.

(2.6) (a) *Er hatte nicht vor, den Löffel abzugeben.*
he have.PAST not forward the spoon deliver
'He did not intend to kick the bucket.'

(b) *?Der Löffel wurde abgegeben.*
the spoon is.PAST deliver
'The spoon was delivered.'

The analyses of the examples imply that we must allocate some level of fixedness to the expression. This is why some MWEs are also referred to as **fixed** or **semi-fixed expressions** (more of them later). We can attribute several constrains to the level of fixedness of *den Löffel abgeben*. For example, the expression requires the NP to be in the accusative case. Moreover, pluralisation is not permissible (e.g., *die Löffel abgeben*, 'give away the spoons', even if the subject of the sentence is in plural), as is modification of the noun by adjectives as in *den silbernen Löffel abgeben*, 'give away the silver spoon'. Syntactically, such operations are of course permitted, but when we apply them to a MWE, it is well possible that we change its meaning.

The role of syntax is thus to license such constructions. But how can it know that an expression like *den Löffel abgeben* is a valid construction? The answer to that question lies in how words are stored in the mental lexicon, which is the topic of the following section.

2.1.3 The Mental Lexicon

Chomsky [1995] argues that there are two systems in the **faculty of language**[4] which complement each other: the **mental lexicon** and the **computational system**. The mental lexicon provides the words which the computational system uses to produce and analyse language. The production module becomes active when we utter sentences, while the analysis takes place when we want to understand what someone has said or written. We are mainly concerned with how words are stored in the mental lexicon here.

What constitutes a lexical item[5]? What is worth being lexicalised or, in other words, stored in the mental lexicon? These are first and foremost questions which have preoccupied researchers in psycholinguistics, and since applications in NLP mimic language production and understanding, computational linguistics can only benefit from work in this field. The fact that there are even branches called lexicology and lexicography in computational linguistics strengthens the assumption that a better understanding of the mental lexicon helps. As such, the study of words and their meaning, combined with

[4]The faculty of language is mainly concerned with how humans produce and use language.

[5]Lexical items are sometimes also referred to as **lexical entries** or **listemes**.

how they are used in natural language should inspire and motivate many computational linguistic theories and applications.

It would make sense to assume that a "word" which consists of several morphemes and that can be deduced by lexical rules or functions[6] does not have to be stored in the mental lexicon. Thus, the past tense of the regular verb *walk*, which is formed by adding the inflectional suffix *-ed* and hence results in *walked*, does not get a separate entry in the mental lexicon. Much in contrast, the past form of an irregular verb like *be*, namely *was*, would probably be listed in a separate lexical entry, with a cross-reference to the infinitive form in order to prevent redundancy in terms of conceptual representations.

However, the assumption made above is difficult to verify in psycholinguistic research. Various studies in psycholinguistics have tried to determine how words are stored. There is, for example, research by Baayen et al. [1993], which is concerned with plural and singular forms of nouns. They find that frequency plays an important role. They use a picture-naming task in which participants are presented with nouns whose plural is more frequent than their singular form, like, for example, the picture of an eye. The researchers notice that the production of the singular form takes more time than the production of the plural form. This is a strong indication for the fact that besides the singular form, the plural form receives a separate entry in the mental lexicon, based on it being more frequent than the singular form. Another, yet less likely theory would be that at some stage, the plural form replaces the singular form in the lexicon, and that now the singular form has to be produced via back-formation.

From the previous section we know that MWEs are complex constructions. In a small scale study, Ströbel [2012] investigates how non-native speakers process idioms like *to kick the bucket* or *to have a skeleton in the closet*. In a reading task in which such expressions appear both in their idiomatic, as well as in their literal sense, participants take less time to process idiomatic expressions. Interestingly, whether the expression occurs in its literal or idiomatic sense does not have an influence. Rather, familiarity and non-familiarity cause faster or slower processing, respectively. These observations hint at the fact that such constructions are indeed saved in the mental lexicon, an assumption that is backed by Chomsky [1981].

Considering the example from (2.4), it is the task of the computational system to assign a V-slot to the expression when it is inserted into the syntax system. This is because we assume that sentences are built from atomic units. As far as generative grammar is concerned, we represent such atomic units as X^0. Thus, the mental lexicon may only consist of such units.

If we turn back to our example (2.1), such lexical units would look like illustrated in (2.7).

(2.7) `N` $\rightarrow$ `Oma`
`P` $\rightarrow$ `sein`
`Adv` $\rightarrow$ `unverhofft`

[6] also see Section 2.1.6

...

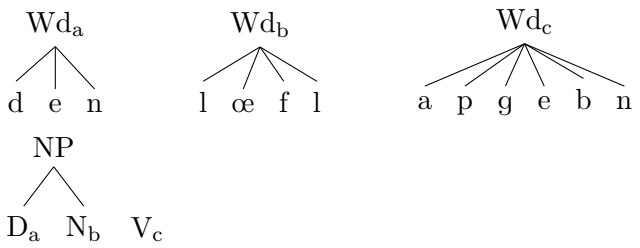
But how to represent the MWE? Compare the two possibilities for *den Löffel abgeben* in (2.8).

(2.8) (a) `VP` → `den Löffel abgeben`

(b) `V` → $[_V[_{VP}[_{NP}[_D den][_N Löffel]][_V abgeben]]$

Represented this way, we can refer to (2.8b) as a lexical verb with internal syntactic structure. This structure, made up of words which are in the mental lexicon already, permits morphological and phonological processes to take place when this expression enters into syntax. In contrast, such variation is not possible if we choose the illustration in (2.8a) to be the correct one. Hence, we now assume that a word is stored in the lexicon not only with its phonological structure (PS), but also with its syntactic structure (SS). We can illustrate this as a tuple $< PS, SS >$. Only then the computational system can fill the slots in syntax. Example (2.9) shows a possible representation of this tuple in the mental lexicon.

(2.9)



We already know the representation of the first line in (2.9) from the concept of phonological structure (see Section 2.1.1). Note that it is still possible to distinguish the single words which constitute the MWE. This is necessary to fill the slots in the syntactic structure, which the second line shows. This kind of "disconnected" syntax permits syntactic flexibility, a characteristic that is not inherent to all types of MWEs. We could also see this structure as an instance of a schema as defined by Booij [2010] (see Section 2.1.1). The alphanumeric indices of the single words in the phonological structure also reappear in syntactic structure, so they denote which syntactic slot can be occupied by which lexical item. Therefore, we assume that there must exist an interface between phonological structrue and syntactic structure, which we call the PS-SS interface (Jackendoff [1997]). In the case of MWEs, the syntactic structure in the computational system does not only have to licence the MWE as X^0, but also the phonetic structure of the constituents it consists of. The slightest modification in one of those words could cause the interpretability of the MWE or the syntactic integrity of the whole sentence to fail. A point in case is the pluralisation of the idiom, as in *die Löffel abgeben*.

To summarise the insights from this section, we can say that items in the mental lexicon consist of X^0 units, which can be used by the computational system to generate and comprehend sentences. Along with the phonological structure, syntactic structure is stored with the lexical items, so that the licensing of certain grammatical structures and phrases can take place. What is missing so far is the role that semantics plays. If we

consider a conventional lexicon, there is no entry without a semantic specification of the lexical entry. This is why we turn to semantics next in order to complete the picture.

2.1.4 How MWEs Construct Meaning - the Aspect of Lexical Semantics

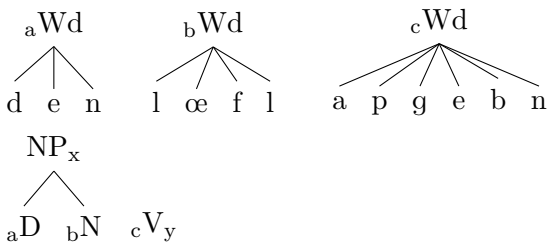
Taking up the discussion in the previous section, it becomes evident that along with phonological and syntactic structure, we have to assign meaning to the lexical entries. Since finding meaning for single word lexical entries is already a difficult task as such, the presence of MWEs further complicates the design of any lexicon. We assume that the mental lexicon has limits and that it is not possible to store the meaning of, for example, all the sentences that exist in the world in the mental lexicon. But how do MWEs behave in this respect?

We have seen in Section 2.1.3 that due to the complexity of MWEs and because they are used as X^0s in the computational system, they have to be stored in the mental lexicon. The presence of phonological and syntactic structure allows us already to produce grammatical sentences. However, since the purpose of language is to communicate in a way that makes sense, there must also exist a representation for meaning in the mental lexicon. Otherwise, non-sense sentences like 'Colorless green ideas sleep furiously' (Lees and Chomsky [1957]) would be produced.

Like the analysis of smaller-scale MWEs suggests, one concept that could help in attaching meaning to MWEs is the **Principle of Compositionality**, which Löbner [2013, p. 15] defines as follows: "The meaning of a complex expression is determined by the lexical meanings of its components, their grammatical meanings and the syntactic structure of the whole." This definition, though, becomes quickly problematic. If we try to infer the meaning of the idiomatic expression in (2.5), which is 'to die', from its constituents, namely *Löffel* and *abgeben* (leaving the function word *den* aside), we arrive at a dead end. Someone who is not familiar with this expression could only infer the right interpretation if she knew that it was common between the 15th and 19th century for servants to receive a spoon from the farmer with whom they worked. In the case that the servants left, they had to hand over the spoon. If a servant died, the spoon went back into the farmer's possession. In either case was the spoon given to the next servant.

Evidently, the majority of native speakers of German do not even know this fact, they just learnt at some stage that *den Löffel abgeben* equals *sterben* 'to die'. This is the case for many MWEs, but not for all. How do we represent this in the mental lexicon? Jackendoff [1997] suggests a visualisation like in (2.10).

(2.10)

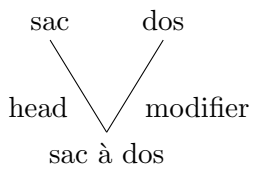


$[STERBEN\,([\quad]_{\mathrm{A}})]_{\mathrm{y}}$

Let us analyse the structure in (2.10), which is slightly different from the structure in (2.9). Our lexical entry is composed of three levels. On the top level[7], the mental lexicon defines the phonological structure of each entry. The syntactic structure occupies the second level, while the third is reserved for what we will henceforth call the conceptual structure. As in (2.9), subscript letters are cross-references. The subscripts of the phonological structure moved to the front and still refer to the specific slots which they can fill in syntactic structure. Subscripts of elements in the syntactic structure now occur after the lexical category, and relate to slots in the conceptual structure. If we put the pieces of the puzzle together, we see that $_{a}$Wd (*den*) from the phonological structure fills the D-position in syntactic structre, $_{b}$Wd (*Löffel*) the N-slot, and $_{c}$Wd (*abgeben*) the position of V. In a similar vein, subscript *y*, which subsumes the whole conceptual structure, signifies that the whole expression maps to a V. Because we assume that the mental lexicon tries to prohibit redundancy, the concept of dying in the conceptual structure could also be replaced by an index[8] which refers to the verb 'to die' in the mental lexicon. The empty slot marked with *A* will be filled with an argument. In case of our example sentence in (2.1), it would be occupied by *seine Oma*.

The visualisation of a lexical entry in (2.10) might suit this special kind of MWEs well. However, there are other kinds of MWEs which might behave differently from our example. In such expressions, the Principle of Compositionality can be applied. Take, for example, the French compound *sac à dos* 'rucksack'. In such a construction, we can decompose the meaning from the single units. That is, *sac* 'bag', *à* 'to', and *dos* 'back', so literally *bag-to-back*, or in other words, a bag that is worn on one's back. We can attribute special kinds of roles to the single constituents of a compound like *sac à dos*. In semantics, we distinguish between **head** and **modifier**. Example (2.11) illustrates this point.

(2.11)



The analysis in (2.11) shows that *sac* is the head, while *dos* is the modifier. Therefore, we speak of a specific kind of bag, namely one that we carry on our back. The pattern *N à N* in French seems to be a very productive one as concerns compounding. It allows constructions like *verre à vin* 'wine glass', *fermeture à glissière* 'zip', or *salle à manger* 'dining room'. Note that the preposition, together with the second noun, does not simply constitute a prepositional phrase, since it is lacking the determiner. Rather, we must take the preposition *à* for acting as some kind of glue which keeps the two nouns together. We refer ones more to the schematic character of such a pattern as highlighted in Section 2.1.1.

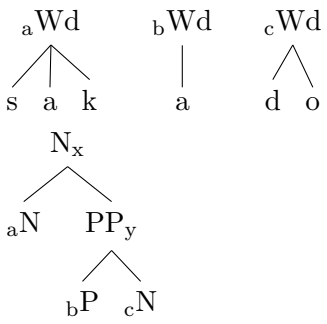
The question at hand is, though, if such expressions are stored in the mental lexicon or if they are produced online when needed (according to the pattern) since the meaning

[7]We do not suggest that order is important here, either level could be the "top" level.

[8]This could also be the case for elements in the phonological structure, as they usually consist of words we already know.

of them can be constructed from their constituents. Given that there is no single word in French which relates to the concept of a rucksack, we must assume that *sac à dos* is a single entry in the mental lexicon. We can represent this as in example 2.12).

(2.12)



$[_{\text{chose}} SAC\, [_{\text{type}} POUR : DOS]_{\text{y}}]_{\text{x}}$

Note that we can deduce a semantic pattern from the French lexical entry in (2.12), which, for a MWE with the same syntactic sturcture as in (2.12), would look like in (2.13).

(2.13)

$[_{\text{chose}}\ (\quad)_{\text{a}}\ [_{\text{type}}\ (\quad)_{\text{c}}]_{\text{y}}]_{\text{x}}$

In a similar fashion, we are now able to construct the senses for *verre à vin* and *fermeture à glissière*.

We now have to adapt the representation of how we store a word in the mental lexicon from a tuple consisting of only phonological and syntactic structure, to one that includes conceptual structure (referred to as CS), as in $< PS, SS, CS >$. This means that we can define each lexical entry by its phonological, syntactic, and conceptual structure. As we have shown, even MWEs, some containing phrasal objects, can obtain lexical status. In the following section, we will further investigate how phrases are constructed.

2.1.5 About Phraseology and MWEs

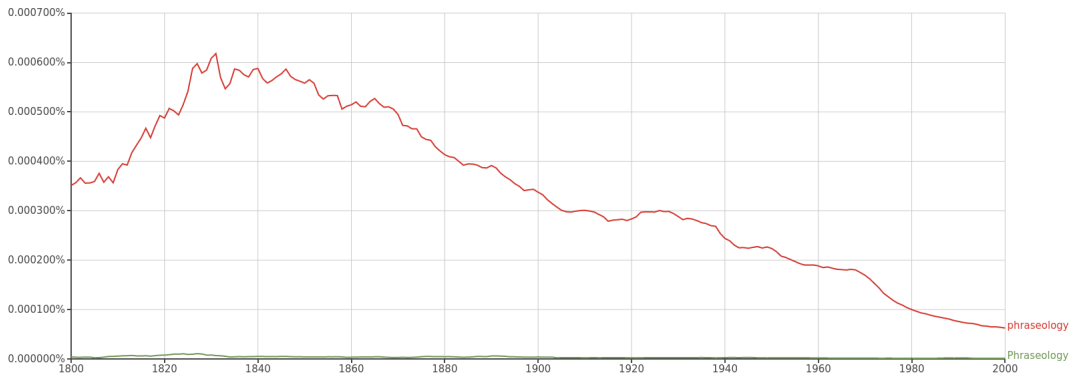


Figure 3: N-gram frequency distribution of the word 'phraseology' over the last 200 years in the Google Ngram Viewer.

From the point of view of a phraseologist, we could say that by looking at a MWE,

we are looking at a chunk, or even a phrase, as we consider two or more "connected"[9] words in a sentence. Indeed, much older than the study of MWEs is the study of phraseology, or at least we can say that the term 'phraseology' has been in use for much longer. As Figure 3 shows, the term 'phraseology' has already been in use in the 19th century. Back then, phraseological studies mainly implied the accumulation of never-ending lists of phrases (see, for example, Duverger [1800]). Still, the overall perception of a phraseologist is that of a phrase collector.

That this view of the field prevails is not astonishing, since phraseology is a field which not only struggles with terminological issues, but also with its stance within linguistics (Gries [2008]; Burger et al. [2007]). Burger et al. [2007], for example, sees phraseology as a subarea of lexicology, while Gries [2008] refrains from a clear positioning of the field. Nonetheless, he makes references to other fields like cognitive linguistics, construction grammar, or corpus linguistics, and hints at the fact that phraseology left its traces in these fields. But what is this field really about?

Phraseology as such is concerned with how groups of words – which are called **phrasemes**, **phraseologisms**, or **phraseological units** – act together in language use. This includes identifying phrasemes, elaborating on their purpose, and defining their meaning. This resembles the work of a lexicologist, which is why we want to employ Burger et al. [2007]'s view here. We visualise this in Figure 4. Figure 4 also clarifies that the core subject of this thesis, namely the MWE, receives the term **phraseme** as an umbrella term, because we perceive every MWE per se as a phraseme. This implies that not all phrasemes are MWEs. From Section 2.1.4, we have learnt that in a MWE, the different constituents contribute to a different degree to the overall meaning of the expression. If we take a look at communicative phrasemes as defined by Stein [2007] and how they are used in discourse, we see that there is a wealth of expressions without compositional meaning. Yet, it is more the pragmatic and communicative purpose which lets us classify the multiword unit as a phraseme (but not as MWE). For example, *mit freundlichen Grüssen* 'with kind regards' is a so-called **routine formula**, which we would classify as a phraseme, but not as a MWE, because of the absence of the compositional meaning component. Note that we differentiate also between MWE and multiword unit. A multiword unit can consist of any sequence of words. A multiword unit can only be a MWE if it shows compositional meaning and fulfils additional criteria which we will describe in the following. Certain multiword units can be phrasemes, but only if they do not possess other characteristics which would allow us to classify the multiword unit as a MWE. Whether or not these kinds of phrasemes have lexical status is an interesting topic for phraseologists who are mainly concerned with how phrasemes come into existence and is beyond the scope of this thesis.

Although the phraseological character of MWEs is obvious, it seems that not many studies with MWEs as their focus make the relation to phraseology. Morevover, researchers who work on MWEs tend to establish their own terminology and are unaware of the actual field they are working in. This affects first and foremost the various phenomena

[9]By "connected" we imply that the words building a MWE are in a certain relation to each other.

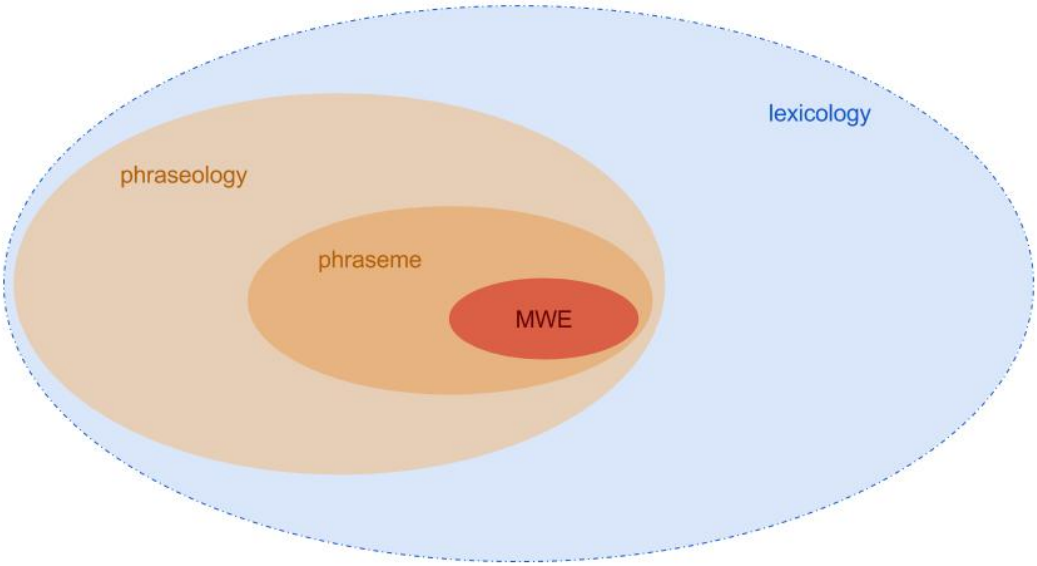


Figure 4: Situating phraseology in linguistics.

that can be observed in this field. Gries [2008, p. 4] even states that "different authors have defined [phraseology] differently, sometimes not providing a clear-cut definition, or conflating several terms that many scholars prefer to distinguish". As a result, the different types of phrasemes were given different names and terms, like **multiword expression**, **idiom**, **fixed expression**, **light verb construction**, **multiword phrase**, etc. Some of these distinctions and classifications may be justified or even necessary. We will come across the one or other term mentioned before as well in Chapter 3. We should bear in mind, though, that if we use these terms, we should take great care in defining exactly what we mean by using them.

Since phraseology in its entierty provides an interesting framework for the characterisation of MWEs, we will elaborate on the definition of a MWE from the point of view of a phraseologist. We adopt the definition of a phraseme that is given by Gries [2008, p. 6] (although he actually calls it **phraseologism**), which says that it "is defined as the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance". Gries [2008] bases this definition on a comparison of several studies which are concerned with phrasemes and lets shine through some criteria which we encounter in corpus linguistics, or more specifically, collocation analysis (see Section 2.1.6). His definition stems from recurring criteria he has found while comparing the studies in the field of phraseology. The basic criteria he identifies ([Gries, 2008, p. 4]) are:

1. the *nature of the units* found in a phrase;

2. the *number of units* in a phrase;
3. the *number of times* the units have to occur together before we can say it is a phrase;
4. the permissible *distance* between units of a phrase and intervening units;
5. the degree of *lexical and semantic flexibility*;
6. the role that *semantic unity* and *semantic non-compositionality* and *non-predictability* play in the definition.

In order to exemplify those criteria, we apply them to our example from (2.1). The first criterion asks for the nature of the different units we find in a MWE. Gries [2008] does not specify exactly what he means by "nature". He mentions, though, that co-occurrences of a lexical item with either another lexical item or a grammatical pattern can be seen as a phraseme. So, what *den Löffel abgeben* makes a phraseme is, among other criteria it fulfils, its syntactic structure, or maybe also its underlying schema if we take into account the insights we have gained from Contruction Morphology.

As concerns the second criterion, the minimum number of units in a phrase must be two. However, not every sequence of lexical units can be denoted as a phraseme. Although the sequence *seine Oma gab* consists of three words, it does not constitute a MWE. Other criteria from above prevent this sequence from being perceived as a MWE. Interestingly, idioms like *the cat's got so.'s tongue* suggest that even sentence-like objects can constitute MWEs, as noted by Jackendoff [1997, p. 163].

Criterion three excludes every combination from being a MWE if the words it is made up of do not occur together often enough. It is well possible that the words in *den Löffel abgeben* occur independently of each other. The deciding factor which lets us perceive this combination as MWE is that they occur more often together as we would expect. The fact that syntactic variance can occur within MWEs poses challenges in determining such frequencies.

As the sentence in (2.1) shows, the distance between units which a MWE consists of can comprise several words. In languages like German, where the boxing of sentences is frequent, there can even be whole sentences in between. Nevertheless, such MWEs as our example still have to be considered as a phraseme. We argue that it is the syntactic structure which justifies this criterion, and that the syntactic structure must be more complex than the one of single words, meaning that it must consist of at least two nodes governed by a mother node before we consider it a MWE.

The fifth criterion refers to changes in conceptual structure. At its core, a MWE consists of words, while each of these words either has its own conceptual structure or fulfils another function (as we have seen with the preposition in French noun compounds). In combination, however, the inherent lexical semantics get blurred. The big picture suggests that we do not necessarily find any of the lexical concepts of the constituents in the expression meaning. Moreover, criterion five hints at the lexical flexibility which happens in the syntactic structure. We have already seen that there are MWEs which are flexible to some extent (take, fore example, the possibility of forming an infinitive

clause with *den Löffel abgeben*), while others, like ‘ more and more’ are much more rigid as concerns their lexical and semantic flexibility.

The last criterion states that the phraseme must function as a semantic unit. The interface between syntactic structure and conceptual structure best exemplifies this criterion. It is, in the end, the combination of the words *den*, *Löffel*, and *abgeben* together with the cultural meaning attached to the MWE which makes this expression a phraseme. However, as we have shown with the routine formula *mit freundlichen Grüssen*, we cannot assign semantic unity to all phrasemes.

Although Gries [2008] mentions that it is not his aim to establish an overall valid terminology that should be used by all linguists, he nonetheless stresses that phrasemes, when analysed, should at least refer to the six criteria which he introduced. In a bigger context, Gries' classification is too strict, since it excludes, for example, routine formulae. However, for the kinds of MWEs which are in the focus of this study we apply Gries' criteria, with the addition of a seventh criterion, namely the **lexical status**. This seventh criterion is motivated by the fact that only atomic units enter the syntax module, and we want to establish the view that MWEs are perceived as atomic units, and that they should be treated as such.

2.1.6 Corpus Linguistics - Collocation Analysis

In corpus linguistics, the study of word co-occurrences which occur together more often than is expected by chance is also referred to as **collocation analysis**. Relating to the field of phraseology, a collocation is thus a kind of phraseme. At the same time, a collocation can also constitute a MWE. This again points to the terminological imbroglio we face when analysing MWEs. Certainly, the term **collocation** is much older than **multiword expression**. Firth [1957] coined it in the late 1950s. With corpus linguistics on its rapid advance from the 1960s onwards, together with technological advancement, phraseology, among other fields, made heavy use of the new techniques for its own purposes. Thus, we can say that, nowadays, corpus linguistic methods are at the core of phraseological analysis.

Mel'ćuk [1998] discusses the issue of collocations and how they relate to phrasemes. Moreover, he explains the importance of **Lexical Functions**[10] when it comes to the deduction of meaning from collocations. He introduces even more terminology: in his view, we have to differentiate between **pragmatic phrasemes** and **semantic phrasemes**. Under pragmatic phrasemes, we find **pragmatemes** and to some degree also idioms, while semantic phrasemes subordinate idioms, collocations, and quasi-idioms.

Where we disagree with Mel'ćuk's analysis is his claim that every phraseme is a linguistic sign in the Saussurean tradition. We regard this statement as too general,

[10] A Lexical Function f assigns a set of synonymous lexical expressions $\{L_i\}$ to the argument of the function, which is generally a lexical unit, in our case a collocation. For example, the lexical function $Degrad(x)$, when applied to *milk*, results in $Degrad(milk)$ = *goes sour* (*Degrad* refers to Latin *degradere* ‘to degrade’). So far, there are around 60 different Lexical Functions (Mel'ćuk [1998]). In terms of collocations, the phrasal verb ‘to give an order’ would be represented as $Oper_1(order) = [to]\ give\ [ART\ \sim\ to\ N]$.

since the representation of the signified in pragmatemes is difficult. If we consider *mit freundlichen Grüssen* once more, which in Mel'ćuk's definition is closest to a pragmateme, we find it difficult to specify the signified of this expression. Moreover, we claim that there is no Lexical Function which expresses the meaning of this phraseme. Thus, although we can think of *mit freundlichen Grüssen* as a collocation, we do not consider it to be a MWE, nor do we share the statement from Mel'ćuk about collocations being **semi-phrasemes** (which hampers the collocability of the pragmateme *mit freundlichen Grüssen*).

Nevertheless, Mel'ćuk [1998, p. 8] also states that "[c]ollocations constitute the absolute majority of phrasemes", which we think is safe to accept. As such, analyses in phraseology greatly profit from corpus linguistic methods concerning collocations. We will explain some of these methods in Chapter 5.

2.2 Summary

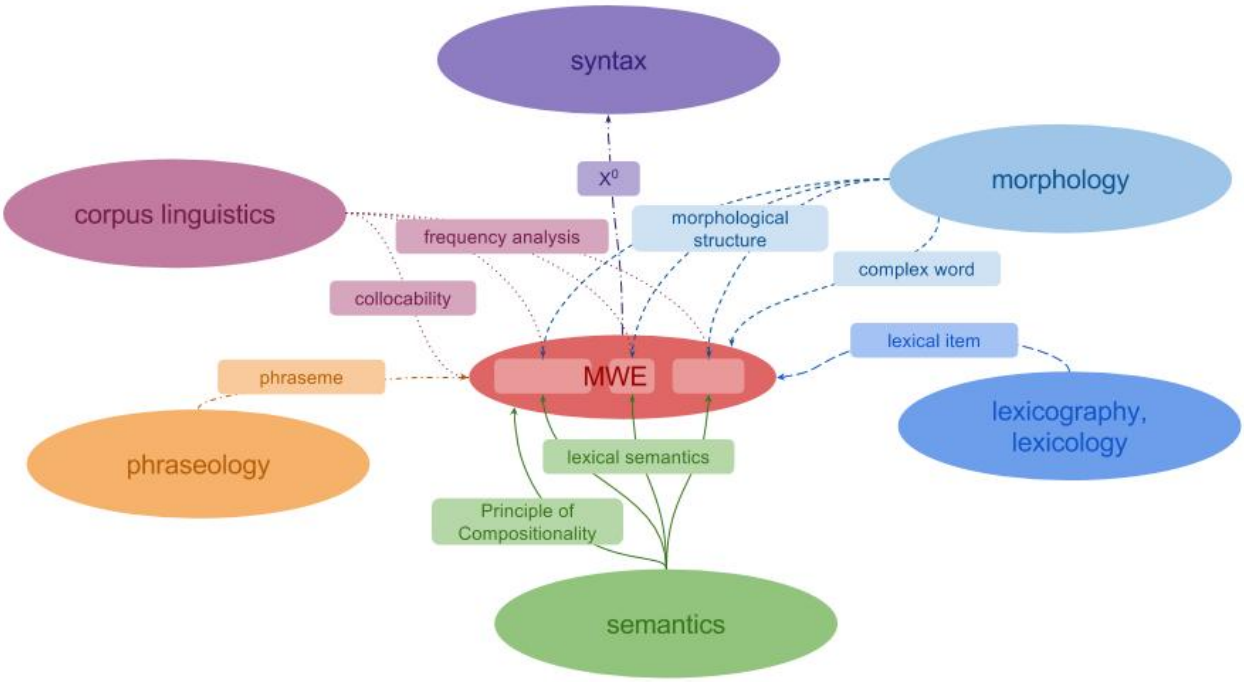


Figure 5: Influences and concepts of various linguistic fields related to the issue of MWEs.

In sum, the purpose of this chapter is to shed light on the many influences MWEs are exposed to. Figure 5 summarises the most important linguistic areas and concepts. Referring to Research Question 1, we would like to briefly skim through them again, starting from the most basic units to bigger ones, while highlighting the essential ideas. In a sense, what follows constitutes the definition of a MWE which we employ in this thesis.

The most basic unit of a MWE is the word. Each word possesses phonological structure and a varying degree of complexity. Furthermore, each word (if we leave function words aside, while they might still serve other purposes) refers to a semantic unit. We say

that each word has a lexical meaning. In compositional MWEs, we can apply the Principle of Compositionality, which allows us to deduce the overall expression meaning from the single components. In all other MWE, the meaning is opaque (as we will see, there are different levels of opacity), and the meaning of a MWE has to be learnt. The MWE constitutes a phraseme as soon as it fulfils seven criteria: (a) the constituents of a MWE form an independent structure, (b) a MWE consists of at least two words, (c) the single constituents occur more often together as predicted by chance, (d) the MWE keeps the permissible distance between its constituents, (e) the MWE and its constituents respect their perhaps limited lexical and semantic flexibility, (f) the constituents of a MWE form a new semantic unit, and (g) the MWE is lexicalised. Then, the MWE can be perceived as an atomic unit (unless it exhibits sentence-like character) and integrates into sentences via the syntax module. Methods of corpus linguistics can be applied to search through languages and extract, identify, and characterise MWEs according to categorisation which we will define in Chapter 3.

In a nutshell, we must place this study first and foremost in the area of phraseology, a subfield of computational lexicography (Heid [2008]), because we exhibit how MWEs and their translations can be extracted via computational methods. What can be taken for granted is the fact that a precise placement of the research of MWEs which takes as much as possible from the related linguistic fields into account is highly beneficial when it comes to the application of computational linguistic methods. The more we know about the nature of this phenomenon, the more precise will our evaluation of the tools be. In this way, we will be able to achieve improvements much faster.

3 The Many Faces of MWEs

Equipped with a clearer concept of MWEs, we turn to the different variants of MWEs that exist in German and French. In general, a common taxonomy of MWEs, or phrasemes, is needed, but especially in phraseology, such classifications are disputed. Müller [1997] postulates a categorisation from the Russian tradition, which classifies phrasemes according to their semantic opacity. The categorisation following in this chapter stems mostly from Sag et al. [2002] and has been extended where appropriate with categories from Baldwin and Kim [2010]. The semantic opacity as described in Müller [1997] also plays a role, but we also consider factors which we have described in the previous chapter. We will illustrate the different types of MEWs with the help of many examples. In the end, we aim at automatically finding translations of MWEs, which is why we will study German and French MWEs, with a specific focus on similarities and differences. Also, we will continuously refer to insights and aspects we gained from Chapter 2. We would like to stress that we take the following criteria for granted, which is why we will not elaborate on them any further when discussing the categories of MWEs: (a) a MWE consists of at least two lexical items, and (b) a MWE constitutes a co-occurrence of lexical items which occur together more often than predicted by chance. Other criteria might not be obvious at first sight, which will justify further explanations.

3.1 Fixed Expressions

As the title of this section already suggests, fixed expressions are MWEs which do not allow variation of any kind. In German *über kurz oder lang* 'sooner or later', *ins Blaue hinein* 'haphazardly', or *im Grossen und Ganzen* 'by and large' are fixed expressions. They can neither undergo variation in the phonological structure (for example, **über kürzer oder länger*) nor in the syntactic structure (for instance, **ins dunkle Blaue hinein*). Moreover, such expressions are fully lexicalised, meaning that they are stored in the mental lexicon. This also implies that in terms of conceptual structure, such expressions have a distinctive meaning attached to them. For instance, *über kurz oder lang* implies that something is going to happen anyway, be it in due time or a bit further in the future. In such a case, the meaning is still compositional, whereas for *ins Blaue hinein*, the meaning is less obvious. Of course, this kind of MWE also occurs in French, where *de temps en temps* 'from time to time', *sens dessus dessous* 'topsy-turvy', or *à perte de vue* 'as far as the eye can reach' are cases in point.

3.2 Semi-Fixed Expressions

To the category of semi-fixed expressions belong MWEs in which the word order is fixed, so syntactic variability is not permitted. Variation in the phonological structure, however, is possible to some degree. We differentiate between three kinds of semi-fixed expressions.

3.2.1 Non-Decomposable Idioms

A non-decomposable idiom is a MWE for which it is not possible to deduce its meaning from its lexical constituents. To this category belongs, for example, the expression *mit jmd. ein Hühnchen zu rupfen haben* 'to have an axe to grind with so.' In French, we find phrases like *faire un tabac* 'to go with the bang', *être/se mettre sur son trente et un* 'to dress up to the nines', or *poser un lapin à qn.* 'to stand so. up'. Moreover, note that variation in the syntactic structure is not possible, and the only possible change in the phonological structure is the inflection of the verb to the past (e.g., *hatte mit ihm ein Hühnchen zu rupfen*) and the variation in reflexive form (e.g., *Vous vous êtes mis sur vos trente et un.*). As concerns conceptual structure, the meaning of such idioms has to be learnt, which is why MWEs of this kind receive a dedicated entry in the mental lexicon.

3.2.2 Compound Nominals

Compound nominals[1] are MWEs that consist of several nouns. In German, the compounding of nouns happens mostly by concatenating nouns, as in *Weinglas* 'wine glass', *Steuerhinterziehung* 'tax evasion', or *Niederlassungserlaubnis* 'settlement permit'. On the top level, all these words consist of two nouns[2]. Compounding as such is a very productive process in German, which is why concatenating more than two nouns is not atypical[3], like in *Arbeiterunfallversicherungsgesetz*, which consists of four nouns, and roughly corresponds to "workmen's compensation act" in English. Note that the nouns which make up these compounds vary in the degree of morphological complexity. While in *Weinglas*, both nouns occur in their simple form, the head in *Steuerhinterziehung*, namely *Hinterziehung* went through different derivational processes (*ziehen* 'to draw' → *Ziehung* 'drawing (of lots)' → *Hinterziehung* 'evasion'). We see as well that sometimes, the linking element *s*[4] occurs to concatenate two nouns, like in *Niederlassung**s**erlaubnis*. Another interesting fact is that, as opposed to French, German noun compounds are right-headed, so in *Steuerhinterziehung*, *Hinterziehung* is modified by *Steuer*, therefore specifying the type of evasion that takes place. As concerns conceptual structure, if for every constituent of the compound a concept is present in the mental lexicon, the inference of the meaning is possible. Therefore, an online decomposition of the compound is probable. There are some nouns,

[1] also called noun compounds by Baldwin and Kim [2010]

[2] Other combinations, like the concatenation of adjectives and nouns is possible as well, as in *Hochhaus* 'sky scraper'.

[3] One of the longest words in the German is *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz*, which consists of seven nouns in total.

[4] Called *Fugen-s* in German.

though, in which the meaning is opaque, like in *Milchgesicht* 'babyface' (which literally translates to 'milk face'). In such a case, the meaning has to be learnt and therefore represented in the mental lexicon.

When we turn our attention to French compound nominals, we notice that they are often linked by a preposition, like in *mal de tête* 'headache', *permis de conduire* 'driving licence' or *marc de café* 'coffee grounds'. This is a typical pattern to build compound nominals in Romance languages and it also occurs with the preposition *à*, like in *verre à vin* 'wine glass'. However, there are also noun compounds which are formed without a preposition, like *chou-fleur* 'cauliflower' or *coupe-papier* 'letter opener'. Note that the word boundaries are marked with a hyphen. Another method to form nouns and which makes use of hyphenation is the concatenation of verbs and nouns, as in *protège-yeux* 'eye protection'. Like the first method to form compounds, compounding by hyphenation is also very productive (Villoing [2012]). Additionally, there are certain nominal compounds which are formed similar to those in English, like *prix normal* 'standard price', or *robinet distributeur* 'tap'. Last but not least, French also makes use of combining nouns and adjectives, which leads to noun compounds like *crise financière* 'financial crisis' (which translates to *Finanzkrise* in German), or *intérêts hypothécaires* 'mortgage interest' (German *Hypothekarzins*).

In contrast to German noun compounds, French compounds are left-headed, as the example in (2.11) shows. As concerns syntactic structure and, therefore, the flexibility of the whole expression, French nominal compounds have to be perceived as nouns which (a) fill an N/NP slot in syntax, and (b) sometimes contain a prepositional phrase. Importantly, the prepositional phrase cannot be moved. Bear in mind that we have already pointed out in Section 2.1.1 that strictly seen, this phrase is not grammatical, but serves the purpose of "gluing" two nouns together to form a new semantic unit. In terms of phonological structure, such types of compounds can be inflected for plural. Interestingly, we would assume that because of the left-headedness of these compound nominals, the pluralisation always happens on the head, as in *verres à vin*. In coordinated noun compounds which do not contain a preposition, however, the plural is marked on both constituents, like in *choux-fleurs*.

The difference of compounding in the two languages poses problems in finding translation equivalents of French and German noun compounds. In the majority of cases, a German noun compound consisting of a single word must be associated with at least two French words. Translating from the other direction, we have to be able to link at least two words to one German word. In the following chapters, we will examine traditional and novel techniques how this can be achieved.

3.2.3 Proper Names

Proper names are a special case of compounding in German in that we usually do not find concatenation of proper names with other nouns. This means that we find occurrences like *Robert Koch Institut*. However, prescriptive approaches postulate that we have to use

hyphenation[5] if we want to link several nouns, which would yield *Robert-Koch-Institut*. This can also be observed in words like *Harry-Potter-Roman*, or names of roads if they contain person names, as in *E.-T.-A.-Hofmann-Strasse*. In French, on the other hand, proper names in compounds occur without hyphenation in the case of institutions, as in *Fondation Albert Viala*, while hyphenation is also present in street names like *Avenue Victor-Hugo*, or places, for example, *place Charles-de-Gaulle*.

3.3 Syntactically Flexible Expressions

Moving from semi-fixed expression to syntactically flexible expressions, we now inpsect how MWEs which allow for syntactic variability behave. To MWEs of this type belong first and foremost expressions which contain verbs. In Baldwin and Kim [2010], such expressions also go by the name 'verb-noun idiomatic combinations', to which also the semi-fixed non-decomposable idioms belong.

3.3.1 Non-Decomposable Idioms

Sag et al. [2002] do not mention this category, as in English, non-decomposable idioms which allow for variability in syntax do not occur. In German, however, we find expressions like *auf den Putz hauen* 'to push the boat out' or our example from Section 2.1, *den Löffel abgeben*. It is not possible to deduce the meaning from its constituents, but they allow some degree of syntactic flexibility as we illustrate in (2.1). In contrast, the English equivalent for *den Löffel abgeben* is *to kick the bucket*. Although being also non-decomposable, the English variant does not allow syntactic variation. Hence, an adaptation of the German syntax as in **His grandmother kicked totally unexpectedly the bucket yesterday* renders the sentence ungrammatical. We would thus assign English non-decomposable idioms to semi-fixed expressions.

3.3.2 Decomposable Idioms

Decomposable idioms comprise expressions like *die Gelegenheit beim Schopf packen* 'to take time by the foreclock', *mit jmdm. ins Gericht gehen* 'to be hard on so.' or *zu tief ins Glas schauen* 'to have had a drop too much'. This type of idiom is decomposable in the sense that it is possible to deduce the meaning from its single words, or rather, from the interplay of the meaning of each constituent. If we take *zu tief ins Glas schauen*, literally *to look too deep into the glass*, we can tell that this must mean that a person drank too much because every time you empty a glass, you will see its bottom, so you are able to look "deep into the glass".

The French language is rich in such expressions, too. This is why we find idioms like *boire comme une trou*, literally 'to drink like a hole', *faire la sourde oreille* 'to turn a deaf ear' and *tirer la couverture à soi* 'to claim everything for oneself'. As with German decomposable idioms, the French variants allow for an interpretation based on each of

[5]This phenomenon is called *Durchkopplung* in German linguistics.

their single constituents. If we want to analyse *boire comme une trou*, *boire* can be taken literally. By specifying the action, which is drinking, with *comme une trou* 'like a hole', we can infer that this must mean that a person drinks too much.

3.3.3 Verb-Particle Constructions

Verb-particle constructions[6] consist of exactly one verb and one compulsory particle. In English, this particle is typically a preposition, like in 'fall out', 'cheer up', or 'wear off', but a verb-adverb combination is also possible, as in 'break down', 'think over', or 'shop around'. The English verb particle constructions most closely correspond to separable prefix verbs in German. Examples for this phenomenon are verbs like *auffallen* 'to stand out', *weitergehen* 'to continue' or 'to walk on', or *aufwarten* 'to come up with sth.' At first sight, these expressions do not look like MWEs. However, when the computational system receives such a word from the mental lexicon and has to integrate it in a sentence, their usage looks like in (3.1a).

(3.1) (a) *Er* ***fiel*** *durch seine gute Arbeit* ***auf****.*
he fall.PAST through his good work up
'He stood out because of his good work.'

(b) *Er auffiel durch seine gute Arbeit.

(c) *Er wird durch seine gute Arbeit auffallen.*
he will through his good work stand-out
'He will stand out because of his good work.'

As we see in (3.1a), the separable verb particle *auf* in the verb *auffallen*, which in this case is a preposition, is located at the end of the matrix clause, while the verb is realised in second position (both parts of the verbs are in bold). This is why Lüdeling [2000] argues for analysing such occurrences as phrasal constructions rather than as single words, an assumption which is backed even from a morphological point of view according to which the particle is perceived as a syntactically independent form in the case it occurs separated from the head verb (Spencer [2015]). Although we do not fully agree with the independence assumption of the separable verb prefix, we acknowledge the fact that it changes its status from a mere prefix morph to a syntactic unit. The construction in which the separable verb prefix and the verb occur together in second position in a sentence yields an ungrammatical sentence, as example sentence (3.1b) shows. The only possibility for separable prefix verbs to not get separated is if we use them in an infinitive construction, as in (3.1c). In terms of phonological structure, separable prefix verbs undergo regular inflectional processes. Their syntactic structure is somewhat harder to capture. Particularly challenging is the representation of the direct objects of transitive separable prefix verbs or intervening prepositional phrases in syntactic structure for which we do not want to stipulate a solution. But the degree of flexibility in certain separable prefix verbs certainly poses problems, not only in the mental representation, but also in terms of NLP, as we will see.

[6] In English as a foreign language education, they go by the name of 'phrasal verbs'.

Interestingly, it is also possible for adjectives to combine with a verb in German, like in of *voll machen* 'to fill sth. up' or 'to round sth. off', or *trocken schlucken* 'to swallow hard'. Both expressions are slightly idiomatic. Although both MWEs consist of an adjective and a verb, they vary in their phonological and syntactic structure. As concerns the former, the adjective in *voll machen* can be compared as example (3.2a) shows.

(3.2) (a) *Dieser Lipgloss macht Lippen voller.*
this lipgloss makes lips fuller
'This lipgloss provides fuller lips.'

(b) *?Er musste trockener schlucken.*
he must.PAST dryer swallow
'He swallowed harder.'

As opposed to (3.2a), a comparison of the adjective in *trocken schlucken* is not possible. The sentence is grammatically sane, but the original meaning is lost. Example (3.2a) shows another peculiarity. While it is possible for *voll machen* to having a noun or a noun phrase inserted, this is not possible for *trocken schlucken*.

In French, we do not find this kind of construction. This must mean that separable prefix verbs, when translated from German into French, are realised as a single word. Indeed, when we take a look at the following example sentences, we see that the French verb *monter*, used in the sense of 'to promote' below, in (3.3) may be translated with the German verb *aufsteigen* in (3.4). The absence of the possibility of modifying a verb by adding verbal prefix to it hints at the fact that either (a) French verbs possess a higher degree of ambiguity or (b) there exist more different verbs in French than in German. For the verb *moter*, option (a) is true, since we can also use this verb to say that we are boarding a train or climbing a mountain.

(3.3) *Le FC Sion monte au premier ligue.*
the FC Sion rises to first league
'The FC Sion promotes to the first league.'

(3.4) *Der FC Sion steigt in die erste Liga auf.*
the FC Sion rises in the first league up
'The FC Sion promotes to the first league.'

However, a construction which resembles the verb-particle construction is the realisation of a verb with an adverb. An example in French is *regarder fixement* 'to stare at so.', which corresponds to the German separable prefix verb *anstarren*. It is important to stress such differences, as they have direct implications on applications in NLP. One such implications is that most probably, German separable prefix verbs will be translated by a single French verb.

3.3.4 Light-Verb Constructions

Light-verb constructions[7] are typically made up of a verb and a noun. They are highly flexible in their phonological structure. The verbal part of a light-verb construction can be inflected for number, aspect, and mood, while the nominal part, which usually occurs in singular and with its article in indefinite form (Samardžić [2008], exceptions are possible), can also be realised as a plural. The syntactic structure of light-verbs first and foremost allows for variation in the nominal part, for example, by modifying it with an adjective. Light-verbs are syntactically more restricted than particle-verb constructions. Examples for light-verb constructions in German are *eine Antwort geben* 'to give an answer', *einen Fehler machen* 'to make a mistake' or *ein Bad nehmen* 'to take a bath'. Taking the first expression, the following example shows different modifications:

(3.5) (a) Modifications in verbal structure

(i) Number

Er	*gibt*	*ihr*	*eine*	*Antwort.*
he	gives	her	a	answer

'He gives her an answer.'

(ii) Aspect

Er	*gab*	*ihr*	*eine*	*Antwort.*
he	give.PAST	her	a	answer

'He gave her an answer.'

(iii) Mood

Er	*muss*	*ihr*	*eine*	*Antwort*	*geben.*
he	must	her	a	answer	give

'He has got to give her an answer'.

(b) Modifications in nominal structure

(i) Pluralisation

Er	*gibt*	*ihr*	*Antworten.*
he	gives	her	answers

'He gives her answers.'

(ii) Adjectival modifier

Er	*gibt*	*ihr*	*nur*	*kurze*	*Antworten.*
he	gives	her	only	short	answers

'He only gives her short answers.'

Note that in such constructions, because of their high idiosyncrasy, the verb cannot always be substituted by a synonym. In this vein, a combination like **einen Fehler tun* is not permissible, while *einen Fehler begehen* is acceptable.

We construe meaning in light-verb constructions differently from regular verb phrases. Consider the example in (3.6). In sentence (a), it is the verb which controls the semantics

[7] Also called *Funktionsverbgefüge* in German.

of the sentence (something is handed over to another person). In sentence (b), we find a light-verb construction, but it is not the verb alone which is responsible for the sentence meaning, which is that somebody answers a question. The noun in the expression takes up much of the semantic space, and the verb becomes, as Samardžić [2008, p. 4] puts it, semantically impoverished. In a sentence like (c), we find that the main verb of the light-verb construction in (b) replaced by a synonymous verb, which does not result in a light-verb construction. This is why we have to interpret the sentence in (c) differently. Instead of someone verbally communicating the answer as in (b), we rather understand that somebody hands over the answer to somebody else in, perhaps, written form. In (c), it is again the verb which projects the semantics of the phrase rather than the noun.

(3.6) (a) *Er gibt ihr einen Stift.*
he gives her a pen
'He gives her a pen.'

(b) *Er gibt ihr die Antwort auf ihre Frage.*
he gives her the answer on her question
'He gives her the answer to her question.'

(c) *Er überreicht ihr die Antwort auf ihre Frage.*
he gives her the answer on her question
'He gives her the answer to her question.'

Of course, we also find such constructions in French. Examples are *prendre un verre* 'to have a drink', *donner un conseil* 'to give advice', or *faire horreur* 'horrify sb.'. They behave similar to their German counterparts.

3.4 Institutionalised Phrases

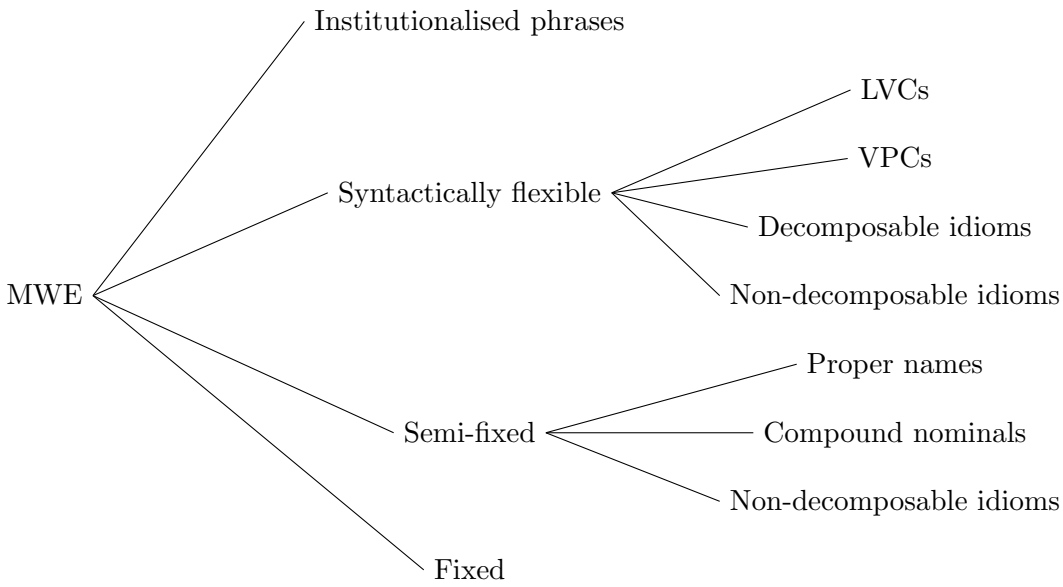
As a last category within MWEs we introduce institutionalised phrases. Both syntactically and semantically, such phrases are fully compositional. From a statistical point of view, however, such expressions are highly idiosyncratic. They correspond closest to what a corpus linguist would call a collocation. Consider the German term *maschinelle Übersetzung* 'machine translation'. There are no rules which dictate that this is the only variant which may denote this concept. The reason why it constitutes a MWE is that the frequency with which the two words occur together is much higher than expected. The same concept could be expressed by the MWE *computergestützte Übersetzung* 'computer-based translation'. However, this term refers to a different concept. In these cases, it is mainly cultural and language developmental aspects which decide which word or expression is used for which concept. In French, the concept of machine translation is expressed with the institutionalised phrase *traduction automatique*. An expression like *traduction machinelle* does not sound too absurd. For the same reason as for the German counter part, if we want to refer to the concept of machine translation, we are obliged to use the former term. The preference over a term is beneficial for native speakers, which are so enabled to express themselves in a clear manner. On the other hand, this poses problems for language learners, who might end up using wrong terms when applying rules and patterns

that are inherent to their own mother tongue. Automatic methods to find translations face exactly the same problems.

3.5 Summary

As we have already mentioned at the beginning of this section, the list of types of different MWEs is not complete. However, we have provided an overview of the major categories. We summarise the categories as an overview in (3.7).

(3.7)



This chapter has highlighted the variety with which MWEs occur in natural language. It has also presented substantial differences between MWEs in German and French, while we have already hinted at issues that are problematic for NLP in some places. The approach taken resembles a contrastive analysis, which is necessary to get a better understanding of how MWEs work in those languages. Only then it is possible to more or less be able to predict how the translation of MWEs between those two languages work and to apply them to other fields of computational linguistics.

4 The Challenge of MWEs in Bilingual Corpora

Multiword expressions are not only of interest to corpus linguists, whose aim it is to study language on the basis of corpus data. Such expressions also pose problems in almost every other field and NLP does not make an exception. With computational linguistics gaining in popularity and applications in many different domains, the way corpora are built and can be accessed has changed dramatically. Corpora grow bigger and bigger, annotations become more faceted and deeper, and multilingual corpus material is not a rarity anymore. With bilingual corpora having become readily available over the last few decades, researchers from different areas are nowadays faced with an even bigger challenge. Due to numerous differences between two or more languages, as small as they may be, the treatment of MWEs has not gotten easier. Problems arise in almost every subfield, beginning from lexicography all the way to statistical machine translation. This chapter summarises recent literature on MWEs in bilingual corpora from different domains.

4.1 MWEs in Lexicography/Lexicology

The task of a lexicographer or lexicologist of a bilingual lexicon consists in finding translations of words and expressions of the source language in the target language. The treatment of MWEs is especially complicated. We infer from the discussions in Chapters 2 and 3 that a lexicographer could be faced with the following questions if she encounters a MWE:

- Is the meaning of a MWE fully compositional or does it need further explanation?
- If the meaning is compositional, is it even worth an entry in the dictionary?
- How is it translated?
- Is there more than one way to translate this MWE?
- If there is more than one way to translate a MWE, how do I find all the possibilities?
- Are there synonymous expressions in the source language and need they be listed in the entry as well?
- If I include a MWE, does this create too much redundancy in the dictionary or can I solve this problem in another way?

We do not intend to answer these questions here. We further claim that this is for sure an incomplete list of questions a lexicographer or lexicologist is faced with when it comes to the translation of MWEs in dictionaries. Of course, with the numerous methods we can choose from to create bilingual or even multilingual dictionaries, the building process can be simplified to some extent.

This has not always been the case. Around 20 years ago, Atkins [1996] dared a look into the future and provided a rudimentary sketch of how the automatic construction of bilingual dictionaries could look like. Keywords that were already present include **extraction**, **comparison**, **partial translation**, and **alignment**. Especially the latter has proven as a useful measure to extract bilingual dictionaries.

However, Atkins' glance into the crystall ball was not absolutely necessary. In the same year, Smadja et al. [1996] presented an algorithm, which they call *Champollion*, for translating collocations from English to French. With their routine, they are able to produce the translation for 'environmental protection', which is *protection de l'environnement* in French. Their algorithm first produces a list of possible collocation candidates in the source language using *XTRACT* (Smadja [1993]). In a subsequent step, Champollion tries to locate the translation of a collocation in the aligned target sentences using the correlation of each possible collocation in the target language sentence. Overall, they report an average precision of 73%. Champollion works without word alignment and without any additional information like, for example, part-of-speech tags.

Nowadays, in order to compile bilingual dictionaries, many researchers take advantage of sophisticated sentence and word alignment algorithms, like the one employed by Dagan et al. [1993] for machine translation. In one such approach, de Caseli et al. [2010] use the alignment to produce a list of Portuguese and English MWEs. After preprocessing their corpus, they extract words from either language which have a 1:n alignment, for example, Portuguese *derrames cerbrais* 'stroke'. Next, they filter the list they have received with the help of a collection of part-of-speech patterns that should not be included. Depending on the frequency of a MWE, they reach a precision of up to 71.1% in automatic comparison to a reference dictionary with subsequent human analysis.

4.2 Terminology Extraction - Finding Exact Matches

In bilingual terminology extraction, a correct and most precise translation of MWEs is desirable, if not essential. As we have seen, statistical measures have their flaws, as do sentence and word alignment techniques, as well as tokenisation, part-of-speech tagging, chunking, and parsing. An interesting fact in terminology extraction is that most MWEs are noun phrases. This fact limits the number of syntactic patterns one has to look for. Moreover, the fact that most part-of-speech taggers, chunkers, and parsers usually perform well with noun phrases benefits terminology extraction. Nevertheless, terminology extraction is still a heavily researched topic, which is why there is a substantial amount of literature which is dedicated to solving this kind of problem.

One of the earliest approaches of bilingual terminology extraction is Dagan and

Church [1994]. They propose a system called *termight*. The aim of this system is the identification and the translation of technical terminology. Termight works out meaningful collocation candidates in the source language based on part-of-speech tags and regular expressions. These candidates are then grouped and sorted. In the bilingual mode, termight tries to find the translation of a term via word alignment. If the found translation is right or not has to be determined by a user, usually a terminologist. Thus, termight is not a fully automatic system to create terminologies, but it is designed to assist a terminologist to find corresponding terms. In the end, termight speeds up the work process, reduces the number of key-strokes, and helps to validate technical terminology.

Daille et al. [1994] focus on the extraction of MWE terminology only and also use a part-of-speech tagged corpus to extract candidates. For this purpose, they use finite-state automata in order to capture morpho-syntactic variation of the predefined part-of-speech patterns. They manage the filtering of the so created list by using statistical measures like mutual information or log-likelihood, among others. They apply this procedure to the English, as well as to the French corpus. Subsequently, they identify translation candidates in two ways:

1. Bilingual count: They determine how many times source and target candidates occur in aligned sentences. By including linguistically motivated features, they are able to fine tune their results.

2. Word alignment: For their candidates on the extracted list, they try to find translations of the terminological multiword unit via word alignment.

They report a precision of 70% for the topmost 1000 candidates on the list.

Moving from rather uninformed terminology extraction to automatic terminology recognition as performed today, Lefever et al. [2009] present a more modern approach. Their aim is to make terminology extraction more language-independent. They claim that it is sufficient to precisely retrieve terminological expressions by linking so-called **anchor chunks**. These links are obtained by a link matrix, which contains the translation of each token in the chunk based on word alignment information. In this way, they create a candidate list which is filtered again by using log-likelihood and mutual information. Compared to a commercial system like *MultiTermExtract*[1], the anchor-chunk-based approach fares up to 20% better.

4.3 Statistical Machine Translation and MWEs

Statistical machine translation is yet another field that suffers from, or lives off, the mere existence of MWEs. How does a statistical machine translation system translate a MWE from one language into the other? In theory, the system should be equipped with additional knowledge. For example, it should know whether it translates from a compounding language into a non-compounding language or vice-versa. However, it is difficult to pass

[1]see `http://www.sdl.com/solution/language/terminology-management/multiterm/extract.html`

along such knowledge in the training phase. One way to tackle this problem is to apply phrase-based machine translation, or even syntax-based systems. Still, problems like long distance relationships are hard to capture.

Nießen and Ney [2004] first and foremost use morpho-syntactic information in order to produce a hierarchical lexicon. Their proposal, which also includes reordering, aims at reducing the amount of bilingual training data. Multiword expressions play a small but important role. They join multiword phrases which their analysers have provided and add them to the dictionary after validation. This has a direct impact on the alignment. With their measures, they are able to reduce the bilingual training data by 10%, while only losing 1.6%[2] in translation quality.

The effect of including MWEs somewhere in the statistical machine translation pipeline is indeed beneficial. Carpuat and Diab [2010] notice that although MWEs might be captured by the phrase tables of state-of-the-art translation systems, they do not treat MWEs explicitly. In their study, they differentiate between a static and dynamic approach. Both methods use a MWE lexicon, which is compiled from *WordNet* (Fellbaum [1998]). The static approach identifies MWEs in the corpus and joins them together into a single token by underscoring. This measure has a direct influence on word alignment and the phrase tables. The dynamic routine, on the other hand, adds MWE knowledge as a feature in the translation lexicon. In this way, the translation system has to decide at coding time whether to translate a word (or a phrase) by a MWE or whether to use a word-by-word translation instead. Evaluation measures do not show significant improvements and the researchers admit that a more sophisticated identification of MWEs would help their take on this subject.

4.4 Summary

We have seen where MWEs pose problems and how they have been tackled up to this point. We can use the insights we gained in the previous chapters and sections for several purposes:

- We can define language-specific syntactic patterns which are able to identify MWEs.
- Due to a contrastive approach, we know which syntactic patterns of one language corresponds to which syntactic entity in the other language.
- We are able to identify weaknesses in the approaches that have been taken so far and provide suggestions for improvement.

We argue that especially linguistic insights, which not all studies take sufficiently into account, will be helpful in the treatment of MWEs in bilingual corpora. As such, if researchers wish to extract and identify MWEs in more than one language at the same time, their methods will heavily benefit from contrastive approaches.

[2]subjective sentence error rate based on human judgement

5 Conventional Measures to Identify MWEs in Corpus Linguistics

This section introduces several measures which have been used to identify MWEs. Because many researchers have developed their own measures which suit their own needs and their own data, the number of such measures is substantial. We decide to only present the most commonly used measures. In the following, we distinguish between rule-based and statistical approaches. Also, getting to know to the peculiarities of the different methods is essential, since we will adopt some of them in our own work.

5.1 Rule-based Measures

Rule-based approaches are common when it comes to either identifying or extracting MWEs from text corpora. For example, in the case of seperable prefix verbs, we can define a routine which re-attaches the separable verb prefix to its head verb. In order to achieve this, we have to assign part-of-speech tags to each word and ideally, we also use parsing information to identify the head verb. Volk et al. [2016b] describe an algorithm (although they do not integrate parsing information) which for each separable verb prefix locates the most recent verb in its finite or imperative form and assigns the correct lemma to the head verb. Example (5.1) shows a sentence in which the identification of the correct lemma is problematic. The third row represents part-of-speech tags corresponding to the Stuttgart-Tübingen-TagSet by Schiller et al. [1995] (STTS). In this sentence, we would like to re-attach the separable verb prefix marked with PTKVZ[1] to the verb, resulting in the lemma *aufwachen* 'to wake up', therefore preventing the incorrect lemma *wachen* 'to keep vigil', which exists in German, but does not represent the meaning of the sentence.

(5.1)	*Er*	*wacht*	*normalerweise*	*um*	*10*	*Uhr*	*auf*	*.*
	he	wakes	usually	at	10	o'clock	up	
	PPER	VVFIN	ADV	APPR	CARD	NN	PTKVZ	$.

'He usually wakes up at 10 o'clock.'

The correct re-attachment heavily depends on the performance of the preprocessing steps. Taggers are not able to recognise long distance dependencies, which is why parsing information would be useful in the case of identifying separable prefix verbs.

We can also formulate rules and algorithms which are based on part-of-speech tag and parsing information for the extraction of nominal compounds or idiomatic expressions.

[1] Called *abgetrennter Verbzusatz* in German.

Nonetheless, there are certain downsides to rule-based approaches, some of which we present here:

- Rule-based approaches which rely on part-of-speech tag or parsing information heavily depend on the performance of the tools which are put to use. If, for example, the tagger had not recognised the PTKVZ in (5.1), the separable prefix verb would not have been identified.
- Sometimes, MWEs are only MWEs when they occur in specific contexts. This means that if we use part-of-speech patterns, chunkers, or parsing information alone to extract MWEs, we cannot distinguish between literal and idiomatic usages of such word combinations.
- Karttunen et al. [1996] start with recognising MWEs at the tokenising step already, using finite-state transducers. In order for this approach to work, they need an extensive lexicon containing MWEs. As we have pointed out above, such an approach is unable to distinguish between, for example, *in general, she is a nice person* and *in general meetings* ([Karttunen et al., 1996, p. 15]). Moreover, the recognition of MWEs at the tokenisation step is problematic, since most part-of-speech taggers assign tags to only single tokens. Nevertheless, it would be practical to pass information along to the tagger and the parser in order to guarantee that the MWEs receive the correct part-of-speech tags and dependencies.

Rule-based approaches can work well for the identification of syntactically simple expressions (as Volk et al. [2016b] show). However, they quickly reach their limits when it comes to complicated phrases which allow for variety in either the phonological or syntactic structure. This is why a vast majority of recent literature has been focusing on statistical methods, including machine learning approaches. We will turn our focus to these measures in the next sections.

5.2 Statistical Measures

Probably the easiest way to find MWEs in a text is using the frequency with which two or more words occur together. But this approach suffers from the fact that combinations like *und der* 'and the', *in die* 'in the', or *von der* 'of the' are much more frequent than real MWEs. Thus, we have to find ways to express the significance of different combinations of words. This is where statistics comes in. Most statistical measures which we nowadays apply to NLP problems have their origins in statistics and information theory, and were probably not intended to solve the problem MWEs pose in the first place. However, we should take a look a the most important measures, also called **association measures** in collocation analysis, which are used in the field.

5.2.1 The *t*-Test

The t-test (Church et al. [1991]) is a widely used means in hypothesis testing. In order for it to be applied, we need a so-called **null hypothesis**. Our aim is to test whether

we can reject it. In the case of statistically verifying the association between a certain combination of words, our null hypothesis is that observing exactly this combination is simply due to chance. If we take the German MWE *über kurz oder lang* 'sooner or later', the null hypothesis states that the combination of exactly these words in exactly this order is coincidental. We calculate the t-value as in (5.2):

(5.2)

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

We understand $\bar{x}$ as the sample mean, s^2 as the sample variance, N as the sample size, and μ as the mean of the distribution. Let us calculate the t-value for the expression above. In the German part of the Text+Berg corpus[2] we find that *über kurz oder lang* occurs 26 times. In total, the corpus comprises 23,051,850 tokens. We also have to determine how many times the single constituents of this MWE occur. For *über* we find 89,142, for *kurz* 5,236, for *oder* 41,717, and for *lang* 3,897 matches. So, what is the probability of this word sequence to occur together? We use the definition of joint probability as shown in (5.3):

(5.3)

$$P(A, B) = P(A)P(B)$$

which says that the probability of two events can be calculated by multiplying the probability of the first event with the probability of the second event. We apply the same formula for more than two events. For our MWE, this gives us the following joint probability:

(5.4)

$$P(\text{über}, kurz, oder, lang) = P(\text{über})P(kurz)P(oder)P(lang)$$

$$= \frac{89,142}{23,051,850} \times \frac{5,236}{23,051,850} \times \frac{41,717}{23,051,850} \times \frac{3,897}{23,051,850} \approx 2.687 \times 10^{-13}$$

We now have the mean of the distribution containing *über*, *kurz*, *oder*, and *lang*, which is $\mu = 2.686 \times 10^{-13}$. The mean of *über kurz oder lang* occurring in the corpus is $\bar{x} = \frac{26}{23051850} \approx 1.128 \times 10^{-6}$. If we insert these values into the equation from (5.2), we get the following result:

(5.5)

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{1.128^{-6} - 2.687^{-13}}{\sqrt{\frac{1.128^{-6}}{23,051,850}}} \approx 5.099$$

[2]as accessed on Nov. 14th 2016 on `http://kitt.cl.uzh.ch/kitt/CQPweb/`

Referring to a confidence level of $\alpha = 0.0005$, for which the critical value is 3.841, we find that our result exceeds this value. Therefore, we can reject the null hypothesis and we can say that there is a very high level of association between the words in our sequence. As useful this test might seem at first sight, Manning and Schütze [1999] state that the t-value itself does not say anything about the "strength" of the collocation, but that the t-test is very useful in ranking collocations. So, if we already have MWE candidates, this test allows us to determine the most relevant MWEs.

5.2.2 Pearson's Chi-Square Test

Because the t-test is based on the assumption that the probability of word pair occurrences are normally distributed, which is generally not true (Church and Mercer [1993, p. 20], Manning and Schütze [1999, p. 169], Evert [2005, p. 83]), Pearson's chi-squared test assumes a χ^2-distribution. The following formula allows us to calculate the chi-square value:

(5.6)

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

where O refers to the observed frequencies and E to the expected frequencies. If we want to determine the collocational bond between, for example, the German adjective-noun collocate $(u, v) = (hoch, Gipfel)$ 'high peak', we can depict the relevant frequencies in a so-called contingency table as in Table 1. Such a table lists the frequencies for all possible

	$v =$ Gipfel		$v \neq$ Gipfel		
$u =$ hoch	1,677	+	31,685	=	33,362
	+		+		
$u \neq$ hoch	29,219	+	22,989,269	=	23,018,488
	= 30,896		= 23,020,954	=	23,051,850

Table 1: Contingency table for observed frequencies of the bigram $(hoch, Gipfel)$ with marginal frequencies.

combinations, which are the occurrence of the lemmas *hoch* and *Gipfel* together (u, v), the occurrence of only *hoch* or *Gipfel* $(u, \neg v)$ and $(\neg u, v)$, and word pairs in which neither *hoch* or *Gipfel* occur $(\neg u, \neg v)$. The numbers next to the table represent the marginal frequencies. The marginal frequencies are simply the sums of the rows and columns, respectively. Taking the sum of the first row, namely 33,362, would mean all occurrences of word pairs in which at least the word *hoch* occurs. We use the marginal frequencies in order to calculate the expected frequencies, which Table 2 contains. We arrive at the expected frequencies by multiplying the marginal frequencies of the respective rows and columns and dividing it by N. For the bigram (u, v) from Table 1, we use the equation $\frac{33{,}362 \times 30{,}896}{23{,}051{,}850} \approx 44.71$. If we put in the observed and expected values from Table 1 and

	v = Gipfel	$v \neq$ Gipfel
u = hoch	44.71	33,317.29
$u \neq$ hoch	30,851.29	22,987,636.71

Table 2: Contingency table for expected frequencies of the bigram $(hoch, Gipfel)$.

Table 2 into the equation from (5.6), we get a χ^2-value of 59752.371. Compare this value to the co-occurrence of any form of the verb *geben* 'to give' with the conjunction *und* 'and' $(geben, und)$, for which the χ^2-value is 0.861. We not only see that it is much lower than the p-value for $(hoch, Berg)$, but we also see that we cannot reject the null hypothesis, since 0.861 is well below the critical value of 3.841. What is quite disturbing and also irritating, though, is that a combination like (und, der) reaches a χ^2-value of 127.568, which would mean that this combination is not due to chance. However, we would not give this co-occurrence the status of a meaningful collocation, or even the one of a MWE. Therefore, in order to determine the strength of the collocation using Pearson's chi-squared test, we are better off using the magnitude of the χ^2-value, besides human judgement. A major drawback, especially for rare collocations, is that Person's chi-square test becomes unreliable when any cell in the contingency tables of observed and expected frequencies contains a value below 5 (Baron et al. [2009]). Thus, there are more elaborate measures to determine collocability.

5.2.3 Log-Likelihood Ratio

Although likelihood ratio tests for collocation analysis (Dunning [1993]) represent a totally different class of test statistics, we can still make use of the contingency tables. The log-likelihood is determined by the sum of the observed frequencies times the log of the observed frequencies normalised by the corresponding expected frequencies, as is shown in (5.7). We refer to the log-likelihood ratio as G^2.

(5.7)

$$G^2 = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

If we apply the formula to the bigram $(hoch, Gipfel)$, the equation returns a value of $G^2 = 3,935.38$. Compared to the $G^2 = 53.89$ for the bigram (und, der), we can see that the collocational bond between $(hoch, Gipfel)$ is much higher. Not only can log-likelihood ratio cope better with sparse data, it is also easier to interpret than the χ^2 statistic ([Manning and Schütze, 1999, p. 173]).

5.2.4 Pointwise Mutual Information

In contrast to the measures presented so far, so-called **point-estimates**, which are based on maximum-likelihood estimates, are not biased towards high-frequency pairs. One such statistical means using point-estimates is pointwise mutual information. We obtain the pointwise mutual information score with the formula in (5.8).

(5.8)

$$pmi = \log \frac{O_{11}}{E_{11}}$$

Note that this formula resembles the log-likelihood statistic. However, we only take the first cell of the contingency tables into account. Thus, we arrive at a pointwise mutual information score of 1.57 for the bigram $(hoch, Gipfel)$. On the other hand, a bigram like (und, der) reaches a score of 0.037, therefore clearly favoring *hoher Gipfel* as a collocaion. A disadvantage of pointwise mutual information is that it tends to be unreliable for low frequency data as well (Evert [2005]).

5.3 Summary

As becomes evident from the discussion above, it is seldom possible to use only rule-based or statistical approaches alone to assess the collocability of a MWE. Rather, we should embrace both methods and use them together. Moreover, we have also seen that the different association measures assume different distributions. We could say that each test makes different assumptions of how a language "works" at its core, therefore returning different results. Thus, we should take great care when we wish to apply association measures, as we cannot randomly choose one. Every method we have encountered so far has certain biases. However, if we choose association measures based on the task at hand, together with informed human judgement, we enable ourselves to gain important insights into how words come and operate together.

6 Word Embeddings

6.1 What are Word Embeddings?

Word embeddings – a buzz-word that drove much of recent research in many domains of NLP. The aim of this chapter is not only to provide a historical account of word embeddings, but also to give an introduction into the technical features of this topic. We will take a glance at the rationale and functionality behind this unsupervised machine learning approach, without focusing on mathematical formulae too much. More important is the concept behind word embeddings and the peculiarities of the different algorithms. Moreover, we will also touch upon applications of word embeddings, especially in the realm of multilingualism, while highlighting research that focuses on MWEs.

6.1.1 A Historical Account of Word Embeddings

In NLP, it has always been a desideratum to automatically learn the semantics of words. With the advent of the internet, search engines like Google wanted to make the web searchable, which boosted research in information retrieval. Given a query, the aim was to find related documents. This is why in the late 1980s and early 1990s, methods like **Latent Semantic Analysis** and **Latent Semantic Indexing** emerged (Dumais et al. [1988], Deerwester et al. [1990]), followed by **Latent Dirichlet Allocation** in the early 2000s (Blei et al. [2001]). These techniques belong to a class called vector space models (we will shortly see why). All of these methods found their way into NLP and considerably helped fields such as topic modeling to flourish. Further refinements led to the success of distributional semantic models, which have since played an important role in word sense disambiguation, sentiment analysis, named entity recognition, and a few other fields in computational linguistics. In general, these distributional models were able to capture the **semantic relatedness** between words and are especially well-suited to determine that the words 'car' and 'road' are related to each other. The question of "how" the two words are related, however, cannot be answered by these methods.

In contrast to latent semantic analysis, word embeddings are a means to capture semantic similarity. Word embeddings are thus able to tell that 'car' and 'vehicle' are very similar words which occur in similar contexts. Basically, we can say that classic distributional methods learn **related** words based on documents in which they occur, while word embeddings are learnt based on the context in which words appear. The distinction between **semantic relatedness** and **semantic similarity** is important and is one of the key differences of distributional semantic models and the models that are used to build word embeddings.

That similar words occur in similar contexts is not a new idea and is also known as the **distributional hypothesis** (Harris [1954]). Both distributional models and word embeddings are driven by the same hypothesis, while the only difference is the definition of what the context is (documents in latent semantic analysis, context words in word embeddings). What is new, however, is how the similarities are learnt. In the past, there have been attempts to learn language models like word embeddings. Especially due to technical complexities these attempts were not practical enough. For instance, in one of the first attempts to train a probabilistic neural language model, it took Bengio et al. [2003] 3 weeks of run-time on 40 CPUs.

This changed in 2013 when Mikolov et al. [2013] released *word2vec* for which they used shallow neural networks to learn word embeddings efficiently. In fact, using neural networks in NLP was not new, but due to some amendments in the algorithms, Mikolov et al. [2013] made word2vec ready to use for the rest of the NLP community. After 2013, word2vec became more and more popular (only take into account the 2500+ papers citing Mikolov et al. [2013]), which is visualised in Figure 6. Their toolkit has been used and further developed ever since and heavily gained in popularity.

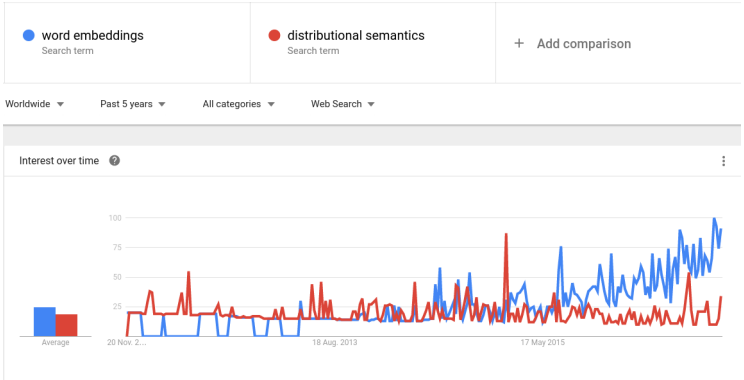


Figure 6: Google Trends search comparing the popularity of 'word embeddings' and 'distributional semantics'.

Most recently, word embeddings have also been used to build multilingual word spaces. The idea behind this approach is that the distributional hypothesis holds across language boundaries, which is why, for example, the German word *Auto* 'car' occurs in similar contexts as the French translation thereof, which is *voiture*. If we now say that each German context word of *Auto* can also be a context word of the French equivalent *voiture*, and vice versa, we only have to find a way to project the German and the French words into the same space, in which, given enough data, translations will eventually occur next to each other. By working with parallel data in which the documents are more or less corresponding translations of each other, it is highly likely that words occur in similar contexts and thus good results for word embeddings are almost guaranteed. In this way, it is not only possible to find word translations, but it is also possible to perform logical deduction, which is accounted for by the most famous equation in (6.1):

(6.1)

$$king - man + woman \approx queen$$

Hence, there is much more to discover in language models which were produced with word embeddings, and finding translations is only one of many useful applications. In the following, we will explain the toolkit developed by Mikolov et al. [2013]. Moreover, we will show what advantages word2vec has over other techniques which deal with semantic distributions, while the focus will be especially on MWEs and multilingual word embeddings. Furthermore, we will encounter some ideas on how certain applications would benefit from semantic models which are able to efficiently capture MWEs.

6.2 What is Behind Word Embeddings?

As already mentioned, word embeddings are generated using unsupervised machine learning algorithms. To be more precise, we use an artificial neural network to produce word embeddings. In the following, we will see how Mikolov et al. [2013] apply neural networks in word2vec.

6.2.1 word2vec

First, we will take a look at how monolingual models are built. Building word embedding models belongs to the realm of unsupervised learning, so basically, we only need huge amounts of raw text. Our aim is the following: given a corpus with a vocabulary V of size $|V|$, for each target word w_t , find its vector $\overrightarrow{v_w}$ so that it minimises the distance of the words which occur within similar context words in context c of size n, while – at the same time – maximising the distance for words which do not occur in the context. n, in general, is a fixed window size, and defines how many words to the left and to the right of the target word w_t are regarded as context. That is a lot of variables which we have to take into account (and there will follow some more). However, we will now do our mise-en-place which allows us to build our word embedding models afterwards.

6.2.1.1 From Words to Vectors

We have already mentioned vectors. They serve us as a valuable tool to represent words in a way which allows us to do calculations with them. Of course, there are different possibilities to represent words as vectors, as already hinted at in the preceding paragraphs. Naturally, each method for a representation has its own peculiarities.

The simplest method to represent a word as a vector is in its so called **one-hot encoding** (or alternatively **1-of-V encoding**). If we take a look at Figure 7, we see this method in a), where we transform 'car' and 'vehicle' into their one-hot encoding. If we traverse through a text, each word can be represented in its one-hot encoding, meaning that the first word is represented as $[1, 0, 0, ..., V]$, the second as $[0, 1, 0, ..., V]$, and so on. Note that in the entire vector, only one value equals 1, while all the other values (standing

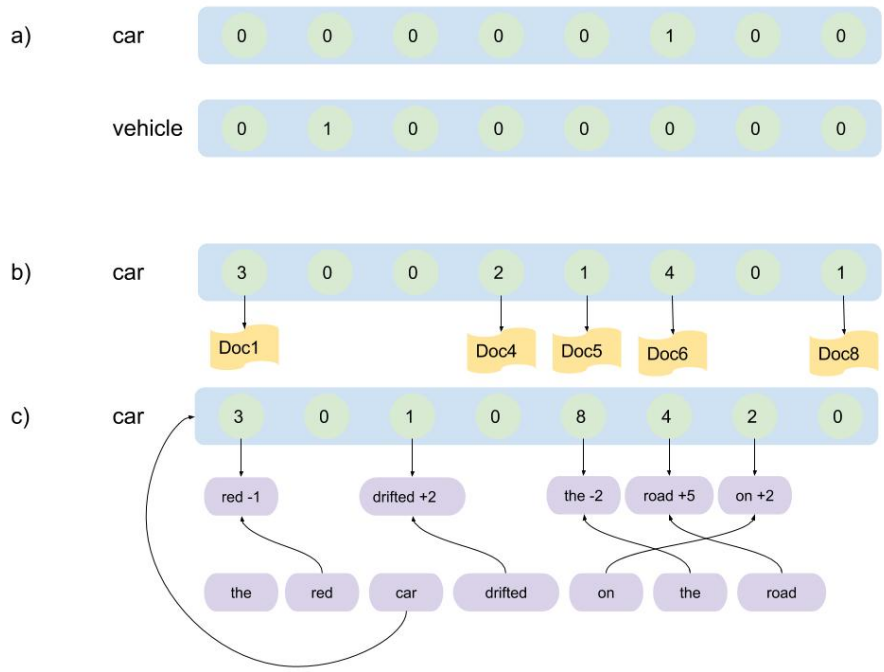


Figure 7: Three different vector representations for the same word. a) is a one-hot encoded vector, b) is based on documents, and c) on the context.

for the rest of the words in the entire vocabulary) are 0s. Given one-hot vectors, it is not possible to prove the fact that 'car' and 'vehicle' are similar words by, for example, calculating the dot product (which would be zero in this case) or the cosine similarity. However, the one-hot encoding will prove useful when dealing with word embeddings, as we will see later on.

If we turn to b) and c) in Figure 7, we see that they both represent two further methods of constructing word vectors for the word 'car'. b) is the representation of the word 'car' based on documents, saying that 'car' occurs three times in document 1, only one time in document 8, etc. c) on the other hand, shows how it is possible to construct a word representation based on the context. The example sentence below the vector represents an instance from which we can build such vectors. We see that, for example, the word 'red' occurs three times together with *car*. This is impossible to say based on just one sentence, but if we assume that we have many more sentences, we would presumably observe that 'red' occurs three times together with 'car'.

Let us stay with the two different methods b) and c) for constructing word vectors. As it turns out, it is vital that we are aware of the distinction between relatedness and similarity mentioned in Section 6.1.1, because depending on which method we choose to construct our vector space model, the similarity (or relatedness) we calculate for two words could be of a different nature. We would like to strengthen this point by looking at an

example involving the four mini-documents in (6.2)[1].

(6.2) 1. seattle seahawks jerseys
2. seattle seahawks highlights
3. denver broncos jerseys
4. denver broncos highlights

Figure 8: Word-document matrix

We would like to know whether 'seattle' is more similar to 'denver', because both of them are cities in the U.S., or if 'seattle' is more similar to 'seahawks', because both the Seattle Seahawks and the Denver Broncos are American Football teams. Depending on the vector representation we choose, we can bring our vector space to either model the former or the latter type of similarity.

Take a look at Figure 8. Because the words 'seattle' and 'seahawks' do not occur in documents 3 and 4, we can infer that documents 1 and 2 are similar to each other, while the same holds for 'denver' and 'broncos'. Hence, 'seattle' is in some way related to 'seahawks'. What this model captures is also referred to as topical relatedness.

In Figure 9, we see that the vectors take context into account. Note that the context matters. If we construct the vectors based on only directly immediate context words (which is why we would disregard that 'highlights' occurs in the context of both 'seattle' and 'broncos'), we would merely see that 'seahawks' and 'broncos' are similar

[1]The idea and some examples for the diagrams were taken from slides by Bhaskar Mitra, which can be found here: `http://www.slideshare.net/BhaskarMitra3/a-simple-introduction-\to-word-embeddings`.

to each other, while we would miss the similarity of 'seattle' and 'denver'. This kind of similarity is typical (because 'seattle' and 'denver' are of the same type, namely cities).

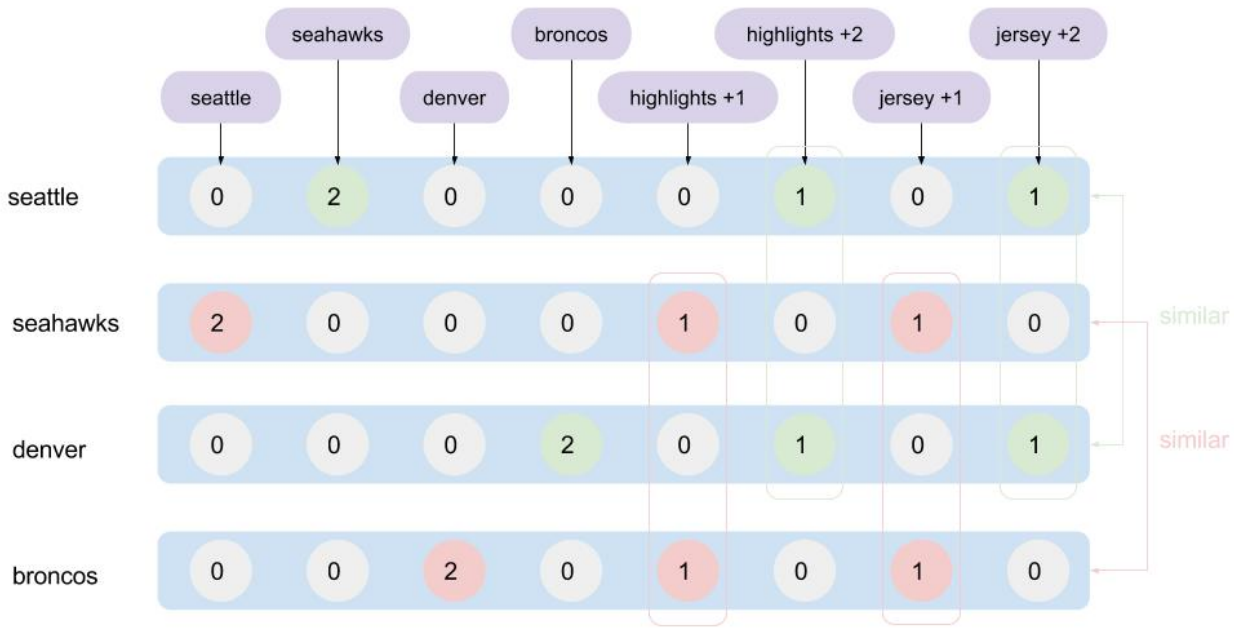


Figure 9: Word co-occurrence matrix which takes different context windows into account

6.2.1.2 N-Gram Language Models

As we have noticed, there are certain design decisions which we have to consider when we build word vectors. One of these design decisions is the size of the context window n. We have seen in Figure 9 that the context window makes the difference. But there are also other factors coming into play when we wish to model a language with word embeddings. Imagine a vocabulary size of several tens of thousands of words. No matter which method we would choose, the resulting vectors will be sparse and high-dimensional. This makes standard approaches impractical. In order to better understand how word embeddings model language, we will make a brief digression to basic language modeling techniques like n-gram models. In fact, word embedding models have very much in common with such language models. Let us take a look at an example sentence (6.3):

(6.3) I stepped on a Corn Flake, now I'm a Cereal Killer.

Language models try to predict the probability of a sequence of words, like, for example, the probability that the sentence in (6.3) occurs. Classic language models are n-gram-based and try to predict the target word w_t (in this case *Flake*) based on n preceding words. Setting n to 3, we would try to predict *Flake* based on the fact that we have already encountered $w_{t-1} = Corn$ and $w_{t-2} = a$, in which case we would speak of a 3-gram model. The relevant probability is depicted in (6.4).

(6.4)

$$p(Flake|Corn, a)$$

If we want to estimate the probability of a whole sentence like (6.3), we simply apply the chain rule under the Markov assumption[2], which means that we multiply the probability of every possible n-gram in the sentence, as described in formula (6.5), partly applied to (6.3) in (6.6).

(6.5)

$$(w_1...w_T) = \prod_i (w_1|w_{i-1}, ..., w_{i-n+1})$$

(6.6)

$$p(I|_-, _-) \times p(stepped|I, _-) \times p(on|stepped, I) \times ...$$

where w_T is a sequence of training words and an underscore symbolises dummy tokens used for the beginning and the end of a sentence. As with word vectors, n-gram models are victims of sparse data. For instance, if we want to translate the sequence in (6.7) from German to French and our trigram language model has never seen the sequence in (6.8), it could be that we end up with a zero score, just because we have never seen the corresponding n-gram during training (note also the bigger n gets, the sparser the data). Of course, we could alleviate this issue by using various smoothing and backoff methods, but this only partly solves the problem.

(6.7) *das* *rote* *Auto*
the red car
'the red car'

(6.8) *la* *voiture* *rouge*
the car red
'the red car'

(6.9) **la* *rouge* *voiture*
'the red car'

Going back to our translation problem, by using backoff or smoothing, our output for the translation could be the one in (6.9), which is obviously wrong. This is where neural network language models come in. They are said to be able to generalize contexts that were not seen during the training phase much better. So, even if (6.9) was not seen in training, a neural network language model would still be able to make predictions about this output, maybe because it has seen *la voiture verte*, from which it infers that the contexts are very similar.

[2]The Markov assumption says that only the previous history of a word matters (so in this case, context to the left) and that this history is limited to k words (Manning and Schütze [1999, p. 193]). We do not make this assumption in word embeddings, since we take context to the right into account, too.

6.2.1.3 Continuous Bag-of-Words

In general, we can say that the continuous bag-of-words model tries to predict a target word w_t based on its surrounding context words. So, if we set n to 3 and try to model on the sentence in (6.3), for the target word *Corn*, the input would be as follows:

- $w_{t-3} = stepped$
- $w_{t-2} = on$
- $w_{t-1} = a$
- $w_{t+1} = Flake$
- $w_{t+2} = now$
- $w_{t+3} = I$

But how does such a model get word embeddings out of this input? The secret lies in the different layers of the model. There are basically three layers in word2vec, each of which will be explained in the following.

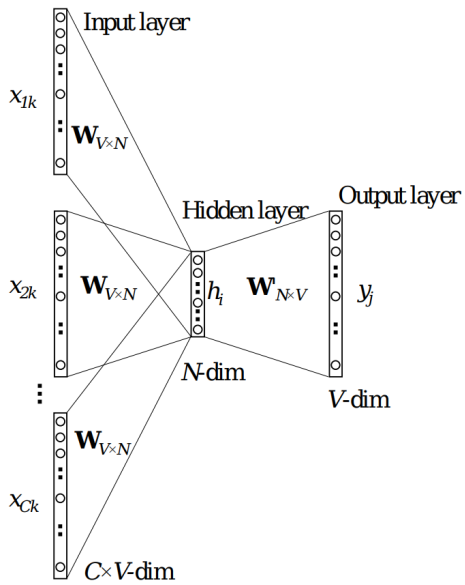


Figure 10: Illustration of continuous bag-of-words model as presented in [Rong, 2014, p. 6].

- **Input Layer** The input layer consists of the context words of target word w_t in their one-hot encoding. These words are represented as $x_{1k}...x_{Ck}$ in Figure 10. The word vectors are passed on to the hidden layer, which is N-dimensional. We will refer to these vectors as input vectors in the following.
- **Hidden Layer** In order to get to the hidden layer, each input vector needs to pass a matrix, called W in Figure 10, which represents weights for each input vector. W is

a $V \times N$ matrix, so there are weights for each word in the vocabulary. Because only one value in each input vector equals one, W, acting as a kind of guard, will pass on the weight vectors to the hidden layer only if the single 1s in the input vectors correspond to a row in W. For example, if the second value in the input vector is one, the second row of weight matrix W will be passed on to the hidden layer (which is essentially just copy/paste).

- **Output Layer** Now we move from the hidden layer to the output layer. In this step, a machine learning algorithm is applied, which for word embeddings typically is stochastic gradient descent. For these calculations, the weights in W' are used. This results in a score for each word in the vocabulary. This is by far not the whole story. In the end, the aim is to maximise the conditional probability of seeing the target word given the input words. This is achieved through backpropagation, so we move back from the output layer to the input layer to update the weights so that similar words get closer, and dissimilar words get further apart. It almost seems as if the weights in matrices W and W' constantly tear at the vectors, making similar words move closer, leaving already similar words untouched and moving dissimilar words further apart. The weights get updated so long until the situation stabilises.

6.2.1.4 Skip-Gram

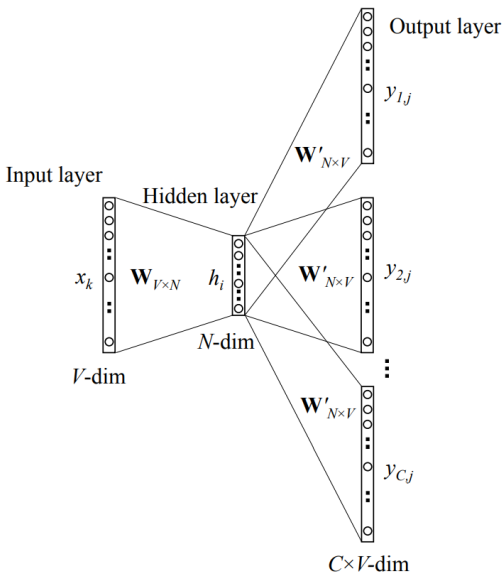


Figure 11: Illustration of skip-gram as presented in Rong [2014, p. 8]

A skip-gram model is the exact opposite of the continuous bag-of-words model. Instead of predicting the target word based on the context words, it tries to predict the context based on a target word w_t. An illustration thereof can be found in Figure 11. Note that the layers are exactly the same (this is the general set up of a shallow neural network model, so we use only one hidden layer), the only difference is that there is only

one word as input vector, but there are several output vectors produced.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Figure 12: Table taken from Mikolov et al. [2013], showing how a vector space model trained with skip-grams captures meaning relationships.

The result, no matter if we apply the continuous bag-of-words or the skip-gram algorithm, is a vector space model which not only captures similar meanings, but also relationships between meanings. Apparently, such models are able to recognise certain structures in the data which cannot be captured by traditional methods like latent semantic analysis. If we take a look at Figure 12, we see that a word embedding model is able to represent the country-capital relationship, or the man-profession relationship. But it also seems to be able to capture grammatical properties, as becomes obvious by looking at 'big - bigger', etc. So, this model successfully represents a positive-comparative relationship.

6.3 Multilingual Word Embeddings and MWEs

The idea of multilingual word embeddings is to project the vocabulary of another language, say L_2, into the same space as the vocabulary of language L_1. Because the distributional hypothesis should hold across languages, translation variants should occur next to each other in the word embedding model. There are different possibilities to construct multilingual word embeddings.

Vulić and Moens [2015] present two algorithms to construct so-called **pseudo-bilingual documents** on which they subsequently train word embeddings. Figure 13 illustrates both of these methods. The first one, called **merge and shuffle**, assumes a parallel and sentence aligned corpus. Going through each sentence, the algorithm randomly chooses between a word from the source language and the target language. The **length-ratio shuffle** takes the mean sentence length of the languages into account. For example, if the mean sentence length of a sentence in the source language is 6 and the one of the target language is 3, the ratio is 0.5, saying that each word of the target language is followed by two words of the source language.

Researchers have already developed toolkits to construct multilingual word embeddings. At the LREC 2016, Berard et al. [2016] presented *multivec*. Multivec unifies various approaches of creating mono- and bilingual word embeddings. We use the toolkit in a test

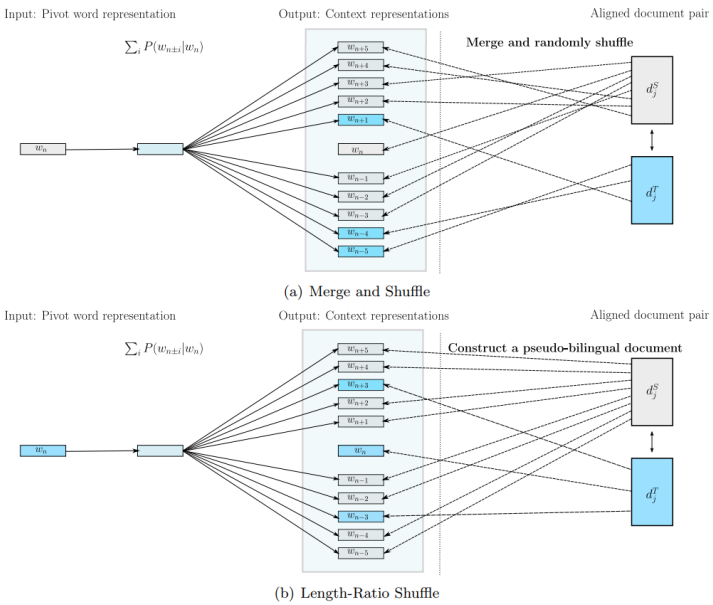


Figure 13: Two different approaches to construct bilingual word embeddings by Vulić and Moens [2015].

on German-English Europarl data (without any preprocessing, which is recommended though). Because word-vectors are high in dimensions (usually between 300 and 1000), we have to reduce their dimensions in order to be able to plot them. We use **Principal Component Analysis**, a common technique to reduce n-dimensional vectors to two dimensions. Figure 14 displays a chosen set of words (English words in green, German words in red). We see indeed that translations occur next to each other. The words *Liebe* and 'love' though, are a bit further apart. This could be due to the dimensionality reduction. In the model, we would not see *Liebe* so close to 'house'. However, we also see that similarities are captured across language boundaries. We thus note that 'Chancellor' and *Premierminister* occur close to each other.

6.4 Word Embeddings and MWEs

Multiword expressions are hard to capture, even for monolingual models. We once again consider Figure 14. We observe that the words 'prime' and 'minister' seem to co-occur relatively often. If we look at the closest words in the model, the two words are indeed not far apart (with a cosine similarity of about 0.77). However, it would be nice to be able to tell that the German word *Premierminister* is translated into English 'Prime Minister'.

Cordeiro et al. [2016] introduce a method which enables us to produce such results. They integrate a preprocessing step which first marks MWEs (preferably in the non-compounding language, so, sticking to our example, we would have done this for English). Next, they construct the multilingual word embeddings on the specially preprocessed

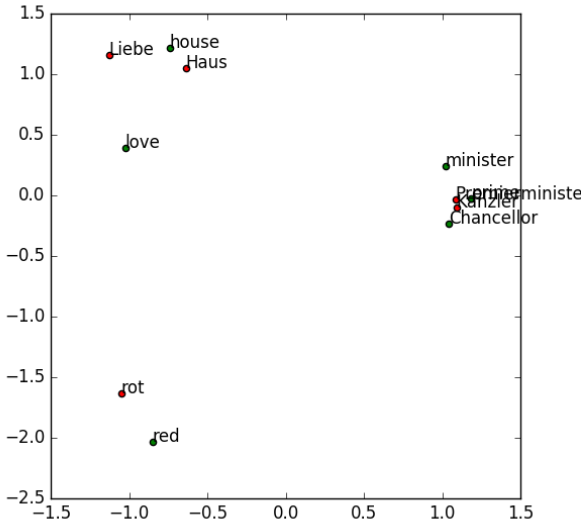


Figure 14: Vector representations in two dimensions of some words from the German-English Europarl corpus constructed with multivec.

data. In our case, this preprocessing step would mark 'Prime Minister' as 'Prime_Minister', which would have the pleasant effect that *Premierminister* and 'Prime_Minister' appeared directly next to each other. In such a way, we can simulate 1:n alignments. With their procedure, Cordeiro et al. [2016] are even able to tell which component of the MWE contributes most to the word meaning (so we would be able to tell which word is the head). However, this method heavily depends on how accurate the detection of MWEs is in the preprocessing step. Cordeiro et al. [2016] make use of the *mwetoolkit*[3] (Ramisch et al. [2010]), in which we have to define patterns with which to extract MWEs ourselves (we will come back to the mwetoolkit in Chapter 7.3.7). This is not a trivial task and also depends on the complexity of the language (for example, when we are talking about long-distance dependencies). It would of course be much more convenient to find a method which automatically recognises MWEs. Nevertheless, their approach is useful and seems to provide promising results.

[3]see `http://mwetoolkit.sourceforge.net`

7 Identifying and Querying MWEs in Bilingual Corpora

The mere availability of corpora is of little use without the proper methods to harvest information from them. As such, one aim of a computational linguist must be to enrich and structure raw text corpora in a way to make them searchable and usable for an end user. The spectrum of users can either be very narrow, for example, if a corpus query system is designed for a particular user group, or very broad, in which case a system is compiled for rather general purposes. Among users who nowadays also rely on corpora in the pursuit of their research are not only linguists per se, but also corpus linguists, lexicologists, language learners, translators, and even historians and geographers.

Our aim is to develop techniques with which we can identify and translate MWEs based on word embeddings. With this goal in mind, we should focus on the development of a corpus qurey system which assists phraseologists, linguists, lexicologists, and foreign language learners. As concerns corpus query systems, there already exist several frameworks, some of which we would like to introduce in Section 7.1. In order to be able to develop such a system, we have to take several steps. Firstly, we have to preprocess the data. Common preprocessing steps are tokenisation, part-of-speech tagging, and parsing. We add the automatic splitting of German compounds, the recognition of MWEs, and the building of bilingual word embeddings to the preprocessing chain. Secondly, much of the intermediate results of the individual steps are stored in a relational database. Thirdly, we develop a search interface which allows to query the database for MWEs. In the following, we will explain each of theses steps with the necessary depth.

7.1 Corpus Query Systems

Online dictionaries like Linguee[1] or Glosbe[2] offer bilingual dictionaries in multiple language pairs. Given single and multiword queries, both systems return probable translation candidates. Moreover, they provide text snippets from the source and target language in which the query terms, as well as their translations, are highlighted. Especially on Linguee the highlighting does not always correspond to the query, which is a disadvantage. In this way, it is much more difficult for a user to capture the context of the word she wants to translate. Context information as such, apart from short text passages, is missing entirely. The user has to decide herself if the translation returned by the system fits the context

[1] see `http://www.linguee.com/`

[2] see `https://glosbe.com/`

she wants to use the word(s) in. Lastly, what both systems lack is an indication of the frequencies of the translations. This forces the user to infer herself which translation is the most probable. However, we see both systems as a valuable resource for finding translations which can be of help to language learners, especially. For linguistic studies, on the other hand, both systems are not sophisticated enough.

Systems which meet the needs of various kinds of researchers are on advance. As the corpora grow bigger and bigger in terms of token size and because the types of annotations – ranging from lemmas over part-of-speech tags, dependencies, phonetic notation, all the way to semantic information – get constantly more diverse, the management of big corpora becomes a challenge. We do not only have to ask ourselves how to store the data, but also how to access it and, perhaps most importantly, how to present it to the user. For purely linguistic motivated usages, there exist systems like the *Corpus Work Bench* (CWB) (Christ [1994]), the update thereof called *Open Corpus Work Bench* (OCWB) (Evert and Hardie [2011]), as well as other frameworks like the one to query gigatoken corpora from the web, the *Corpora from the Web*[3] tool chain (Schäfer [2015]). Many web-based corpus query systems operate with CWB or OCWB in the background. Yet another framework is *SketchEngine*[4] which not only provides ready-to-use corpora for linguistic inquiries, but also offers tools to to compile and query corpora (see Kilgarriff et al. [2014] for a detailed description of the framework and its functionalities). Volk et al. [2014] give a detailed account of corpus query systems, while their focus lies explicitly on parallel concordance systems.

Such parallel concordance systems act as role model for this current thesis. One such system is called *Bilingwis* (**bilingual** **w**ord **i**nformation **s**ystem) (Volk et al. [2011b]) which allows for the exploration of translations in the language pairs German-French, English-Chinese, and German-Rumansh. The text sources differ, covering alpine literature, Swiss law texts and TED talks. Due to the fact that Bilingwis includes diachronic data, sorting translation chronologically is possible, next to the sort option by frequency. Like with query systems based on CWB or OCWB, the lookup of words according to their lemma is implemented as well.

Multilingwis (Clematide et al. [2016]) goes one step further and exploits sentence and word alignment to return results in five languages (i.e., English, French, German, Italian, and Spanish). It is a web-based search tool and stores the entire information on tokens, types, part-of-speech tags, lemmas, parsing information, and sentence and word alignments in a relational database. In contrast to Bilingwis, Multilingwis supports searching for multiword units. For such units Graën et al. [2016] report retrieval times of under 1 second, which is due to the use of materialized views, composite indexes, and pre-planned search functions.

Since Graën et al. [2016] achieve promising results especially in the realm of MWEs, we employ many of their techniques. First and foremost, we use the architecture of the relational database management system. We will elaborate on this further in Section 7.4,

[3]see `http://corporafromtheweb.org/`

[4]see `https://www.sketchengine.co.uk/`

but first we explain the data we use.

7.2 Data

The data which we use for the purpose of identifying MWEs stems from various sources. The following Sections (7.2.1 to 7.2.5) describe the data in more detail. Section 7.3 will elucidate the preprocessing steps. We will round this Chapter off with a few lines on how we integrate word embeddings in the query system in Section 7.5.

7.2.1 Europarl

The largest part of the corpora constitutes the German and French parts from the Europarl corpus (Koehn [2005]). The Europarl corpus consists of written protocols from proceedings of the European Parliament. The current version includes text collections of variable length in 21 European languages. Graën et al. [2014] provide a cleaned version of the Europarl corpus, in which they have corrected encoding and punctuation issues, cleaned the corpus from meta-information occurring in the text, and retrieved lost information from an erroneous treatment of HTML tags. Their procedure results in a loss of about 20% of unnecessary text material, which dramatically enhances the quality of alignments. The so collected German and French part of the Europarl corpus comprises over 40 million words in more than 1.5 million sentences. Table 3 gives a summary of the corpus data.

	de	fr
tokens	41,107,495	47,270,236
sentences	1,753,145	1,691,967

Table 3: Total of tokens and sentences in the German-French parallel part of the Europarl corpus.

7.2.2 Text+Berg

The Text+Berg corpus is a diachronic, multilingual corpus which contains texts from the yearbooks and magazines published by the Swiss Alpine Club[5]. The first publication dates back to 1864. Volk et al. [2010a] describe the acquisition process of the printed versions, text extraction methods and their issues, as well as language identification and named entity recognition. Since the initiation of the Text+Berg project[6], numerous undertakings have been aimed at improving the corpus, such as correction of optical character recognition errors (Volk et al. [2010c], Volk et al. [2011a] and the related KOKOS project[7]), named entity recognition (Volk et al. [2009], Volk et al. [2010b], Ebling et al. [2011]), code-switching detection (Volk and Clematide [2014]), and the correction of part-of-speech tags

[5]see `http://www.sac-cas.ch/`

[6]see `http://textberg.ch/site/en/welcome/`

[7]see `http://kokos.cl.uzh.ch/index.php`

and lemmas (Aepli and Volk [2013], Volk et al. [2016b]), only to name a few. The corpus comes in an XML format and is annotated with article, sentence, word, part-of-speech tag, and lemma information.

Since 1957, the Swiss Alpine Club has published its yearbooks as parallel text. Therefore, the corpus can also be used for research in statistical machine translation (Sennrich and Volk [2011]) and contrastive language studies (Volk et al. [2011b]). It is these texts we use to build the Text+Berg corpus for this thesis. Taking the parallel German and French texts from 1957 to 2015 results in over 400,000 sentences and more than 7.5 million tokens, as Table 4 depicts.

	de	fr
tokens	7,584,124	9,007,498
sentences	407,569	418,188

Table 4: Total of tokens and sentences in the German-French parallel part of the Text+Berg corpus.

7.2.3 Swiss Law Corpus

As another source of parallel texts, we use the Swiss Law Corpus as described by Höfler and Sugisaki [2014]. The corpus comprises the statuary law of the Swiss Confederation, including federal and cantonal constitutions, federal acts, ordinances, and decrees and treaties between the Confederation and individual cantons or municipalities. Table 5 provides an overview of the exact numbers.

	de	fr
tokens	5,996,440	7,792,464
sentences	376,680	376,680

Table 5: Total of tokens and sentences in the German-French parallel part of the Swiss Law Corpus.

7.2.4 Credit Suisse Corpus

The last source texts to compile the final corpus stem from the Credit Suisse Bulletin, which is the world's oldest banking magazine. The first issue of the Credit Suisse Bulletin was published in 1895. As the Credit Suisse Corpus building is still work-in-progress, the corpus only contains text material from 1998 onwards. Included in this corpus is also a large section of the Credit Swiss online news[8]. The compilation of the corpus is described in Volk et al. [2016a]. Being the smallest of all three corpora, it still contains more than 275,000 sentences. In Table 6, we see the token and sentence counts.

[8]Called the Credit Suisse News Corpus. When we refer to the Credit Suisse Corpus, we mean digitized PDFs, as well as the corpus consisting of news articles.

	de	fr
tokens	3,132,943	3,652,304
sentences	283,248	275,816

Table 6: Total of tokens and sentences in the German-French parallel part of the Credit Suisse Corpus.

7.2.5 Summary of the Data

For the experiments in this thesis, we concatenate the corpora described above into one big "super corpus". In sum, we see that this corpus consists of data from four domains: parliament proceedings, alpine literature, law texts, and the finance sector (although the Credit Suisse Corpus also covers topics from economics, lifestyle, sports, etc.). We also see that the proportions are different, as Figure 15 shows. Note that the French part

	de	fr	total
tokens	57,821,002	67,722,502	125,543,504
known lemma types	208,831	31,504	240,335
unknown lemma types	431,179	233,582	664,761
sentences	2,820,642	2,762,651	5,583,293

Table 7: Total of tokens, known and unknown lemmas, and sentences in the German-French parallel "super corpus".

of every parallel corpus is bigger than the German part, in numbers between 15% and 30%. This is due to the non-compounding nature of the French language. A complete overview presents Table 7, including lemma type counts and gross totals. There are several other remarks to be made concerning the data. A fact that might have an influence on the outcome of the experiments is the unproportional distribution of corpus data. The majority of texts stems from the Europarl corpus, while there is less data available from the other corpora. Another problem for the processing of the corpora is that they all come in different formats and have been exposed to different preprocessing methods. We will return to this issue in Section 7.3. Nevertheless, there are several reasons for this kind of corpus design. For one, a credo that is often heard in NLP is "The more data, the better." By including the Text+Berg corpus, the Swiss Law Corpus, and the Credit Suisse Corpus, we see an increase of over 16 million tokens for German while for French, we can add more than 20 million tokens to the final token count. Moreover, the inclusion of texts from other domains might benefit the identification of MWEs. Either we find more MWEs which are different, and thus provide for a greater variation of different MWEs or we find more instances of the same kind of MWEs, therefore gathering more evidence for MWE candidates, or, ideally, both.

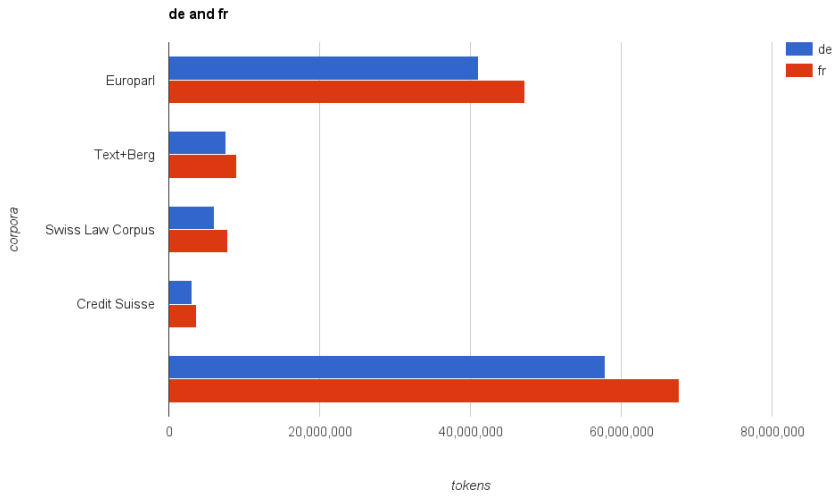


Figure 15: Size comparison of the different corpora.

7.3 Preprocessing

All corpora introduced above have been exposed to different methods of preprocessing. This is mainly due to the fact that the data comes from different sources and in different formats. The base data of the Europarl corpus comes in an XML format, where the data is structured according to chapters and speaker turns. Text+Berg data, on the other hand, is structured in articles. Together with the Swiss Law Corpus and the Credit Swiss Corpus, it has already undergone tokenisation and part-of-speech tagging. However, for some of these corpora, we choose to apply further preprocessing steps. We describe the whole preprocessing procedure in the sections below. We provide a representation of the whole pipeline in Figure 16.

7.3.1 Article Alignment and Cleaning of Text+Berg Data

The Text+Berg corpus relies on automatic procedures for its article alignment. A valuable help is the table-of-contents. However, the article alignment is not perfect. Many article boundaries get lost during the conversion from HTML to XML. Moreover, there are many text passages which are not in the right order. As the quality of the sentence alignment heavily depends on the correct identification of article boundaries, we decide to manually correct the article alignment of the Text+Berg corpus (although the published versions of the corpus provide already tokenised and part-of-speech tagged text in XML format). This results in an increase of parallel articles of almost 30%, from 4952 parallel articles in the uncorrected version to 6393 parallel articles in the corrected version.

During the correction of the article boundaries, we also adjust the order of para-

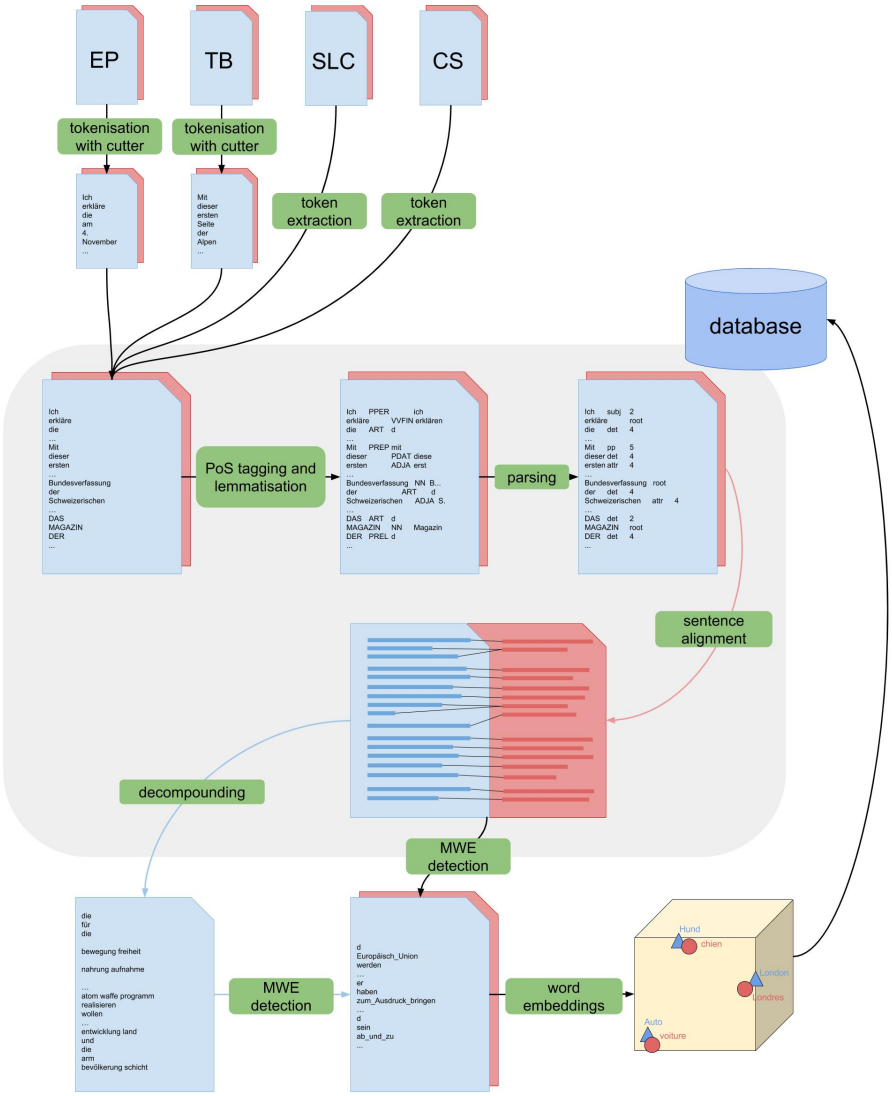


Figure 16: Graphic representation of the preprocessing steps of the corpora used in this study (EP = Europarl, TB = Text+Berg, SLC = Swiss Law Corpus, CS = Credit Suisse Corpus).

graphs where we notice any irregularities. Moreover, we include articles which are in fact translated, but have not been recognised by the pipeline. Furthermore, we exclude large tables consisting only of numbers and only little text. Such tables include, for example, the final revision which the Swiss Alpine Club publishes once a year. We also delete lists and registers of addresses of Swiss Alpine Club huts and tour guides, as they are in fact not or only rarely translated. If they are, they still do not provide essential information, since they mainly consist of names, addresses, and phone numbers. The files prepared in this way then advance to the tokenisation step.

Although the pipeline to create the alignments in the Credit Suisse Corpus looks similar, there is a significant difference. For the PDF issues from 1998 onwards, [Volk et al.,

2016a, p. 91] verify the article boundaries semi-automatically, which results in precision and recall values of about 0.96. We waive a manual correction of the article boundaries for the preceding 15 years mainly because it is a very time consuming endeavour. However, we expect that we would also observe an increase in parallel texts if we manually corrected the article boundaries.

7.3.2 Tokenisation

In this step, we tokenise the data. The Swiss Law Corpus and the Credit Suisse Corpus already come in tokenised form, so the tokenisation step only concerns Europarl and the Text+Berg corpus. For this step, we use a tokeniser called *cutter*[9]. The advantage of this tokeniser is that it uses a hierarchical set of rules and that it leaves tokens which have already been identified untouched. Cutter also provides information on sentence boundaries, which we used to segment the text. We apply cutter to the data from the Europarl and Text+Berg corpus, while for the Swiss Law Corpus and the Credit Suisse Corpus, we use the token information provided by the XML format. By extracting sentence and token information from the two latter corpora, we obtain the same format as is provided by the tokenisation with cutter. Figure 16 exemplifies the tokenisation procedure.

7.3.3 Part-of-speech Tagging

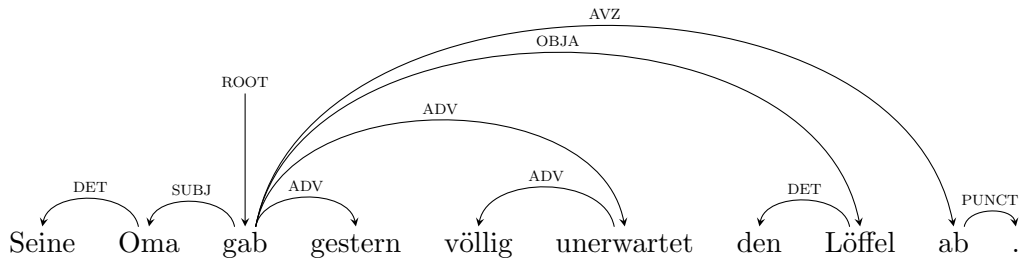
After tokenisation, all the data undergoes part-of-speech tagging. As tagger, we use the *TreeTagger* (Schmid [1994]). In order to reduce the number of unknown lemmas we include the tagger lexicon from Text+Berg. The German parameter file was trained on the German TüBa-D/Z treebank (Telljohann et al. [2004]), which uses the STTS. For the part-of-speech tagging of the French part of the corpus we use the parameter file which was trained on the Le Monde treebank (Abeillé et al. [2003]).

7.3.4 Parsing

The parsing step uses the part-of-speech information and analyses the syntax structure of each sentence. For the parsing process, we use the MaltParser (Nivre et al. [2006]), which takes part-of-speech tagged sentences as input and produces the dependency structure in CoNLL-format[10]. A dependency tree for our introductory example sentence from Chapter 2 produced with the MaltParser looks like in (7.1). It immediately becomes evident why the parsing information can prove useful to detect MWEs. In (7.1) we see that the idiomatic expression *den Löffel abgeben* is realised as ROOT+OBJA+AVZ. We can use this pattern in the next step to extract similar candidates, while it is of course necessary to distinguish between MWE and a conventional grammatical construction.

[9] see `https://pypi.python.org/pypi/applepycutter/1.0.2`

[10] The CoNLL-format is a commonly used format and was first used in the shared tasks of the Conference on Computational Natural Language Learning. Every data row consists of an ID, the surface form, a lemma, a coarse-grained part-of-speech tag, a fine-grained part-of-speech tag, syntactic features, information on the head of the word, the dependency relation to the head, a projective head and a projective dependency relation to the projective head. If positions are unspecified, they are marked with an underscore.



(7.1)

7.3.5 Compound Analysis

Another preprocessing step consists of splitting German noun compounds into their single constituents. Assume we have a German-French sentence pair. In the German sentence, we find the word *Bewegungsfreiheit* 'liberty of action', while in the corresponding French sentence, it is translated as *liberté de manœuvre*. By splitting the German compound into *Bewegung* and *Freiheit*, we not only reduce the number of types, but also make it possible to have projected *Bewegung* and *manœuvre* as well as *Freiheit* and *liberté* close to each other in the bilingual vector space. We will tackle the problem from two sides: on the one hand, by analysing and decompounding German nouns, we will be able to map the constituents of a German compound noun to the corresponding French words, while on the other hand, we use the recognition of MWEs as described in section 7.3.7 to recognise French multiword noun expressions which should map to the noun compound in German.

In our experiments, we use the German morphology analyser *GerTwol* (Haapalainen and Majorin [1994]), which is based on finite-state transducers. We apply it not only to unprocessed, but also to the lemmatised input texts. During the applications phase, we only split words which contain a strong word boundary (marked by a #, as, for example, in *Haus#dach* 'rooftop'). This results, in many cases, in correct splits like *Finanz* 'finance' and *Instrument* 'instrument' from *Finanzinstrument* 'financial instrument'. However, incorrect analyses also occur due to the internal ambiguity of a word. For instance, the word *Haushaltsplan* 'budget' gets analysed as *Haus* 'house', *Halt* 'stop', and *Plan* 'plan', whereas it should only be *Haushalt* 'household' and *Plan*. Nonetheless, decompounding has found to be beneficial not only for statistical machine translation (Brown [2002]), but also for cross-lingual information retrieval (Chen [2002]). The more concrete setup of the experiment will be detailed in Section 7.3.8, where we will explain how we incorporate the result from this analysis in the training of word embeddings. The outcome of the decompounding step applied to our "super corpus" results in an input file consisting of decomposed tokens and an input file for decomposed lemmas.

7.3.6 Sentence Alignment

For the purpose of sentence alignment, we use *hunalign*, an unsupervised sentence aligner developed by Varga et al. [2007]. According to Abdul-Rauf et al. [2010], hunalign can keep up with other state-of-the-art aligners. Sentence alignment is necessary in our work in order to reduce search space given a query. In this fashion, we only look for translations

in sentences in the target language which are directly aligned with languages of the source language, and, at the same time, contain the query term.

7.3.7 Recognition of MWEs

In order to be able to identify MWEs in the multilingual word embedding space, we decide to mark MWEs in the training data, thereby following the approach from Cordeiro et al. [2016]. For this purpose, we use the *mwetoolkit* devolped by Ramisch et al. [2010]. In the following, we will outline which patterns we implement for the mwetoolkit. The motivation for the use of these patterns stems from Chapters 2 and 3. We focus on the recognition of some fixed expressions, noun compounds, verb-particle constructions, and lexical verb constructions. We disregard idiomatic expressions, as they are hard to capture, mainly because many of such expressions can take almost sentence-like form. The recognition of MWEs happens in two steps: first, we need to extract MWE candidates. From these candidates, we will compile a list with which we annotate the input files for the word embeddings in a second step. The exact procedure is described in the following.

7.3.7.1 Patterns for mwetoolkit

In the case of fixed expressions, it is difficult to find regular patterns which allow for a systematic identification of such expressions. However, the fact that this type of MWE can only occur in one specific order already restricts the search space. Another limiting factor is that there are certain patterns which we could use both for German and French. We focus on bi-particle adverbs for the experiments. In German, a pattern which describes such a type of adverb is adverb-conjunction-adverb (ADV KON ADV), in which the adverbs can also be replaced by adjectives (ADJD KON ADJD). In French, we can use the corresponding pattern (A_qual C_C A_qual).

As concerns noun compounds in German, the implementation of a pattern that locates noun or named entity sequences introduces too much noise. We therefore exclude noun-noun compounds which occur separate from each other from further analysis. In French, on the other hand, there exist several productive patterns which can be used to identify MWEs. Among the most prominent is the noun-preposition-noun (N_C P N_C) combination. Moreover, we include the lookup of patterns of the form noun-adjective (N_C A_qual). In very rare cases, the stringing together of nouns occurs as well. Hence, we add the pattern (N_C N_C), as it might be helpful to identify proper names.

German verb-particle constructions offer the opportunity to include parsing information. When separable prefix verbs occur with their separable prefix located elsewhere in the sentence, the parser marks the verb-prefix relation as AVZ. In adjective-verb combinations like *voll machen* 'to fill sth. up', the relation to the verb receives a PRED tag. As such, useful relations are ROOT-AVZ and ROOT-PRED, while the AVZ must be tagged as either PTKVZ or ADJD (adverbial or predicative adjective). The French language does not use such verbs.

Last, lexical verb constructions also heavily depend on information provided by the

parser. Generally, we are interested in verb-noun occurrences in which the meaning of the verb alters. The two parts of this construction do not have to follow each other, meaning there can be variation in syntactic structure. The pattern defined for the identification of lexical verb constructions is thus rather general. Nevertheless, and mainly due to the fact that the mwetoolkit proves the collocability of a MWE by applying different association measures, we include the pattern ROOT-OBJA for German, and root-obj for French.

Table 8 provides an overview of the patterns used with the mwetoolkit[11]. In some cases, we use regular expression style notation in the table, for example, for matching at least two succeeding nouns or proper nouns. Moreover, the asterisk must be interpreted as a Kleene star. It only occurs once, namely in the pattern matching lexical verb constructions, where we want to match every possible verb form.

German		
MWE type	pattern	example(s)
Fixed expressions		
	ADV (KON|KOKOM) ADV	*mehr oder weniger* 'more or less', *hie und da* 'from time to time'
	ADJD (KON|KOKOM) ADJD	*weit und breit* 'far and wide', *näher und näher* 'closer and closer'
VPCs		
	ROOT → (VVFIN|VVIMP) AVZ → PTKVZ	*aufmachen* 'to open', *zusehen* 'to watch'
	ROOT → V* PRED → ADJD	*vollmachen* 'to fill sth. up', *trocken schlucken* 'to swallow hard'
LVCs		
	ROOT → VV* OBJA → NN	*Antwort geben* 'to give an answer', *in Anspruch nehmen* 'to make demands on'
French		
MWE type	pattern	example(s)
Fixed expressions		
	ADV C_C ADV	*plus ou moins* 'more or less', *tel ou tel* 'this or that'
	A_qual C_C A_qual	*sains et saufs* 'safe and sound', *jeunes et vieux* 'young and old'
Noun compounds		
	N_C P N_C	*camp de base* 'base camp', *chemin de fer* 'railway'
	N_C A_qual	*crise financière* 'financial crisis', *intérêt hypothécaires* 'mortages interest'
	N_C N_C	*état membre* 'member state', *robinet distributeur* 'tap'
Proper names		
	(N_C|N_P){2,}	*Maurice Zwahlen*
	N_P P N_P	*Mer du Nord* 'North Sea'
LVCs		
	root → V* obj → N_C	*prendre une décision* 'take a decision', *donner une réponse* 'to give an answer'

Table 8: Patterns for the automatic extraction of MWEs. For the German fixed expressions, we additionally include the alternative of a comparing conjunction (KOKOM), as well as the following alternatives for particles: pronomial adverb (PAV), adverbial interrogative or relative pronoun (PWAV), preposition (APPR), and separable verb prefix (PTKVZ).

7.3.7.2 Evaluation of Extracted Patterns

Not every multiword unit that the mwetoolkit extracts with the aforementioned patterns is a meaningful MWE. Therefore, we have to rank the candidates returned by mwetoolkit. Conveniently, the toolkit comes equipped with the implementation of five different association measures. In this way, we rank the candidate list from each category (fixed

[11]Both German and French patterns are listed in Appendix A.

expression, noun compounds, etc.) according to the five built-in metrics, which are log-likelihood, pointwise mutual information, the t-test, Maximum Likelihood Estimation, and the Dice coefficient. This leaves us with five ranked lists for each category. Bear in mind that we have mentioned in Chapter 5 that certain methods are biased towards certain distributions or frequencies. On inspection of the candidate list for German light-verb constructions sorted according to pointwise mutual information scores, for example, we see that the MWE *zum Taugenichts abstempeln*, which roughly translates to 'make so. a good-for-nothing', is among the top 20 candidates. This candidate is indeed a very good example for a MWE, but it only occurs once in the corpus, which is why we must not rely too much on the scores of this measure (remember that pointwise mutual information tends to be unreliable for low frequency data, in this case, it seems to favour rare events). However, our aim is to annotate as many true MWEs in the input data as possible, which is why we deem it necessary to compare the measures and manually inspect the candidate lists in order to decide which measure captures best a distinct pattern. This evaluation is based on human judgement from one annotator. For every sorted candidate list of each category, the annotator looks at the top 100 candidates and decides for each candidate whether it belongs to the category or not. We provide a summary of the evaluation in Table 9[12].

As Table 9 shows, the t-score captures German fixed expressions and light-verb constructions best, while maximum likelihood expectation is the most reliable measure for verb-particle constructions. For French, we find that the t-test outperforms the other measures in the identification of relevant fixed expressions and light-verb constructions (similar to German). For nominal compounds, two measures are even, namely the Dice coefficient and maximum likelihood estimation. On further evaluation we notice that the Dice coefficient favours rare events, and nominal compounds consisting of proper nouns like 'Burkina Faso' especially. For this reason, we decide to continue with the candidate list ranked according to maximum likelihood estimation scores. For the other lists, we choose the one with the score in bold in Table 9.

Table 10 shows the total number of MWEs for each category returned by the mwe-toolkit in the first column. Including the entire candidate lists in the MWE annotation part would result in many wrongly annotated MWEs. For this reason, we decide to compile a restricted set of MWEs. This step also simplifies the comparison to the identification of MWEs using word alignment information in a subsequent evaluation (see Chapter 8). In order to arrive at a clean list of only true positives, we manually filter the remaining six ranked lists. The manual filtering is necessary because there are still false positives among the top candidates. The final MWE list should thus contain only candidates which are true positives. This leaves us with a much smaller number of MWE candidates, as we show in the second column of Table 10. Note that the cut-off point in the final column is arbitrary and that, given further evaluation, we do not per se exclude the integration of additional MWEs.

[12] We are aware of the fact that for representative results, it would have been necessary to carry out an evaluation experiment with several annotators.

German	*expression type*	*Dice coefficient*	*MLE*	*LLR*	*PMI*	*t-test*
	fixed expressions	27	53	N/A	0	**58**
	LVCs	0	42	49	0	**59**
	VPCs	44	**74**	64	35	73
French						
	fixed expressions	18	22	N/A	0	**25**
	nominal compounds	***97***	***97***	91	0	96
	LVCs	33	31	35	0	**38**

Table 9: Small human evaluation of different association measures offered by the mwetoolkit. The evaluation shows the precision of the 100 top ranked candidates in each list. The information for the log-likelihood ratio of fixed expressions is not available, since it has only been implemented for bi-grams in the mwetoolkit.

German	*expression type*	*total count*	*final candidates*	*cut-off value*
	fixed expressions	22,831	120	$t = 2.35$
	lexical verb constructions	404,940	306	$t = 8.92$
	verb particle constructions	14720	187	$mle = 1.21e^{-7}$
	total	442,491	513	
French				
	fixed expressions	46,528	36	$t = 2.82$
	noun compounds	983,694	1,101	$mle = 9.45e^{-7}$
	lexical verb constructions	312,063	59	$t = 26.18$
	total	1,342,258	1196	

Table 10: Overview of total extracted MWEs per category and final number of MWEs for further processing and evaluation.

7.3.7.3 Annotation of MWEs

In the last step, we annotate the German and French part of our "super corpus" with mwetoolkit and the restricted set of patterns we describe above. We apply the recognition of MWEs once for the raw "super corpus" consisting of lemmas, and once for the raw "super corpus". In the German part, this procedure results in the annotation of 115,890 MWEs for raw text and 118,493 MWEs for lemmatised input material, respectively. The annotation of MWEs in the part of the corpus which we have treated with GerTwol is less successful, with only 96 annotated MWEs for raw input and 23,818 for lemmatised text. This is due to the fact that many nouns have been split during the application of GerTwol. For the French part of the corpus, we reach a value of annotated MWEs of 834,198 for the lemmatised input, while for raw text input the total count amounts to 826,889 MWEs. The imbalance between the total numbers of annotated MWEs in the German and French part of the corpus arises from that on the one hand, we include twice as many pattern

in French than we do for German, and on the other hand, most French MWE candidates are nouns, which we assume to be more frequent than, for instance, German lexical verb constructions.

As input and output for the annotation serves the CoNLL format. On finding a MWE in a sentence, the mwetoolkit "glues" the strings together. For example, if the mwetoolkit finds the *bateau de pêche* 'fishing boat' in the input, it writes all constituents of the MWE on one line. Figure 17 exemplifies this procedure.

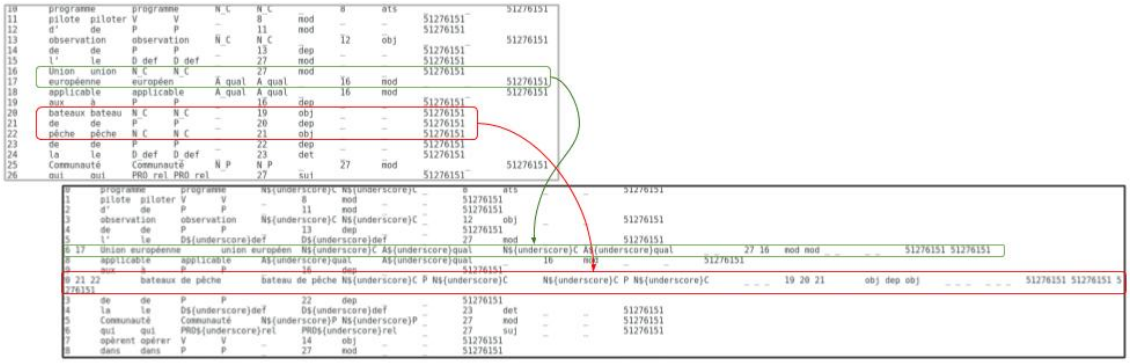


Figure 17: Concatenation of a MWE performed by *mwetoolkit*

In order to include this information in the training phase for the bilingual word embeddings, we concatenate the strings by underscoring, transforming the MWE separated by spaces from the CoNLL format to *bateau_de_pêche* in the input file for the word embeddings training phase. An evaluation on the annotation of MWEs with our candidate lists in both the German and the French corpus consisting of only lemmas show that we reach a precision of over 99%. The source of misidentification lies in either wrongly assigned part-of-speech tags, mistakes during parsing, or a greedy search by mwetoolkit. From these sources of errors follow wrong German MWEs like *Entscheidung_treffen_treffen* (greedy search), *und_aus* (probably wrongly tagged), and *Verantwortung_übernehmen_tragen*, as well as French false positives like *de_énergie*, *temps_de_travail_travail*, or *prendre_acte_le_initiative* (greedy search). Recall, on the other hand, is difficult, if not impossible to determine for such a big corpus.

7.3.8 Word Embeddings

For the training phase, we use different parameters in order to train different models. With *multivec* (see Section 6.3 of Chapter 6), it is possible to assign different values to parameters like the size of the context window, dimensionality, or if multivec should use the skip-gram or continuous bag-of-words algorithm. Since we do not know beforehand which configuration works best for the identification of MWE translations, we train multiple models. Figure 11 gives an overview of the variations of parameters. We keep the learning rate at 0.05 and the minimum word count at 5 for all models (the minimum word count excludes all the words from training which occur with a smaller frequency than defined).

This training procedure yields a total of 48 models. In order to choose the best model for the system, we perform the following evaluation: We randomly choose 100 MWEs from the German and French candidate list we used for training. We then calculate the closest word for each of these candidates in the other language based on cosine similarity. Next, we determine if this is a valid translation based on human judgement. For example, if for the German MWE *Anerkennung zollen* 'to pay tribute to so.' a word embedding model produces the translation 'rendre hommage', we award 1 point to this specific translation. If at least a part is correctly translated, for instance, if the model returns *réponse* 'answer' for the German MWE *Antwort geben* 'to give an answer', we still assign half a point. A translation receives zero points if it is wrong in its entirety.

For the models with GerTwol data in the input, we calculate the closest word for each of the constituent of the MWE and determine if the so generated output is a valid translation. To make an example, if we want to find the translation of *Gelegenheit geben* 'to give so. the opportunity', we assess the translation quality of the model on its return of the best translation for *Gelegenheit* and *geben*. In this concrete case, the system might return *occasion donner*, which we would rate with 1, despite of the wrong word order. Note that we do not have similarity scores for this category, since there is no similarity score for the whole MWE. We present and further elaborate on the results of this evaluation in Chapter 8.

German input		**French input**	**algorithm**	**context window**	**dimensions**
raw+MWE	↔	raw+MWE	sg	4	300
raw+GerTwol	↔	raw+MWE		8	
lemmatised+MWE	↔	lemmatised+MWE	cbow	12	600
lemmatised+GerTwol	↔	lemmatised+MWE			

Table 11: Overview of model parameters to train word embeddings (sg = skip-gram, cbow = continuous bag-of-words).

7.4 The Database

When it comes to the management of large corpora, the questions of how to store and how to access them arise. These questions are also influenced by the purpose of the corpora. Raw text files might suffice if we just want to store the text, or if we want to train statistical machine translation systems. Corpora in XML-format can also be useful for storage purposes. However, if we want to make corpora searchable in an efficient way and if we want to retrieve information from them, text or XML files are not practical. A good means for storing larger corpora provide relational databases. Graën et al. [2016] have successfully implemented a corpus query system based on a relational database with data from the Europarl corpus. The Multilingwis system is capable of a parallel search which returns results for no less than five languages, while more languages are going to be available in due time. We mainly apply Graën et al. [2016]'s database design, with the

exception of adding one table for word embedding information, and dropping the table for word alignments (since we want to rely on word embedding information only). In the following, we introduce the database design, which we visualise in Figure 18.

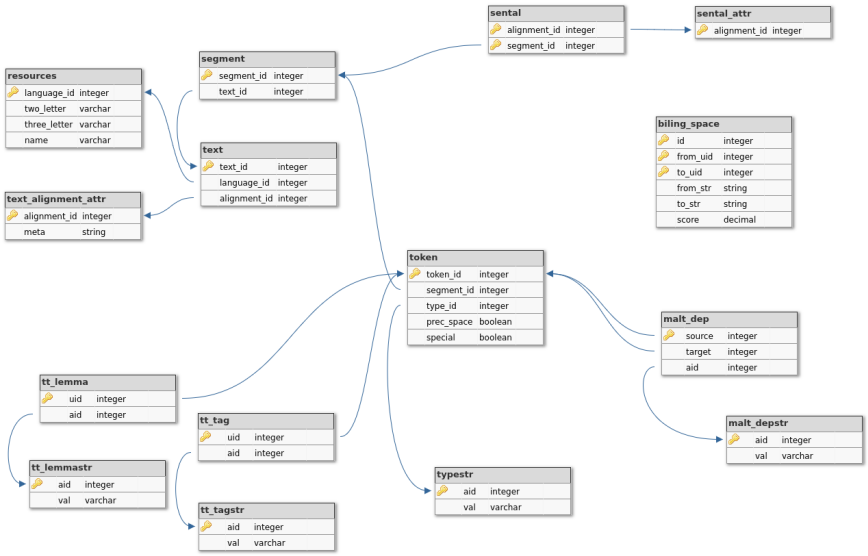


Figure 18: Illustration of database structure.

7.4.1 Representation of Texts in the Database Structure

Every single text in the corpus, be that sessions in the Europarl corpus, articles in the Text+Berg corpus, or law section in the Swiss Law Corpus receives a dedicated entry in the text table. The text table is linked to the resources table, which provides information about the language a text is written in, as well as to the text_alignment_attr table, which tells us to which article in the other language a text is aligned. This information is needed for the sentence alignment. Moreover, a foreign key (blue arrows in Figure 18) assigns each segment from the segment table to a specific text.

7.4.2 Token Representation in the Database

After tokenisation, each token receives an entry in the token table. Every token is assigned an ID with which it can be identified, while we also equip it with a type ID. The type ID points to the typestr table, which contains the type string. This measure prevents redundancy in that we do not include the type string every time the type appears in the token table. Moreover, we can link each token to a specific segment, which usually is a sentence. Other attributes in the token tables are prec_space and special. Prec_space provides information about spacing, for example, if the token is preceded by a space or not. This information makes string aggregation easier and retains the spacing as it occurs in the raw corpus data. We do not provide information for the special attribute.

7.4.3 Part-of-speech Tag Information on Database Level

Part-of-speech tag information is split up in two tables, the tt_tag and the tt_lemma table (where the tt stands for TreeTagger). Each token is associated with a lemma from the tt_lemma table and a part-of-speech tag from the tt_tag table. If a lemma is unknown, it does not receive an entry in the tt_lemma table, but usually, the TreeTagger provides a tag, which is why the token would nonetheless receive an entry in the tt_tag table. The string information for both lemmas and tags – for the same reason as for the tokens – is again separated in the tt_lemmastr and the tt_tagstr table, respectively.

7.4.4 Representation of Parses in the Database

We store the information from the MaltParser in two tables. In the malt_dep table we keep the information which source token is related to which target token. The malt_depstr provides the information about the type of relationship between two tokens. In general, this would make the querying for parsing structures possible. As of yet, this is neither implemented in Multilingwis, nor do we rely on parsing information during query time of our system.

7.4.5 Sentence Alignment in the Database

Since we perform the sentence alignment based on types and lemmas, we store this information in two separate tables. The segments itself are stored in the segment table. The sentence alignment tables hun_lemmas_sental and hun_types_sental assign an alignment ID to each alignment from hunalign and lists this ID twice, in a first row with the source segment, and in a second row with the target segment (hence the composite primary key over the alignment and segment ID). The segment_count row provides information about how many segments are aligned, which allows for only pairwise occurrences of the alignment IDs. The sentence alignment tables also point to an additional table in which we find the confidence score of the alignment provided by the sentence aligner.

7.4.6 Integrating Word Embedding Information

Word embedding models are stored as binary files, containing huge matrices. The querying can be made more efficient, although this is not the primary aim of this study. We try the following method: for our best performing model, we integrate the top 6 translation candidates for each word in the vocabulary of the model. All this information is stored in a separate table called biling_space. Together with an id, we store each translation with the source and the target id from the tt_lemmastr table, along with the source and target string and the similarity score from the word embedding model. A cleaner solution would have been to assign foreign keys to the lemma table or to create a materialised view. The former method, however, would have required changes in the properties of the database design of which we were unaware until we tried this method. Nevertheless, the method of singling out the data into a stand-alone table still is beneficial to the final system, since

we do not have to load the whole vector space models, which amount from 400MB to 2GB (depending on the configuration) for the data at hand, into memory for each query.

7.5 BIPHI – A Bilingual Phrase Identifier

In order to make the information we have calculated accessible to users, we have written an interface which allows to query the database. We get our inspiration from Bilingwis and Multilingwis, although the interface is much simpler and not as sophisticated as the aforementioned tools. Nevertheless, the possibility to query a parallel corpus in which words are "aligned" with word embeddings enables a user to get a good impression of what is possible, and what not.

We use *django*[13] for the development of the application. Django is a high-level Python Web framework. In the following, we describe the core functionalities and present our system as a whole. The evaluation of the techniques which underlie the corpus query tool is subject of the following chapter.

7.5.1 The Life Cycle of a Query and how it is Processed

The starting page of BIPHI contains nothing more than the text field in which a user can enter her query. Figure 19 gives an impression of the search interface. A user can choose whether she wants to search for German or French MWEs. In either case she must indicate the translation direction in order to get results, since we have not implemented automatic language detection like is the case for, e.g., Multilingwis. Upon hitting the 'Search'-button, the query gets processed and the user is linked to the result page.

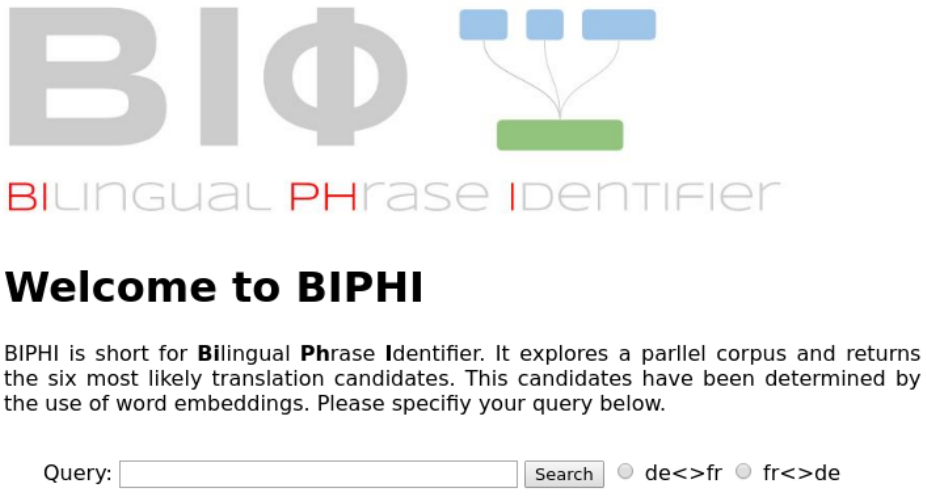


Figure 19: The query interface of BIPHI.

.

[13]see https://www.djangoproject.com/

Figure 20 shows the basic processing of a query entered into the system. Given a query, we first lemmatise it. Next, we identify all the segments in which the query term or, in the case of a multiword query, constituents of it occur. In this respect, we tried to mimic the query strategy implemented in Multilingwis and described in Clematide et al. [2016]. The method is simple: for each term in the query, we search the database index and subsequently join the results obtained for every following lemma. This procedure has the pleasant effect to constantly reduce the search space, since we only look for the next lemma within a token range of 4 tokens. At the same time, we search for translation candidates in the imitation of the bilingual vector space in the database. Because the model which underlies the system is trained on lemmas (see Chapter 8 for an explanation), we also use the lemmatised form of the query. This gives us a list of 6 translation candidates in total.

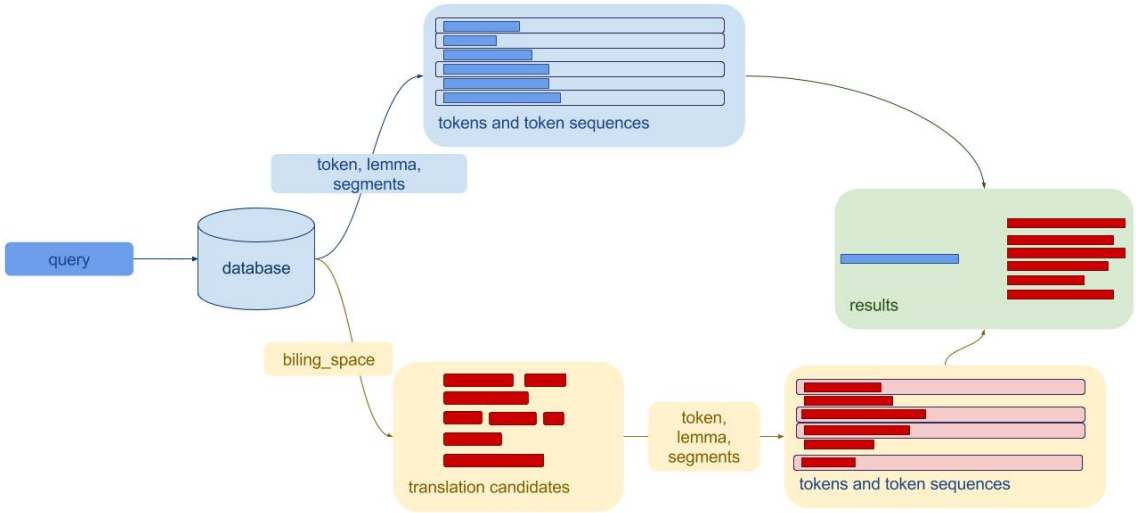


Figure 20: Processing steps of a query.

The result page gives an overview of how many times the query entered by the user occurs in the corpus. It also provides an example sentence in which the query term is used. In the overview of the results, the user finds the most probable translations identified by the bilingual word embedding model, along with two example sentences each, in which the terms should be highlighted.

We are aware that the system is very basic. But it nevertheless allows for more or less fast lookups of relevant translation candidates. Figure 21 and Figure 22 show the results of two different queries. We see that also single word queries return probable results. There are several methods to improve the system.

- The integration of language identification of the query term makes the system easier to use for end users.
- In order to be able to tell whether the bilingual word embedding model has “aligned” the words correctly, it would be useful to present the aligned sentence next to the one in which a translation candidate occurs.

Overview

Your query **tempête** occurs **1417** times in the corpus.

Example sentence: La **tempête** s' est déchaînée contre l' avion , le ballottant de -ci de -là au gré des trous d' air .

Query results

Translation candidate	Frequency	Example sentences
Unwetter	198	-(B4-0939/97) der Abgeordneten Miranda, Novo, Ribeiro und Gutiérrez Díaz im Namen der GUE/NGL-Fraktion zu den gravierenden Folgen des Unwetters in Portugal;
		Doch dann kam das **Unwetter**.
Schneesturm	188	2008 waren Überschwemmungen und Regenfälle in Australien , die zu Exportunterbrechungen führten , Produktionsausfälle in Südafrika und schwere Schneestürme in China , die ebenfalls Exportengpässe nach sich zogen .
		Jahr für Jahr trotzen diese Bergbewohner sowohl der Kälte und den Schneestürmen eines endlosen Winters als auch dem grellen Licht und den zerstörerischen UV-Strahlen eines kurzen, aber intensiven Sommers.
Orkan	100	Modernste Messmethoden haben gezeigt , dass im Innern eines meditierenden tibetischen Mönchs bezüglich seiner Hirnwellen ein eigentlicher **Orkan** tobt .

Figure 21: Overview and top section of the results for the French single word query *tempête* 'storm'.

- The computation of aligned sentences in the way we tried to implement it was not efficient enough. This is why we limit the number of example sentences to two. The solution of this problem would also be beneficial to finding alignments (see the second point).
- We sometimes observe that the highlighting of the query and translation terms does not work as expected. This is due to the fact that we only mark them if they occur in their lemmatised form. Moreover, we sometimes highlight words which do not directly belong to a MWE (for instance if the French word *gaz* 'gas' occurs elsewhere in a sentence which at the same time also contains the MWE *gaz à effet de serre* 'greenhouse gas'). Correct alignment facilitates the recognition of the context for users and makes it easier for them to decide whether a translation is relevant or not.
- The frequencies of the translation candidates reflect the total number of occurrences of the word or multiword unit in the total corpus. Much more interesting would be the fact how many times the query term was translated as the variant proposed by the system.

We see that there is much more room for improvement. Moreover, bilingual word embedding models are able to capture more than just the translation. If we want to exploit these features as well, we will have to rethink our implementation method, and look for other ways to make the word embedding models searchable.

Overview

Your query **Entscheidung treffen** occurs **1520** times in the corpus.

Example sentence: Aber das Parlament hat eine wichtige **Entscheidung** zu **treffen**, und zwar muß es seine Position festlegen, damit der Rat der Landwirtschaftsminister die dringend erforderlichen Änderungen vornehmen kann.

Query results

Translation candidate	Frequency	Example sentences
décider	15330	Naturellement, je sais très bien que l'Europe pourrait **décider** de se passer de tout système, quel qu'il soit. Mais la vie en a décidé autrement.
décision	39194	2Elle peut déléguer l' instruction de procédures et la préparation de décisions à un ou à plusieurs de ses membres . nombre de décisions sont prises conjointement.
prendre un décision	2965	Aussi voudrais-je souligner que dans ces conditions on ne peut, à mon avis, prendre aujourd'hui de décision sur le rapport Funk.

Figure 22: Overview and top section of the results for the German MWE *Entscheidung treffen* 'to make a decision'.

7.6 Summary

This chapter has provided an outline of the different preprocessing steps which we applied to the data. It becomes clear that especially the identification of MWEs based on patterns and association measures is problematic in the sense of accuracy. Still, there is much manual labour involved. One method to alleviate this problem is to include dictionaries of MWEs in the identification process. However, chances are that such dictionaries are not complete, which is why it is probable that MWEs which do not occur too often will be missed. We are aware of the fact that we miss these MWEs too. Rather then focusing on these instances, the aim of this thesis is to explore whether word embeddings are capable of capturing MWEs in the first place. The inclusion of rarer MWEs, or rather the identification of them, is subject to further research.

What is more, we have established the benefits of using relational databases to store large corpora. We have introduced the corpus query system BIPHI, which builds on top of such a relational database and is intended to find MWEs. We have given a detailed account of its core functionalities and searching strategies. The evaluation of the model which underlies the system follows in the following chapter.

8 Analysis and Interpretation

The results which we present in this chapter are three-fold. Firstly, we evaluate the word embedding models, which is necessary to determine which configuration captures MWEs best. Secondly, we will test the performance of the best word embedding model against word alignments, or, to be more precise, a phrase table which is based on word alignment information. Lastly, we will asses the quality of the translations of the best word embedding model compared to translations provided by Multilingwis. This evaluation process should give us an impression of what bilingual word embeddings are capable of and also how they perform compared to established techniques.

8.1 Evaluation of Word Embedding Models

Table 13 summarises the results of the evaluation of the bilingual word embedding models using different parameters (as described in Section 7.3.8 of Chapter 7). The column 'matches' describes the percentage of returns a vector space model managed for 100 German and 100 French MWEs. It seems that some MWEs, although being present in the corpus, are missing in certain word embedding models. This must be due to the fact that words which do not occur frequently enough in the training data are excluded from the training step. The column 'avg de → fr' gives the translation accuracy for 100 randomly chosen German MWEs , while the column labelled 'avg fr → de' shows the values for another 100 French MWEs translated to German (all of which occur in our candidates lists). The 'avg' column contains the average of the two preceding columns and constitutes what we will call the overall accuracy of a specific model. The similarity scores in the last two columns present the average of all the similarity scores returned, given the same 100 German and French MWEs as we have used to calculate the scores in columns 8 and 9.

For each of our input text type (i.e., lemmas, GerTwol+lemmas, raw input, and GerTwol+raw text), we mark the model which performs best in bold in Table 13 (if several scores rank highest, we additionally put them in italics). Our evaluation shows that model 4 has the highest accuracy score with 0.7 and an overall similarity score of 0.48 for translations from German to French and 0.62 in the opposite direction. It uses word lemmas as training input and is trained with the skip-gram algorithm with a window size of 8 context words and 600 dimensions. Interestingly, other models provide higher average similarity scores (e.g., models 1, 3, 5, 7, 9, 11, 25, 27, and 29). In contrast to the best performing model 4, however, their translations were not always as precise.

To give a few examples on how the models perform differently, consider the translation of the French MWE *apporter son soutien* 'to give support'. Model 4 translates this

MWE as *unterstützen* 'to support', while model 10 gives *uneingeschränkt* 'boundless' as most probable translation. The only, but crucial difference between the two models is that model 4 was trained with the skip-gram approach, whereas model 10, on the other hand, was trained with the continuous bag-of-words algorithm. Upon closer inspection, we find *unterstützen* as second-best translation in model 10, followed by *vorbehaltlos* 'unreserved' and *rücksichtslos* 'ruthless'. These are all words which likely occur together with *unterstützen*, *Unterstützung*, or *Unterstützung geben* and if we consider that the continuous bag-of-words model tries to find similar words based on their context, it is only logical that it favours words which occur around the target word. The skip-gram algorithm, which predicts the context based on a target word, seems to be able to pinpoint translations more accurately.

There are other peculiarities which are worth taking a look at. The German MWE *Früchte tragen* 'to bear fruit', for instance, is only translated correctly by model 42 as *fruits portent* (although the word order is not right). For the same MWE, model 40 gives the translation **fruits supporter*. We can prevent this translation by checking against the cosine similarity between *fruits* and *porter* and the one between *fruits* and *supporter* in the French model only. The latter returns a value of 0.12, while the former is 0.33. The higher similarity score is an indicator for the fact that *fruits* occurs more often together with *porter* than with *supporter*.

Another characteristic of word embedding models is that some are more suitable to reflect type similarity, while others provide more information about the topical similarity (see Section 6.2.1.1 of Chapter 6). A point in case is the translation of the French MWE *banque centrale*. Model 35 returns as top-most translation candidate *EZB*, which is an acronym for *Europäische Zentralbank* 'European Central Bank'. Model 30, on the other hand, puts *Zentralbank* 'central bank' as its first translation. The major difference between the two models is again that they were trained with different algorithms, model 35 with continuous bag-of-words and model 30 with skip-gram. We can thus assume that models trained with skip-gram give more weight to words which are topically related, while the nature of the similarity in models trained with continuous bag-of-words is rather typical (because the *EZB* is an instance of a central bank). In order to verify this fact, we would of course have to conduct further experiments.

If we compare the performance of the models based on the translation direction, it immediately strikes the eye that the translation from French to German is often more accurate than the translation from German to French. Upon applying a t-test in order to determine the significance of this difference, we find a p-value < 0.0001, which ascertains a highly significant difference. We therefore have to assume that translating from a non-compounding language to a compounding language with the help of word embeddings is more successful than the other way around. Put another way, it is easier to find n:1 alignments than 1:n alignments in the vector space model. This finding is in line with the capability of IBM word alignment models, which are able to generate many-to-one alignments, but not one-to-many (Brown et al. [1990]).

As a last comparison, we are interested in whether any of our preprocessing steps

actually benefits the translation performance. We present the results of four comparisons in Table 12. As we can observe, lemmatisation has a significant impact on the performance of the models. We reckon that this measure should always be applied in order to achieve better results. It is likely that lowercasing would also have a beneficial effect, since it reduces again the number of types and therefore the vocabulary size the word embedding models are trained on.

input	avg	std	t-value	p
lemmas	.63	.03	1.7961	0.0862
lemmas GerTwol	.61	.05		
raw text	.56	.03	2.0375	0.0538
raw text + Gertwol	.59	.03		
lemmas	.63	.03	5.8793	**>0.0001**
raw text	.56	.03		
lemmas GerTwol	.61	.05	1.3365	0.1951
raw text + GerTwol	.59	.03		
lemmas + lemmas GerTwol	.62	.04	4.4353	**>0.0001**
raw text + (raw text + GerTwol)	.57	.04		

Table 12: Significance test for different model groups. Statistically significant values are in bold.

If we take a closer look at Table 12, we notice that decompounding does not influence the quality of translations, not even if we compare lemmas against the us of raw text in the fourth row. This comes as a surprise, since we have expected that decompounding allows for a more precise alignment of constituents of MWEs. We have not evaluated the compound analysis with GerTwol. It is probable that the tool does not always perform as we expected and thus the potential benefits of decompounding is missing in the models. By looking at the translation of the French MWE *gaz à effet de serre* 'greenhouse gas', which translates to German *Treibhausgas*, in model 16, we see the components *Treib* and *gas* among the top 3 translation candidates. *haus* is probably missing because it generally does not occur in the environments of *Treib* and *gas*. We notice too that GerTwol decompounded the noun wrongly, since it should correctly be *Treibhaus* and *gas*. If we translate the single components into the other direction, we see that the correct translation occurs in second position at the latest. Hence, decompounding could still be beneficial if applied more carefully than we have done here.

We want to present the insights we gained from this experiment in a nutshell:

- The training parameters do have an influence on the translation accuracy. The biggest impact has the choice over the underlying algorithm. We have seen that skip-gram-based models provide translation variants more accurately than models that have been trained with the continuous bag-of-words algorithm. What we have not tried, but what would certainly be interesting, is to integrate **negative sampling** in the training steps. In short, negative sampling compiles bad examples for each word

No.	*tokens*	*algorithm*	*window size*	*dimensions*	*matches*	*avg de ↔ fr*	*avg fr ↔ de*	*avg*	*sim. scores de ↔ fr*	*sim.scores fr ↔ de*
1	lemmas	sg	4	300	.99	.55	.72	.63	.56	.68
2				600	.99	.56	.73	.65	.46	.59
3			8	300	.99	.60	.71	.65	.58	.71
4				600	.99	.60	.81	**.70**	.48	.62
5			12	300	.99	.57	.76	.66	.61	.73
6				600	.99	.59	.67	.63	.43	.57
7		cbow	4	300	.99	.55	.70	.62	.50	.64
8				600	.99	.54	.72	.63	.43	.56
9			8	300	.99	.50	.67	.59	.50	.65
10				600	.99	.58	.67	.62	.44	.57
11			12	300	.99	.55	.64	.59	.50	.64
12				600	.99	.59	.67	.63	.43	.57
13	lemmas GerTwol	sg	4	300	1.0	.57	.74	***.66***	-	-
14				600	1.0	.57	.70	.63	-	-
15			8	300	1.0	.59	.69	.64	-	-
16				600	1.0	.57	.64	.60	-	-
17			12	300	1.0	.60	.71	***.66***	-	-
18				600	1.0	.62	.67	.65	-	-
19		cbow	4	300	1.0	.55	.55	.55	-	-
20				600	1.0	.56	.70	.63	-	-
21			8	300	1.0	.54	.56	.55	-	-
22				600	1.0	.53	.69	.61	-	-
23			12	300	1.0	.53	.55	.54	-	-
24				600	1.0	.54	.56	.55	-	-
25	raw text	sg	4	300	.96	.51	.51	.51	.56	.63
26				600	.96	.45	.52	.48	.47	.56
27			8	300	.96	.59	.57	.58	.59	.64
28				600	.96	.55	.56	.55	.48	.56
29			12	300	.96	.54	.56	.55	.60	.65
30				600	.96	.61	.58	***.59***	.49	.56
31		cbow	4	300	.96	.49	.58	.54	.48	.56
32				600	.96	.54	.58	.56	.42	.50
33			8	300	.96	.55	.64	***.59***	.43	.50
34				600	.96	.55	.64	***.59***	.43	.50
35			12	300	.96	.53	.65	***.59***	.48	.55
36				600	.96	.53	.61	.57	.42	.49
37	raw text + GerTwol	sg	4	300	.97	.56	.51	.53	-	-
38				600	.97	.56	.64	.60	-	-
39			8	300	.97	.60	.59	.59	-	-
40				600	.97	.60	.57	.59	-	-
41			12	300	.97	.61	.66	**.63**	-	-
42				600	.97	.57	.58	.58	-	-
43		cbow	4	300	.97	.53	.67	.60	-	-
44				600	.97	.57	.68	.62	-	-
45			8	300	.97	.57	.59	.58	-	-
46				600	.97	.53	.65	.59	-	-
47			12	300	.97	.50	.54	.52	-	-
48				600	.97	.54	.65	.59	-	-

Table 13: Evaluation of the bilingual word embedding models using different parameters.

and tries to adjust the word's vector in such a way that it becomes more dissimilar to these words. This approach strengthens the relation to similar words and thus exploits the training data to the full.

- A big influence has the lemmatisation of the data. Other normalising steps, like lowercasing, the treatment of special characters, and many more could also have positive effects on the training of word embeddings.

- The decompounding of German nouns does not seem to have an influence on the quality of the word embedding models. This should, however, be verified by further experiments, since other normalising steps as described above have an impact on the word embedding models.

8.2 Word Embeddings vs. Phrase Table

Upon having identified the best performing model, we need to test it against traditional methods to find MWEs. One such method in statistical machine translation is using information from phrase tables. The assumption is that the phrase tables, which consist of translations of phrases, capture MWEs to some degree. In order to do the comparison, we train a phrase table with our lemmatised "super corpus". We use the lemmatised version since the best word embedding model is based on lemmas as well. This allows for a fair comparison. We generate the phrase table with the *Moses*[1] toolkit by Koehn et al. [2007]. We use the standard configuration. Next, we randomly extract 150 MWEs each from the German, as well as from the French MWE candidates list, and filter the phrase table for lines in which these MWEs occur. We only consider exact matches. For instance, during the extraction of lines containing the German MWE *Bemerkung machen* 'to make a remark', we only extract lines lines if the first field of the line contains this extract string. The same applies for French MWEs, except that we scan the second field of the phrase table for the exact MWE string.

During the extraction, we observed that many translation candidates in the phrase table receive unrealistic translation probabilities of 0.5 or even 1, although there are other candidates which are in fact better translations. Consider the example in (8.1). The first row shows a perfect translation of the German MWE, while the translation in the second example is unusable. Nevertheless, the first value in the third field, which is the probability of the French sequence given the German sequence, is 1, whereas the probability of the translation in the first row is much smaller (only 0.00952381). We notice too, however, that the unlikely sequence in the second row only occurs once (the first value in the fifth field) as a "translation" of the MWE. We thus exclude all such occurrences in order to receive more probable translations.

(8.1)
- `bemerkung machen ||| faire un remarque ||| 0.00952381 0.0173189 0.00265957 4.58811e-05 ||| 0-2 1-2 ||| 105 376 1 ||| |||`
- `bemerkung machen ||| observation à formuler en ce qui concerner ||| 1 0.000537771 0.00265957 4.82644e-15 ||| 0-0 1-1 1-2 1-3 1-4 1-5 0-6 ||| 1 376 1 ||| |||`

The evaluation is again based on human judgement and proceeds similar as the evaluation of the vector space model. One major difference is that we do not only rate the top-most translation candidate, but the top 6 translation candidates of the word embedding model and the phrase table. After having excluded 2 MWEs each from the German

[1] see `http://www.statmt.org/moses/`

and the French list since they did not really constitute MWEs from further analysis, we are left with 148 MWEs each. This amounts to a total amount of 888 translations we need to consider (6×148). For the translation direction from German to French, there are 35 cases in which neither the word embedding model nor the phrase table provide meaningful translations. This leaves us with 113 MWEs and thus 678 translations. In the other translation direction, we exclude 11 MWEs for the same reason, resulting in 822 translations. We then proceed as follows: for each MWE, we first look at each of the six translation candidates suggested by the word embedding model and decide whether it is a correct translation (it receives 1 point), a partially correct translation (it receives 0.5 points), or incorrect (it receives no points). Next, we carry out the same rating for the translations provided by the phrase table. In a last step, we compare the two ratings. The translation method which scores higher and has at least one correct translation (preventing the preference of one method if it produces only partially correct translations when the total score of the other method is lower, but has at least one correct translation) receives one point. If both scores are even, we assign half a point to each system.

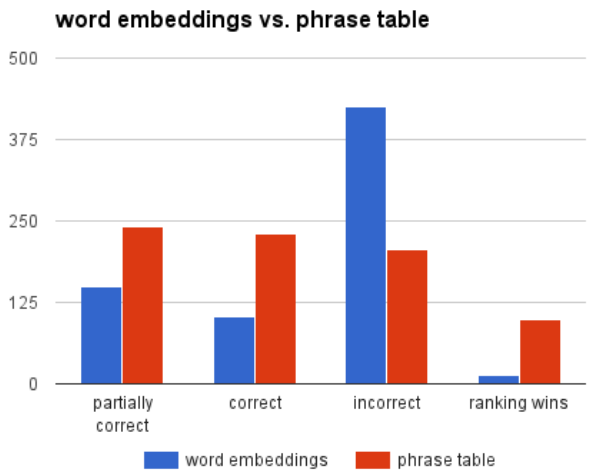


Figure 23: Performance of word embeddings vs. phrase table for the translation of MWEs from German to French.

Figure 23 summarises the evaluation of word embeddigns vs. phrase table in the translation direction from German to French. It is more than evident that the translations provided by the phrase table outperform the word embedding model. In percentages, the word embedding model only provides partially correct translations in 22.12% of all cases, it gets 15.04% of all translations correct, and only wins 11.4% of the rankings. The phrase table performs better in every respect, achieving 35.55% partially correct translations, 33.92% correct translations, and it wins 85.97% of the rankings (2.63% are even).

For the translation direction from French to German, Figure 24 presents the summary. We can see that not only word embeddings, but also the phrase table performs better as opposed to translating in the other direction, which is again to the inherent characteristics of word alignments per se. Still, the phrase table outperforms the word embedding

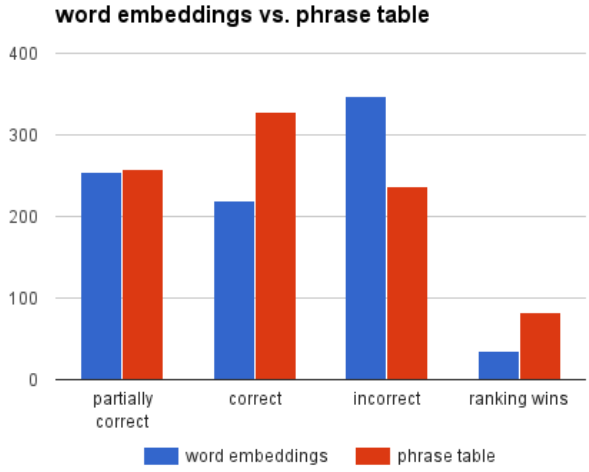


Figure 24: Performance of word embeddings vs. phrase table for the translation of MWEs from French to German.

model, although in terms of partially correct translations, both systems achieve comparable results. We also observe that for the word embedding model, the amount of correct translations sees an increase of almost 12%. It also wins a little more than a quarter of the rankings. The performance gain is much lower for the phrase table, for which the identification of correct translations increases by only about 6%. The percentage of won rankings substantially decreases to a value of 60.58%, meaning that in about 15%, word embeddings and phrase tables score evenly.

The above evaluation allows us to draw the following conclusion: Word embeddings cannot yet cope with techniques based on word alignment information. However, if word embedding models have enough information, they can outperform the translations of a phrase table sometimes. If we take the German MWE *unter Beweis stellen* 'to give prove of sth.', we see that the word embedding model provides not only the correct MWE in French, which is *faire preuve*, but also synonyms like *démontrer*, *prouver*, or *montrer*, while the phrase table comes up with translation suggestions like *donner le preuve*, or *apporter le preuve*. Thanks to the inherent properties of word embeddings, which is to find similar words, we assume that word embeddings are better suited to detect synonyms, or closely related words. This is indicated by the fact that for the French MWE *guerre mondial* 'world war', we find translation candidates like *1945* or *40er-Jahr*. We would not mark such translations as correct, but if we take into consideration that the Second World War was fought in the 1940s, we must acknowledge that word embedding models also capture other useful information. The phrase table is not able to make such abstractions, or at least, it does not see the correlation between *guerre mondial* and *1945*. As such, word embeddings cannot replace phrase tables, but they could prove as a valuable tool for gaining more, and especially more relevant, information from them.

8.3 Word Embeddings vs. Multilingwis

We could say that it is more logical to rely on phrase tables when we want to identify translations of MWEs, since the phrases in such a table are most likely to correspond to MWEs. However, phrase tables are not able to capture long distance dependencies. This is why word alignments could be more reliable. Multilingwis intelligently exploits word alignment information. A comparison of the system against the best word embedding model could provide valuable insights.

The set up for this evaluation is almost identical to the one described in Section 8.2. We also evaluate in both translation directions. Instead of randomly choosing the MWEs from the candidates list, we select 50 MWEs each for German-French and for French-German from the MWEs we have already evaluated against the phrase table. We are particularly interested in the MWEs which achieve high scores in that comparison, and thus we mainly include MWEs which have won a ranking or which provide a high number of (partially) correct translations. The selection process results in a total of 300 translation candidates for each language we need to evaluate. Again, we take the top 6 candidates from the word embedding model and from Multilingwis into consideration. The assignment of points follows the same schema as outlined in Section 8.2.

Figure 25 shows the result for the translation direction from German to French. As is the case with phrase tables, Multilingwis outmatches the word embedding model. Whereas the word embedding model only provides correct translations in almost 30% of all cases, Multilingwis is correct in over 50%, with only 17.33% of incorrect translations (compared to the 46% error rate of the word embedding model). As concerns the rankings, the word embeddings win over Multilingwis in 12% of all comparisons, while Multilingwis wins 76% of the rankings.

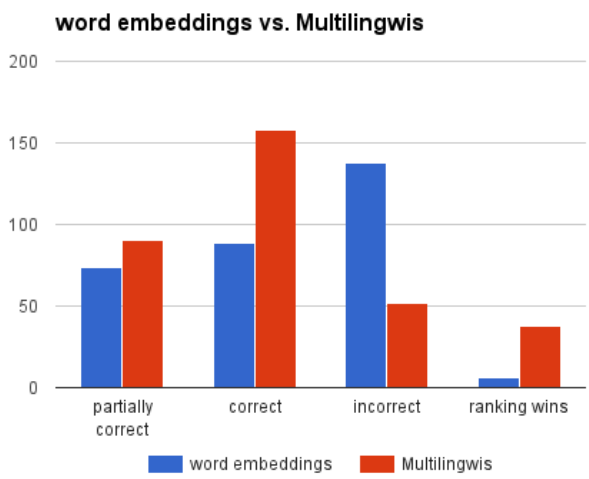


Figure 25: Performance of word embeddings vs. Multilingwis for the translation of MWEs from German to French.

In Figure 26, very much to our surprise, we observe that Multilingwis performs

worse in the identification of French MWEs than it does for German MWEs. In contrast, the percentage of correct translations the word embedding model provides increases from 29.67% to 34.67%. The race for the ranking also gets closer: the word embedding model wins 28% of them, leaving Multilingwis 68%, which is a loss of 8% as compared to the translation direction German-French.

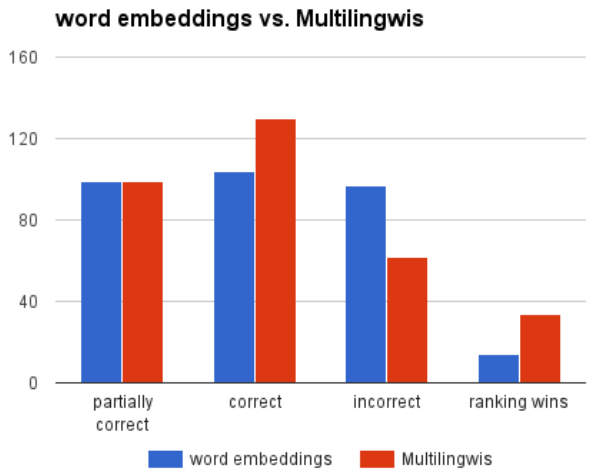


Figure 26: Performance of word embeddings vs. Multilingwis for the translation of MWEs from French to German.

Similarly to the comparison of word embedding models to phrase tables, bilingual word embeddings can only cope with word alignments in limited ways, at least as concerns the identification of MWEs. What we did not consider is the alignment of single words, for which the results we have obtained here could look very differently. Nevertheless, there are cases in which word embeddings can either provide additional information, or in which they outperform word alignment. For instance, if we consider translations for the French MWE *pays de développement* 'developing country', the word embedding model returns the correct translation *Entwicklungsland*, as well as *Industrieland*, *AKP-Land* (in which *AKP* stands for 'African, Caribbean and Pacific Group of States'), *AKP-Staat*, *arm* 'poor', and *Afrika* 'Africa'. All these words are clearly associated with each other. Multilingwis is not capable of producing this output. It would only make the connection of *pays de développement* and *AKP-Land* if any of the words in the French MWE is alignend to *AKP-Land*, which is only once the case.

As such, word embeddings prove once more that they can reflect relationships among words and that at times, they are even able to produce correct translation. However, they do not yet reach the performance of word alignments. Bear in mind that the primary aim of word embeddings is not to align words, but to find relationships between them. This fact still holds when we project another language into the same space, as we have shown with a few examples above.

8.4 Limitations and Shortcomings

There are a few limitations and shortcomings originating from the corpus design, the preprocessing, as well as the evaluation method. Some of them are minor issues, while others clearly have a big influence on the performance for bilingual word embeddings. This is why we highlight the most important shortcomings in this section. At the same time, we would like to suggest solution strategies in order to eliminate eminent weaknesses.

8.4.1 Data Preparation

Since the data originates from various sources, they differ greatly in their cleanliness. The Europarl corpus used in this thesis, for example, has been exposed to a lot more cleaning steps than the Credit Suisse Corpus. This is unsatisfactory, because ideally, we would like to work with data as consistent as possible.

Another reason for the weak performance of word embeddings could also be the amount of data we have used to train the models. We assume that more data would have been beneficial to the identification of MWEs. Other researchers use large parts of Wikipedia to train their monolingual models. We are not sure whether a model trained on a comparable corpus made up of Wikipedia articles which is much larger than the corpus we use in this thesis would actually perform better.

As concerns data preparation, we would need to assess the quality of the decompounding step in future. This prevents wrong "alignments" in the word embedding models. Moreover, we should try whether lowercasing the data has any outcome on the experiments.

8.4.2 The Searchability of Word Embedding Models

We still lack a practical solution to make bilingual word embedding models searchable. An approach as we have introduced in Section 7.4 of Chapter 7 is not satisfactory, since we cannot perform comparisons based on cosine similarity and other measures any longer once we store the data in the database as we have done it. On the other hand, loading the whole model into memory every time we issue a query cannot be a practical solution either. Hence, we must find other solutions to solve this problem. The one we implemented in this thesis just so serves the need to find relevant translations. If we want to get more than 6 translation candidates, however, we will not be able to produce them, since we only store the six top-most translations.

What is more is the fact that we have not linked the keys in the biling_space table to any other table in the database, meaning that it is poorly integrated into the rest of the database design. This can have implications on the performance of queries in the database. A more solid approach would at least make references to other tables in the database. Nevertheless, this integration serves the need of a most basic corpus query tool.

8.4.3 Training and Evaluation of the Word Embedding Models

We have not taken all possible training parameters into consideration during the training of our bilingual word embedding models. The integration of negative sampling, as well as other values for the dimensionality reduction could have proven useful. However, the inclusion of additional parameters and the training with a greater variety of such parameters result in many more values to evaluate. Since our evaluation is mostly based on human judgement, it is not possible to evaluate more models with just one annotator.

This directly leads us to the presumably biggest weakness, which is the evaluation procedure. This is for two reasons: firstly, human judgement is not perfect. Secondly, human evaluation is a time consuming endeavour. Moreover, in order to produce more reliable results, we should have included judgements from other annotators. This problem calls for the need of more elaborate and maybe also automatic evaluation methods which are dedicated to measure the performance of bilingual word embedding models. There are already approaches which compare the performance of monolingual models based on similarity tasks. When it comes to the judgement of the correctness of translations provided by bilingual models, there do not exist standard measures yet. Nevertheless, there is promising work on its way to remedy this deficiency. Apart from manually compiling a gold standard and testing against it as in Vulić and Moens [2015], there is also the possibility to extract data from a multilingual WordNet as proposed by Bond and Foster [2013] and test against this. This evaluation method is applied by [Upadhyay et al., 2016] who perform bilingual lexicon induction based on word embeddings. Still, their work includes some manual evaluation too. This explicitly stresses the need for new and elaborate measures which aim at determining the correctness of bilingual word embeddings.

9 Conclusion

In conclusion, this thesis provides an extensive account of MWEs and their peculiarities, especially on how they are translated from German to French and vice versa. We emphasise that we cannot look at MWEs in isolation, but that many linguistic fields influence their existence. Each of these fields has a different way of treating MWEs. By highlighting these differences, it is possible to provide a more concise definition of what constitutes a MWE. Hence, the detailed account of this linguistic phenomenon allows us to exactly position the work of this thesis in the realm of computational phraseology.

As phraseology is essentially a subfield of lexicology and lexicography, we are also concerned with the translation of MWEs. In order to make accurate predictions on how we must translate a MWE from one language to another, we need to be aware of the different types of MWEs. Only then we see that verb-particle constructions do not occur in French and that the noun compounding in French behaves very differently from the method the German language uses.

When it comes to the automatic processing of MWEs in NLP, these insights allow for a careful design of patterns to extract MWEs from corpora. Upon integration of information of MWE in the training data we use for bilingual word embeddings, we are able to successfully find translations of MWEs. Our evaluation shows that bilingual word embedding models are not capable yet to compete with information gained from word alignments or phrase tables.

However, the admittedly expected defeat of word embedding models over word alignments and phrase tables should give no reason to abandon this technique altogether. The few good translations the word embedding models manage should rather be the motivation to investigate more effort in this research branch. Although we will probably never see the day when bilingual word embedding models replace word alignments, word embeddings can help the word alignment.

As such, it would be wise to unite these techniques. In this fashion, it would be possible to generate models which not only provide translations, but at the same time also synonyms and related terms. Such work would greatly benefit lexicographers and lexicologists, and it would most certainly aid in statistical machine translation.

All in all, the problem of MWEs and how to translate them remains once more unresolved. But research on this topic definitely goes in the right direction. What we think this thesis reveals best is that we cannot rely on one technique alone in order to get to grips with MWEs. Every method has its advantages and disadvantages. Even if it might seem that bilingual word embeddings have more disadvantages than advantages, we assume that together with other techniques, they could help to better cope with MWEs.

9.1 Future work

In the section about limitations and shortcomings in Chapter 8 we already hint at possible future work. On aim must be the development of measures to assess the quality of bilingual word alignments. Human judgement is too labour-expensive and impractical if we need to evaluate a great number of models. We should try to find ways to combine different techniques in order to find translations of MWEs. For instance, the integration of word embedding information in Multilingwis could help identify relevant translations. At the time of writing, Multilingwis ranks translation candidates according to its frequency. It can thus happen that a relevant translation, if it is a rare one, is not among the top 10 candidates. Using word embedding information could influence the ranking in such a way that the rarer but nonetheless relevant translation gets a higher rank. This would prevent users from scrolling through long translation lists. Moreover, word embeddings could provide context information which makes the disambiguation of translation candidates possible. This calls for a hybrid system in which we focus on the strengths of several techniques in order to get the most out of them.

Glossary

This glossary provides a short description of the key words and phrases that are used in this thesis.

association measure Statistical method to determine whether two or more words occur together by chance or if they constitute a collocation.

collocation Words which occur together more frequently than is predicted by chance.

corpus A collection of machine-readable texts, usually segmented up to word level and annotated with part-of-speech and lemma information.

lemma Base-form of a word, usually found in lexical entries.

lexicography The study of collections of words.

morpheme Most basic unit with which the formation of words is possible.

morphology The study of words and how they are formed.

multiword expression A lexicalised chunk consisting of two or more words. Its meaning can be either compositional or non-compositional.

part-of-speech tagger A tool to automatically assign words their part-of-speech.

parsing In computational linguistics, parsing is understood as the automatic annotation of sentences with their syntactic structure.

phoneme Distinctive (and atomic) units in the analysis of spoken language which distinguish one word from another.

phraseology The study of phrasemes and phraseological units.

semantics The study of word meaning.

syntax The study of how lexical items are formed into sentences.

vector A representation of numeric values with an origin, end point, length, and direction.

word embeddings Vector representation of words in a low-dimensional vector space.

References

S. Abdul-Rauf, M. Fishel, P. Lambert, S. Noubours, and R. Sennrich. Evaluation of sentence alignment systems. In *Technical report, Fifth MT Marathon*, 2010.

A. Abeillé, L. Clément, and F. Toussenel. Building a treebank for French. In *Treebanks*, pages 165–187. Springer, 2003.

V. Adams. *Complex Words in English*. Routledge, 2014.

N. Aepli and M. Volk. Reconstructing complete lemmas for incomplete German compounds. In *Language Processing and Knowledge in the Web*, pages 1–13. Springer, 2013.

K. Aijmer and B. Altenberg. *English Corpus Linguistics*. Routledge, 2014.

B. T. S. Atkins. Bilingual dictionaries: Past, present and future. *EURALEX'96 Proceedings. Göteborg: Department of Swedish, Göteborg University*, pages 515–546, 1996.

H. Baayen, W. Levelt, and A. Haveman. Research reported in annual report of the Max-Planck Institut für Psycholinguistik, 6. *Nijmegen: Max-Planck Institut*, 1993.

T. Baldwin and S. N. Kim. Multiword expressions. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

A. Baron, P. Rayson, and D. Archer. Word frequency and key word statistics in corpus linguistics. *Anglistik*, 20(1):41–67, 2009.

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(2):1137–1155, 2003.

A. Berard, C. Servan, O. Pietquin, and L. Besacier. Multivec: a multilingual and multilevel representation learning toolkit for nlp. In N. C. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 601–608, 2001.

F. Bond and R. Foster. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362, 2013.

G. Booij. Construction morphology. *Language and linguistics compass*, 4(7):543–555, 2010.

P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

R. D. Brown. Corpus-driven splitting of compound words. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, 2002.

H. Burger, D. Dobrovol'skij, P. Kühn, and N. R. Norrick. Phraseology: Subject area, terminology and research topics. *Phraseology. An International Handbook of Contemporary Research*, 1, 2007.

M. Carpuat and M. Diab. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245. Association for Computational Linguistics, 2010.

A. Chen. Cross-language retrieval experiments at CLEF 2002. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 28–48. Springer, 2002.

N. Chomsky. Lectures on government and binding. *Dordrecht: Foris*, 1981.

N. Chomsky. *The Minimalist Program*, volume 28. Cambridge University Press, 1995.

O. Christ. A modular and flexible architecture for an integrated corpus query system. *CoRR*, 1994. URL `http://arxiv.org/abs/cmp-lg/9408005`.

K. Church, W. Gale, P. Hanks, and D. Kindle. Using statistics in lexical analysis. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, page 115, 1991.

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

S. Clematide, J. Graën, and M. Volk. Multilingwis – a multilingual search tool for multi-word units in multiparallel corpora. In G. C. Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 513–522, 2016.

S. Cordeiro, C. Ramisch, and A. Villavicencio. mwetoolkit+sem: Integrating word embeddings in the mwetoolkit for semantic mwe processing. In N. C. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk,

and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

I. Dagan and K. Church. Termight: Identifying and translating technical terminology. In *Proceedings of the Tourth Conference on Applied Natural Language Processing*, pages 34–40. Association for Computational Linguistics, 1994.

I. Dagan, K. W. Church, and W. A. Gale. Robust bilingual word alignment for machine aided translation. In *In Proceedings of the Workshop on Very Large Corpora*, 1993.

B. Daille, É. Gaussier, and J.-M. Langé. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 515–521. Association for Computational Linguistics, 1994.

H. M. de Caseli, C. Ramisch, M. d. G. V. Nunes, and A. Villavicencio. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2): 59–77, 2010.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391, 1990.

S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–285. ACM, 1988.

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

W. Duverger. *A Comparison Between the Idioms, Genius, and Phraseology of the French and the English Languages*. Juigné, 1800.

S. Ebling, R. Sennrich, and D. Klaper. Digging for names in the mountains: Combined person name recognition and reference resolution for German alpine texts. In *Language and Technology Conference*, pages 189–200. Springer, 2011.

S. Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2005.

S. Evert and A. Hardie. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham, 2011.

C. Fellbaum. *WordNet*. Wiley Online Library, 1998.

J. Firth. A synopsis of linguistic theory 1930–55. *Studies in Linguistic Analysis*, pages 1–32, 1957.

J. Graën, D. Batinic, and M. Volk. Cleaning the Europarl corpus for linguistic applications. In *KONVENS*, pages 222–227, 2014.

J. Graën, S. Clematide, and M. Volk. Efficient exploration of translation variants in large multiparallel corpora using a relational database. In *4 th Workshop on Challenges in the Management of Large Corpora Workshop Programme*, pages 20–23, 2016.

S. T. Gries. Phraseology and linguistic theory: A brief survey. *Phraseology: An interdisciplinary perspective*, pages 3–25, 2008.

M. Haapalainen and A. Majorin. Gertwol: ein System zur automatischen Wortformerkennung deutscher Wörter. *Lingsoft, Inc*, 1994.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

U. Heid. Computational phraseology. an overview. *Phraseology: An Interdisciplinary Perspective*, pages 337–360, 2008.

S. Höfler and K. Sugisaki. Constructing and exploiting an automatically annotated resource of legislative texts. In *LREC*, pages 175–180, 2014.

R. Jackendoff. *The Architecture of the Language Faculty*. Number 28. MIT Press, 1997.

L. Karttunen, J.-P. Chanod, G. Grefenstette, and A. Schille. Regular expressions for language engineering. *Natural Language Engineering*, 2(4):305–328, 1996.

A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlỳ, and V. Suchomel. The sketch engine: ten years on. *Lexicography*, 1(1):7–36, 2014.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

R. B. Lees and N. Chomsky. Syntactic structures. *Language*, 33(3):375–408, 1957.

E. Lefever, L. Macken, and V. Hoste. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 496–504. Association for Computational Linguistics, 2009.

S. Löbner. *Understanding Semantics*. Routledge, 2013.

A. Lüdeling. Particle verbs in nlp lexicons. In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000: Stuttgart, Germany, August 8th-12th, 2000*, pages 625–630, 2000.

C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language processing*, volume 999. MIT Press, 1999.

I. Mel'ćuk. Collocations and lexical functions. *Phraselogy. Theory, Analysis, and Applications*, pages 23–54, 1998.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

G. Müller. Beschränkungen für Binomialbildung im Deutschen. *Zeitschrift für Sprachwissenschaft*, 16:5–51, 1997.

S. Nießen and H. Ney. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204, 2004.

J. Nivre, J. Hall, and J. Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219, 2006.

C. Ramisch, A. Villavicencio, and C. Boitet. mwetoolkit: a framework for multiword expression identification. In *LREC*, volume 10, pages 662–669, 2010.

X. Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer, 2002.

T. Samardžić. *Light Verbs and the Lexical Category Bias of Their Complements*. PhD thesis, Department of Linguistics, University of Geneva, 2008.

R. Schäfer. Processing and querying large web corpora with the COW14 architecture. *Challenges in the Management of Large Corpora (CMLC-3)*, pages 28–35, 2015.

A. Schiller, S. Teufel, and C. Thielen. Guidelines für das Tagging deutscher Textcorpora mit stts. *Manuscript, Universities of Stuttgart and Tübingen*, 66, 1995.

H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994.

R. Sennrich and M. Volk. Iterative, MT-based sentence alignment of parallel texts. In *18th Nordic Conference of Computational Linguistics, NODALIDA*, 2011.

F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1): 143–177, 1993.

F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38, 1996.

A. Spencer. Derviation. In P. Müller, editor, *Word-Formation: An International Handbook of the Languages of Europe*, Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK). De Gruyter, 2015. ISBN 9783110375732.

S. Stein. Phraseme im Diskurs. *Phraseology. An International Handbook of Contemporary Research*, 1:220–236, 2007.

P. Ströbel. The processing of formulaic sequences by nonnative speakers of English. 2012.

H. Telljohann, E. Hinrichs, S. Kübler, and R. Kübler. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004*, 2004.

S. Upadhyay, M. Faruqui, C. Dyer, and D. Roth. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*, 2016.

D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh, and V. Trón. Parallel corpora for medium density languages. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 292:247, 2007.

F. Villoing. French compounds. volume 24, pages 29–60. 2012.

M. Volk and S. Clematide. Detecting code-switching in a multilingual alpine heritage corpus. *EMNLP 2014*, page 24, 2014.

M. Volk, N. Bubenhofer, A. Althaus, and M. Bangerter. Classifying named entities in an alpine heritage corpus. *Künstliche Intelligenz*, 4:40–43, 2009.

M. Volk, N. Bubenhofer, A. Althaus, M. Bangerter, L. Furrer, and B. Ruef. Challenges in building a multilingual alpine heritage corpus. In *LREC*, 2010a.

M. Volk, A. Göhring, and T. Marek. Combining parallel treebanks and geo-tagging. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 192–196. Association for Computational Linguistics, 2010b.

M. Volk, T. Marek, and R. Sennrich. Reducing ocr errors by combining two ocr systems. In *ECAI-2010 workshop on language technology for cultural heritage, social sciences, and humanities*, pages 61–65, 2010c.

M. Volk, L. Furrer, and R. Sennrich. Strategies for teducing and correcting ocr errors. In *Language Technology for Cultural Heritage*, pages 3–22. Springer, 2011a.

M. Volk, A. Göhring, S. Lehner, A. Rios, R. Sennrich, and H. Uibo. Word-aligned parallel text: A new resource for contrastive language studies. In *Supporting Digital Humanities, Conference*, pages 52–206, 2011b.

M. Volk, J. Graën, and E. Callegaro. Innovations in parallel corpus search tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3172–3178, 2014.

M. Volk, C. Amrhein, N. Aepli, M. Müller, and P. Ströbel. Building a parallel corpus on the world's oldest banking magazine. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 288–296, 2016a.

M. Volk, S. Clematide, J. Graën, and P. Ströbel. Bi-particle adverbs, pos-tagging and the recognition of german separable prefix verbs. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 297–305, 2016b.

I. Vulić and M.-F. Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. ACL, 2015.

Curriculum Vitae

Personal Details

Phillip Ströbel
Rotfarbweg 8
9213 Hauptwil
phillip.stroebel@uzh.ch

Education

2006-2010	Grammar School at Kantonsschule Romanshorn
2010-2014	B.A. studies in English Literature and Linguistics, General Linguistics and Computational Linguistics at the University of Zurich
since 2014	M.A. studies in Multilingual Text Analysis at the University of Zurich

Professional Activities

2014	Student Assistant in the SwissTXT project at the Institute of Computational Linguistics at the University of Zurich
2015	Student Assistant in the Archimob project at the Corpus Lab of the University of Zurich
since 2015	Student Assistant in the Text+Berg project at the Insitute of Computational Linguistics at the University of Zurich
2016	Tutor in lecture “Einführung in die Multilinguale Textanalyse”

A Miscellaneous

A.1 German patterns for mwetoolkit

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE dict SYSTEM "dtd/mwetoolkit-patterns.dtd">
<patterns>
	<!-- find bi-particle adverbs -->
	<pat>
		<either>
			<pat>
				<w id="10" pos="ADV"/>
			</pat>
			<pat>
				<w id="11" pos="ADJD"/>
			</pat>
			<pat>
				<w id="12" pos="PAV"/>
			</pat>
			<pat>
				<w id="13" pos="PWAV"/>
			</pat>
			<pat>
				<w id="14" pos="APPR"/>
			</pat>
			<pat>
				<w id="15" pos="PTKVZ"/>
			</pat>
		</either>
		<either>
			<pat>
				<w id="16" pos="KON"/>
			</pat>
			<pat>
				<w id="17" pos="KOKOM"/>
			</pat>
		</either>
```

```
        <either>
                <pat>
                        <w id="18" pos="back:10.pos"/>
                </pat>
                <pat>
                        <w id="19" pos="back:11.pos"/>
                </pat>
                <pat>
                        <w id="20" pos="back:12.pos"/>
                </pat>
                <pat>
                        <w id="21" pos="back:13.pos"/>
                </pat>
                <pat>
                        <w id="22" pos="back:14.pos"/>
                </pat>
                <pat>
                        <w id="23" pos="back:15.pos"/>
                </pat>
        </either>
</pat>
<!-- find a LVC with object only -->
<!-- *** order N V *** -->
<pat>
        <w id="20" pos="NN" syndep="OBJA:21"/>
        <w id="21" pos="VV*"/>
</pat>
<!-- *** order V N *** -->
<pat>
        <w id="30" pos="VV*"/>
        <pat ignore="true" repeat="*">
                <w id="31"/>
        </pat>
        <w id="32" pos="NN" syndep="OBJA:30"/>
</pat>
<!-- find LVC with object only and a SVP -->
<pat>
        <w id="40" pos="VV*"/>
        <pat ignore="true" repeat="*">
                <w id="41"/>
        </pat>
        <w id="42" pos="NN" syndep="OBJA:40"/>
```

```
            <w id="43" pos="PTKVZ" syndep="AVZ:40"/>
        </pat>
        <!-- find LVC with prepositional phrase -->
        <!-- *** order V P N *** -->
        <pat>
            <w id="50" pos="VV*"><neg lemma="haben"/></w>
            <pat ignore="true" repeat="*">
                <w id="51"/>
            </pat>
            <w id="52" pos="APPR" syndep="OBJP:50"/>
            <w id="53" pos="NN" syndep="PN:52"/>
        </pat>
        <!-- *** order P N V *** -->
        <pat>
            <w id="60" pos="APPR" syndep="OBJP:62"/>
            <w id="61" pos="NN" syndep="PN:60"/>
            <w id="62" pos="VV*"/>
        </pat>
        <!-- find all VPCs -->
        <pat>
            <w id="70" pos="VV*"/>
            <pat ignore="true" repeat="*">
                <w id="71"/>
            </pat>
            <either>
                <pat>
                    <w id="72" pos="PTKVZ"
                    syndep="AVZ:70"/>
                </pat>
                <pat>
                    <w id="73" pos="ADJD"
                    syndep="PRED:70"/>
                </pat>
            </either>
        </pat>
</patterns>
```

A.2 French patterns for mwetoolkit

```
<?xml version="1.0" ?>
<!DOCTYPE dict SYSTEM "dtd/mwetoolkit-patterns.dtd">
<patterns>
        <!-- match ADV/A_qual C_C ADV/A_qual -> bi-particle adverbs -->
```

```
<pat>
        <either>
                <pat>
                        <w id="1" pos="ADV"/>
                </pat>
                <pat>
                        <w id="2" pos="A_qual"/>
                </pat>
        </either>
        <w id="3" pos="C_C"/>
        <either>
                <pat>
                        <w id="4" pos="back:1.pos"/>
                </pat>
                <pat>
                        <w id="5" pos="back:2.pos"/>
                </pat>
        </either>
</pat>
<!-- match noun compounds -->
<pat>
        <w id="10" pos="N_C"/>
        <pat ignore="true" repeat="{1}">
                <w id="11" pos="A_qual"/>
        </pat>
        <pat repeat="{1,}">
                <w id="12" pos="P"/>
                <w id="13" pos="N_C"/>
        </pat>
</pat>
<pat repeat="{2,}">
        <w id="20" pos="N\*"><neg lemma="art."/><neg
        lemma="article"/><neg pos="N_card"/><neg
        lemma="du"/></w> <!-- ignore the words "article",
        "du" and numbers -->
</pat>
<!-- match N A compound -->
<pat>
        <w id="30" pos="N_C"/>
        <w id="31" pos="A_qual"/>
</pat>
<!-- match LVCs -->
```

```
<!-- *** order V N *** -->
<pat>
        <w id="40" pos="V"/>
        <pat ignore="true" repeat="*">
                <w id="41"/>
        </pat>
        <either>
                <pat>
                        <w id="42" pos="N_C" syndep="mod:40"/>
                </pat>
                <pat>
                        <w id="43" pos="N_C" syndep="obj:40"/>
                </pat>
        </either>
</pat>
<!-- *** order N V *** -->
<pat>
        <either>
                <pat>
                        <w id="50" pos="N_C" syndep="mod:53"/>
                </pat>
                <pat>
                        <w id="51" pos="N_C" syndep="obj:53"/>
                </pat>
        </either>
        <pat ignore="true" repeat="*">
                <w id="52"/>
        </pat>
        <w id="53" pos="V"/>
</pat>
<!-- *** order V D N *** -->
<pat>
        <w id="60" pos="V"/>
        <pat ignore="true" repeat="*">
                <w id="61"/>
        </pat>
        <either>
                <pat>
                        <w id="62" pos="D*"/>
                        <w id="63" pos="N_C" syndep="mod:60"/>
                </pat>
                <pat>
```

```
                        <w id="64" pos="D*"/>
                        <w id="65" pos="N_C" syndep="obj:60"/>
                    </pat>
                </either>
            </pat>
            <!-- *** order D N V *** -->
            <pat>
                <either>
                    <pat>
                        <w id="70" pos="D*"/>
                        <w id="71" pos="N_C" syndep="mod:75"/>
                    </pat>
                    <pat>
                        <w id="72" pos="D*"/>
                        <w id="73" pos="N_C" syndep="obj:75"/>
                    </pat>
                </either>
                <pat ignore="true" repeat="*">
                    <w id="74"/>
                </pat>
                <w id="75" pos="V"/>
            </pat>
</patterns>
```

B List of tools and corpora

This appendix contains a list of tools and corpora we used in this thesis.

- Corpora
  - Credit Suisse News Corpus (`http://kitt.cl.uzh.ch/kitt/b4c/en/corpora.php`)
  - Credit Suisse PDF Bulletin Corpus (`http://kitt.cl.uzh.ch/kitt/b4c/en/corpora.php`)
  - Europarl (`http://www.statmt.org/europarl/`)
  - Swiss Law Corpus
  - Text+Berg Corpus (`http://textberg.ch/site/de/willkommen/`)
- Tools
  - cutter (`https://pypi.python.org/pypi/applepycutter/1.0`)
  - django (`https://www.djangoproject.com/`)
  - GerTwol (`http://www2.lingsoft.fi/doc/gertwol/`
  - hunalign (`http://mokk.bme.hu/en/resources/hunalign/`)
  - MaltParser (`http://www.maltparser.org/`)
  - multivec (`https://github.com/eske/multivec`)
  - mwetoolkit (`http://mwetoolkit.sourceforge.net/PHITE.php`)
  - TreeTagger (`http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`)

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Philosophische Fakultät
Studiendekanat
Rämistrasse 69
CH-8001 Zürich
www.phil.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass die Masterarbeit von mir selbst ohne unerlaubte Beihilfe verfasst worden ist und ich die Grundsätze wissenschaftlicher Redlichkeit einhalte (vgl. dazu: http://www.uzh.ch/de/studies/teaching/plagiate.html).

Hauptwil, 05.01.2017

.....

Ort und Datum Unterschrift