Flexible Language Independent Multiword Expression Analysis

A thesis presented
by

Bahar Salehi

to
The Department of Computing and Information Systems
in total fulfillment of the requirements
for the degree of
Doctor of Philosophy

The University of Melbourne
Melbourne, Australia
November 2016
Declaration

This is to certify that:

(i) the thesis comprises only my original work towards the PhD except where indicated in the Preface;

(ii) due acknowledgement has been made in the text to all other material used;

(iii) the thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed: ___________________________ Date: ________________
Flexible Language Independent Multiword Expression Analysis

Abstract

A multiword expression (MWE) is a combination of single words which together act as a single unit. Multiword expressions pose several challenges not only because they are estimated to be as frequent as simple words, but also because of their unpredictable lexical, syntactic, semantic and pragmatic behaviour in language. Given that MWEs can be found in every language and are very productive, general approaches are required to model them and their behaviour, without making language or type-specific assumptions.

This thesis addresses three research goals: (1) to model the semantic compositionality degree of MWEs in way that is applicable to a wide range of languages and types of MWEs; (2) to integrate the compositionality degree of MWEs into machine translation evaluation, in order to examine the importance of measuring compositionality in a natural language processing application; and (3) to determine the MWE inventory of a language, for which we do not have any linguistic background or annotated data.

In order to model the compositionality of MWEs, we propose four approaches using (1) translation-based string similarity, (2) translation-based distributional similarity, (3) word-embeddings and (4) Wiktionary provided definitions and translations. In these approaches, we avoid making language-specific or type-specific assumptions to examine their generality. For each approach, we compare with state-of-the-art methods and perform parameter sensitivity analysis to determine its strengths and weaknesses.

Then, we discuss integrating MWE compositionality into machine translation evaluation, focusing on English noun compounds. The underlying hypothesis here is that we should be less forgiving of partial matches with the reference translation for MWEs which are less compositional.

Finally, we propose a cross-lingual method to use resource-rich languages to determine the common MWE patterns of a resource-poor language. Most previous studies on MWE identification constrain the method based on prior knowledge of the common MWEs in a given language, which we are able to determine automatically with this method.
# Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Title Page</td>
<td>i</td>
</tr>
<tr>
<td></td>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>Table of Contents</td>
<td>iv</td>
</tr>
<tr>
<td></td>
<td>List of Figures</td>
<td>vi</td>
</tr>
<tr>
<td></td>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td></td>
<td>Citations to Previously Published Work</td>
<td>xi</td>
</tr>
<tr>
<td></td>
<td>Acknowledgments</td>
<td>xii</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Background and Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Aim and Scope</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Contributions</td>
<td>8</td>
</tr>
<tr>
<td>1.4</td>
<td>Thesis Outline</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Literature Review</td>
<td>13</td>
</tr>
<tr>
<td>2.1</td>
<td>What is a multiword expression?</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Types of MWEs</td>
<td>21</td>
</tr>
<tr>
<td>2.3</td>
<td>Computational studies of multiword expressions</td>
<td>28</td>
</tr>
<tr>
<td>2.4</td>
<td>Compositionality of MWEs</td>
<td>32</td>
</tr>
<tr>
<td>2.5</td>
<td>Applications</td>
<td>54</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary</td>
<td>57</td>
</tr>
<tr>
<td>3</td>
<td>Resources</td>
<td>59</td>
</tr>
<tr>
<td>3.1</td>
<td>Primary Datasets</td>
<td>59</td>
</tr>
<tr>
<td>3.2</td>
<td>Secondary Datasets</td>
<td>62</td>
</tr>
<tr>
<td>3.3</td>
<td>Classification and regression</td>
<td>70</td>
</tr>
<tr>
<td>3.4</td>
<td>Evaluation metrics</td>
<td>72</td>
</tr>
<tr>
<td>3.5</td>
<td>Summary</td>
<td>78</td>
</tr>
<tr>
<td>4</td>
<td>Predicting the Compositionality Degree of MWEs</td>
<td>79</td>
</tr>
<tr>
<td>4.1</td>
<td>Translation-based String Similarity</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>Translation Distributional Similarity Approach</td>
<td>111</td>
</tr>
<tr>
<td>4.3</td>
<td>Word Embeddings Approach</td>
<td>124</td>
</tr>
</tbody>
</table>
## Contents

4.4 Summary .................................................. 140

5 Detecting Non-compositional MWE Components using Wiktionary 141
  5.1 Introduction ............................................. 142
  5.2 Methodology ............................................. 143
  5.3 Experiments ............................................. 148
  5.4 Error Analysis .......................................... 151
  5.5 Summary ................................................. 152

6 Incorporating Compositionality into MT Evaluation 153
  6.1 Machine Translation Evaluation ......................... 154
  6.2 Using compositionality scores in Tesla ................. 156
  6.3 Dataset .................................................. 159
  6.4 Results .................................................. 159
  6.5 Discussion .............................................. 162
  6.6 Summary ................................................ 164

7 Determining the Multiword Expression Inventory of a Language 165
  7.1 Introduction .............................................. 166
  7.2 Background .............................................. 168
  7.3 Resources ............................................... 170
  7.4 MWE Patterns .......................................... 171
  7.5 Methodology ............................................ 175
  7.6 Results .................................................. 177
  7.7 Error Analysis .......................................... 184
  7.8 Conclusion ............................................... 186

8 Conclusion 188
  8.1 Summary of Chapters and Contributions ................. 189
  8.2 Future work ............................................. 194

A List of Languages 235
List of Figures

2.1 Continuum of semantic compositionality and semantic idiosyncrasy. . 16
2.2 An example of a 2D semantic space with three vectors. . . . . . . 42
2.3 CBOW vs. skip-gram model . . . . . . . . . . . . . . . . . . . . . . 50
2.4 An example of word-alignment. The idiom *kick the bucket* in English
in aligned to one word in Persian . . . . . . . . . . . . . . . . . . . . 52

3.1 An example of indirect translation . . . . . . . . . . . . . . . . . . . 65
org/wiki/Help:Export) . . . . . . . . . . . . . . . . . . . . . . . . . . . 66
3.3 A snapshot of Wiktionary entry for *natural language processing* . 68
3.4 SVM optimal hyperplane . . . . . . . . . . . . . . . . . . . . . . . . 70
3.5 True positive, False positive, False negative and True negative. . . 75

4.1 Schematic outline of our proposed string similarity approach using the
ith target language . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86
4.2 Schematic outline of our proposed method to combine the scores from
multiple languages . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
4.3 Relation between the compositionality of the first component and the
whole noun compound over the ENC dataset . . . . . . . . . . . . . . . 89
4.4 Relation between the compositionality of the second component and
the whole noun compound over the ENC dataset . . . . . . . . . . . . . 90
4.5 Relation between the coverage and the correlation score of target lan-
guages for the ENC dataset . . . . . . . . . . . . . . . . . . . . . . . . . . 93
4.6 Relation between different values of *α* and the correlation score for the
compositionality of English noun compounds . . . . . . . . . . . . . . . 94
4.7 Correlation between the coverage and the correlation score of target
languages for the verb component of the EVPC dataset . . . . . . . . 99
4.8 Relation between the coverage and the correlation score of target lan-
guages for the second component of the EVPC dataset . . . . . . . . 99
4.9 Relation between the compositionality of the first component and the
whole noun compound over the GNC dataset . . . . . . . . . . . . . . . 100
4.10 Relation between the compositionality of the second component and the whole noun compound over the GNC dataset ........................................... 101
4.11 Relation between the coverage and the correlation score of target languages for the GNC dataset .............................................................. 104
4.12 Relation between different values of $\alpha$ and the correlation of the compositionality score for the GNC dataset .......................................................... 106
4.13 Outline of our approach to computing the distributional similarity (DS) of translations of an MWE with each of its component words, for a given target language. $score_1$ and $score_2$ are the similarity for the first and second components, respectively. .................................................. 114
4.14 Outline of the method for combining distributional similarity scores from multiple languages, across the components of the MWE. $CS_{method}$ refers to any of the methods described in Section 4.2.2 for calculating compositionality. ................................................................. 115
4.15 Histograms displaying how many times a given $N$ is selected as the best number of languages over ENC, EVPC and GNC dataset. ............... 117
4.16 Token frequency effect on corpus-based approaches. WORD2VEC-$i$ and MSSG-$i$ represent the $i$th parameter settings of each method shown in Table 4.17. ................................................................. 132
4.17 The effect of window size parameter on WORD2VEC-1 ................. 134
4.18 The impact of vector size ................................................................. 135
4.19 The effect of the number of prototypes on the results with MSSG .... 136

6.1 Integrating compositionality scores of noun compounds into TESLA (shown in BOLD) ............................................................... 157

7.1 Overall patterns of co-occurrence of POS tags across compound and mwe dependency tuples in the selected languages ....................... 171
7.2 Cross-lingual similarity of MWE pattern distributions using JSD .... 172
7.3 Cross-lingual similarity of MWE pattern distributions using MSE ..... 174
7.4 Selecting a language as a source language ........................................ 178
7.5 Distribution of ROC AUC scores when combining source languages . 180
7.6 Combining source languages given the noisy dependency relations. In (a)–(c), the English treebank is used as the source language for the cross-lingual parser, and in (d)–(f) Swedish is used. ............. 183
## List of Tables

1.1 Multiword expressions (American vs. British English)  
1.2 Examples of compositionality scores  
2.1 Examples of statistical idiosyncrasy (adopted from Cruse (1986)). The correct, neutral and incorrect usages are shown with ‘+’, ‘?’ and ‘-’.  
2.2 English Examples of MWE types  
2.3 The semantic relation between the components of English noun compounds ($N_1$ is the modifier and $N_2$ is the head), adopted from Kim and Baldwin (2005).  
2.4 Compositionality degree of English noun compounds with their components and compound mean.  
2.5 Examples of English verb particle constructions with binary annotations for each component.  
2.6 Mutual information of red carpet and economic impact and their variants  
2.7 An example of semantic vectors for silver screen and its components.  
3.1 Examples from English noun compound (ENC) dataset, with their components and compound mean±deviation scores  
3.2 Examples from the English verb particle construction (EVPC) dataset, with the annotations for each component.  
3.3 Examples from the German noun compound (GNC) dataset, with their components and compound mean±deviation scores  
4.1 Examples of English MWEs and their components with translations into Persian. Direct matches between the translation of a MWE and its components are shown in **bold**; partial matches are *underlined*  
4.2 Correlation on the ENC dataset. N1, N2 and NC, are the first component of the noun compound, its second component, and the noun compound itself, respectively. Correlation using only distributional similarity (DS) is $r = 0.700$.  
4.3 The 10 best languages for the ENC dataset using LCS.
4.4 Correlation on EVPC dataset, based on the best-10 languages for the verb and particle individually. Correlation using only distributional similarity (DS) is \( r = 0.177 \).

4.5 Results for the classification task.

4.6 The 10 best languages for the verb/particle component of the EVPC dataset using LCS.

4.7 Correlation on GNCs dataset. N1, N2 and NC, are the first component of the noun compound, its second component, and the noun compound itself, respectively. Correlation using only distributional similarity (DS) is 0.141.

4.8 The 10 best languages for German noun compounds.

4.9 The results after ignoring missing translations on GNC dataset.

4.10 The 10 languages with the highest translation coverage for ENC, EVPC and GNC (Common languages with the supervised selected languages are shown in bold).

4.11 Correlation on ENC, EVPC and GNC datasets using supervised and unsupervised string similarity approach.

4.12 Results of distributional similarities using 10 best languages on ENC dataset (N is window size).

4.13 The 5 best languages along with the language family and the number of times each language is selected in our 10×10 cross-validation folds setup, for ENC, EVPC and GNC.

4.14 Pearson’s correlation on the ENC, EVPC and GNC datasets.

4.15 Results (%) for the binary compositionality prediction task on the EVPC dataset.

4.16 Parameter settings of the examined methods (\( d = \)vector dimensionality and \( w = \)window size). The \textsc{word2vec}-1 and MSSG-1 parameters were set based on the default recommendations of Mikolov et al. (2013a) and Neelakantan et al. (2014). The settings were changed in other examined methods, accordingly.

4.17 Pearson’s correlation (\( r \)) for the different methods (Table 4.16) over the three datasets.

4.18 The frequency information for each dataset excluding the zero values.

4.19 Pearson’s correlation (\( r \)) for the different operators to combine vector representations provided by MSSG, over the three datasets. SS stands for the string similarity approach.

5.1 Lexical coverage of WordNet, Wiktionary and Wiktionary+Wikipedia over our two datasets.
5.2 Compositionality prediction results over the ENC dataset, relative to the first component (the modifier noun) and the second component (the head noun). The precision and recall calculated for the first and second component differ because the compositionality judgements differ between components. ........................................ 149
5.3 Compositionality prediction results over the EVPC dataset, relative to the first component (the head verb) and the second component (the particle) ........................................ 150

6.1 An example of a reference translation, and the output of two machine translation systems. ........................................ 155
6.2 Kendall’s ($\tau$) correlation over WMT 2013 (all-en), for the full dataset and also the subset of the data containing a noun compound in both the reference and the MT output ........................................ 159
6.3 Pearson’s ($r$) correlation results over the WMT all-en dataset, and the subset of the dataset that contains noun compounds ................................................... 160
6.4 The number of judgements that were ranked correctly by TESLA originally, but incorrectly with the incorporation of compositionality scores (“P→N”) and vice versa (“N→P”), and the absolute improvement with compositionality scores (“$\Delta$”) ........................................ 161
6.5 Two examples from the all-en dataset. Each example shows a reference translation, and the outputs of two machine translation systems. In each case, the output of MT system 1 is annotated as the better translation. ........................................ 163

7.1 List of examined languages in this paper with their respective percentage of MWE tokens. Languages with a higher proportion of MWE tokens and patterns are shown in bold. ........................................ 170
7.2 Average AUC scores for each source language. ........................................ 179
7.3 The best source languages for predicting the MWE patterns in each of the 5 selected languages. ........................................ 181
7.4 The top predicted MWE patterns in English, by combing all other languages. ........................................ 181
7.5 AUC scores when English is used as the source language to transfer dependency links and to train our reranker model. In “Baseline + gold”, the trained model is applied to the gold-standard annotation of the target language rather than the parsed corpus. ........................................ 182
7.6 Top-ranking Persian and Croatian MWE patterns extracted using English and Swedish as the source language. Those patterns which match the top-ranking gold standard patterns are shown with “†”. ........................................ 184

A.1 List of languages used in Sections 4.1 and 4.2 ........................................ 236
Citations to Previously Published Work

Large portions of Chapter 4 have appeared in the following papers:


Large portions of Chapter 5 have appeared in the following paper:


Large portions of Chapter 6 have appeared in the following paper:


Large portions of Chapter 7 have appeared in the following paper:

Acknowledgments

Firstly, I would like to offer my sincere gratitude to my two deeply knowledgeable, extensively experienced, and truly friendly supervisors, Tim Baldwin and Paul Cook. I am grateful for their great patience, continuous support, professional integrity and insightful feedback on my academic and even personal development. I could not have imagined having better supervisors for my Ph.D. study. I am also thankful to Chris Leckie and Steven Bird, my thesis committee, who have always provided me with helpful and practical suggestions on my research and future career. My sincere thanks also goes to my master’s thesis supervisor, Afsaneh Fazly, for introducing me to multiword expressions, and encouraging me to do a Ph.D. in this area.

During my candidature, I met brilliant, dedicated and yet humble and friendly colleagues in the Melbourne NLP group. In particular, I wish to name and thank Miji Choi, Ping Ping Tan, Nitika Mathur, Doris Hoogeveen, Ekaterina Vylomova, Long Duong, Fei Liu, Ned Letcher, Yitong Li, Afshin Rahimi, Julian Brooke, Oliver Adams, Andrew Bennett, Huizhi Liang, Bo Han, Jey Han Lau, Marco Lui, Meladel Mistica, Li Wang, Florian Hanke, Richard Fothergill, Spandana Gella, Yvette Graham, Su Nam Kim, Karin Verspoor and Trevor Cohn. Their invaluable suggestions on professional subjects as well as their support on personal matters made this journey especially pleasant and memorable.

I am grateful for the financial and travel supports I received from the University of Melbourne, NICTA VRL and Google during my research.

I would like to extend my gratitude to my Iranian friends Azadeh Ghari Neiat, Shaahin Madani, Mahtab Mirmomeni, Yousef Kowsar, Sargol Sadeghi, Mohsen Laali, Reza Soltanpour, Mahsa Zamani, Amin Sadri and Mohammad Oloomi for their companionship.

I wish to express my sincerest thanks, love and deep appreciation to my family. My uncle Mehran and his lovely family, Khatereh, Shervin and Artin, and my cousin Solmaz were always there to help me and to feel less homesick. I appreciate all the love and support I have ever received from my parents Vahid and Jamileh, my lovely sister Setareh, and my Borhan who happens to be my dearest friend.
Chapter 1

Introduction

This thesis is about multiword expressions (hereafter, MWEs) and determining their properties in not only English, but any language, where we have access to a lexicon or text corpus. In this chapter, we briefly introduce MWEs, and their importance in language and language processing. Then, we explain the scope and contributions of this thesis, and finish the chapter by introducing the upcoming chapters.

1.1 Background and Motivation

MWEs are groups of words that function as a single unit. In this thesis the single words that make MWEs are called components. The most obvious and typical example of MWEs are idioms, where the meaning of the idiom cannot be understood from knowledge of its components.

A few examples of idioms are:

(1.1) We were shooting the breeze for a long time.
(1.2) The poor man *kicked the bucket* yesterday.

(1.3) Buying a house *costs an arm and a leg*.

*Shoot the breeze* in (1.1) involves neither *shooting* nor any *breeze*. There was no *bucket* in (1.2) and the man did not actually perform the act of *kicking*, and buying a house in (1.3) does not lead to any physical injuries!

However, idioms are not the only MWEs in a language. Some MWEs are more transparent, in that the meaning is to some extent understandable from the components. However, the choice of components is fixed and cannot be replaced by similar words to convey the same meaning. For example, for a non-native speaker, it is not clear why we have *strong coffee*, *gin and tonic* and *home sweet home* in English, but not *powerful coffee*, *tonic and gin* and *house sweet house*. There is no apriori reason why the components occur in this particular combination. A native speaker uses them without thinking twice about it. In fact, the famous American comedian, George Carlin, uses this characteristic of MWEs (*near miss*, *get on* and *non-stop flight*) to comedic effect by pondering about them:

... Here’s a phrase that apparently the airlines simply made up: **near miss**. They say that if two planes almost collide, it’s a near miss. ... It’s a near hit! A collision is a near miss ...

... I’m told to **get on** the plane ... And I think for a moment. On the plane? No, my friends, not me. I’m not getting on the plane; I’m getting in the plane! ... It seems less windy in there ...

... Then they mention that it’s a **non-stop flight** ... Call me old-fashioned, but I insist that my flight stop!...

One way of diagnosing whether a particular combination of words constitutes an MWE is to consider it in other languages. A meaningless word-for-word translation
Chapter 1: Introduction

#### Table 1.1: Multiword expressions (American vs. British English)

<table>
<thead>
<tr>
<th>British English</th>
<th>American English</th>
</tr>
</thead>
<tbody>
<tr>
<td>telephone box</td>
<td>telephone booth</td>
</tr>
<tr>
<td>car park</td>
<td>parking lot</td>
</tr>
<tr>
<td>take away</td>
<td>take out</td>
</tr>
<tr>
<td>lay the table</td>
<td>set the table</td>
</tr>
<tr>
<td>lady bird</td>
<td>lady bug</td>
</tr>
<tr>
<td>have a nap</td>
<td>take a nap</td>
</tr>
</tbody>
</table>

of a phrase is an indication of being an MWE. The word-for-word translation of *zamin khordam* from Persian to English is *I ate the ground*, which is not the correct translation and shows that we have encountered an MWE in Persian. This should be instead translated to *I fell down* to make sense in English. Another example is *evil eye* in English which is expressed as *salty eye* in Persian, *bad sight* in Hindi and *envious eye* in Arabic. In this example, the translations are each slightly different from the word-for-word translation.

While MWEs can be expressed differently in different languages, they can also be used to distinguish speakers of various dialects of the same language. Table 1.1 shows MWE examples used differently in British and American English.

One of the challenges with MWEs is that they can be found in different forms:

(1.4) *shoot the breeze, take away, lady bug*

(1.5) *non-stop, long-term, self-esteem*

(1.6) *honeymoon, postman, lighthouse*

(1.7) *give somebody applause, piss someone off, switch something off*
In (1.4) and most of our examples so far, the components are contiguous and recognizable by a space in between the components. In (1.5) the components are concatenated with a hyphen. In (1.6) however, the components are contiguous but with no white space. This phenomenon is commonly observable in Germanic languages. For example in German *Rechtsschutzversicherungsgesellschaft* is the longest German word in everyday use, according to the 1995 Guinness Book of World Records, with 39 characters, and means “an insurance company that provides legal protection”. Detection of MWEs can become more complicated when the components do not occur continuously such as the examples in (1.7). Additionally, there exist MWEs which can be observed in multiple forms such as *tradeoff* and *trade-off*.

Some MWEs can be used in both literal and idiomatic senses. Detecting which sense is used in a context, is another challenge. For example, in English *to kick the bucket* can have two meanings: (1) “to hit a bucket with the foot”, or (2) “to die”. Another example in Persian is *to drink cold water* which has the literal meaning as well as “to get jailed”. Even more challenging MWEs are those with more than one non-literal sense such as *look up*, which has more than one non-literal sense apart from the literal sense of looking upwards:\(^1\)

- (Of a situation) Improve: things seemed to be *looking up* at last
- (informal) Make social contact with someone: *he would look her up when he was in the area*
- Search for and find a piece of information in a book or database: *the translation*

\(^1\)Definitions and examples are from [http://www.oxforddictionaries.com/definition/english/look](http://www.oxforddictionaries.com/definition/english/look)
The number of MWEs in a language is estimated to be of the same magnitude as the simplex words (Jackendoff 1996). This, besides the above-mentioned specific behaviour of MWEs, shows the importance of considering and modeling them in natural (human) language processing systems. For example, in an information retrieval system, if we were looking for documents relating to ivory tower, we would not be interested in documents on tall buildings. Machine translation is another application, where the importance of MWEs has been widely studied and many approaches have been proposed to integrate MWEs into machine translation systems.

1.2 Aim and Scope

The fact that MWEs are widely observable in every language, and the need for language technologies across all languages, makes it necessary to propose methods which are as general as possible, and can be applied to a wide range of languages.

Among the MWE related tasks, there are three tasks that we focus on in this thesis and will introduce in this section:

- Modeling the compositionality degree of MWEs
- Incorporating the compositionality of MWEs into a natural language processing (NLP) application
- Determining the MWE inventory of a surprise language
### 1.2.1 Modeling the Compositionality Degree of MWEs

Our first focus is on the semantics of MWEs, in measuring to what extent the meaning of the components is reflected in the meaning of the MWEs. This degree is called *compositionality*, signifying the extent to which the meaning of the MWE is a composition of its components. For example, *climate change* is highly compositional, and by knowing the meaning of *climate* and *change*, it is easy to guess the meaning of *climate change*. On the other hand, *silver screen* is less compositional, because *silver screen*, which refers to “cinema”, has the meaning of *screen*, but not *silver*. *Rat run* (in the sense of “a small road used as a shortcut to avoid the traffic on major roads”) is very non-compositional and more idiomatic than the other two examples. Obviously, by knowing only the meaning of *rat* and *run*, it is hard to guess the meaning of *rat run*. Table 1.2 shows annotated compositionality scores for these three MWEs and their components, as reported by the annotators of Reddy *et al.* (2011).

In this thesis, we specifically explore methods which are language-independent and can be applied to any kind of MWE across a range of languages. To this end, we make use of resources which are available in a wide range of languages.

#### Table 1.2: Examples of compositionality scores

<table>
<thead>
<tr>
<th>MWE example</th>
<th>Degree of compositionality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First component</td>
</tr>
<tr>
<td>climate change</td>
<td>98%</td>
</tr>
<tr>
<td>silver screen</td>
<td>28%</td>
</tr>
<tr>
<td>rat run</td>
<td>8%</td>
</tr>
</tbody>
</table>
1.2.2 Incorporating compositionality scores into NLP applications

As mentioned in Section 1.1, multiword expressions are estimated to be as frequent as simplex words, and therefore an NLP application with any degree of linguistic precision should be able to handle them. When it comes to semantics, the compositionality of MWEs is particularly important. Recent studies (Tsvetkov and Wintner 2010; Weller et al. 2014; Carpuat and Diab 2010; Venkatapathy and Joshi 2006) have shown that incorporating MWEs compositionality into machine translation systems can improve translation quality. In this thesis, we examine the incorporation of compositionality scores into other NLP applications, namely a machine translation evaluation system.

In this thesis, our analysis of compositionality is limited to specific kinds of MWEs which are easy to find in a given body of text.

1.2.3 Determining the MWE inventory of a surprise language

Different patterns of MWEs are observable in different languages, such as Noun–Noun compounds in English, Adj–Noun compounds in Persian and Verb–Verb compounds in Japanese. Some of these patterns are very common in one language but rare in another.

Most research on finding MWEs in textual data has made explicit assumptions about what patterns are more common in that language (Lapata and Lascarides 2003;
Tanaka and Baldwin 2003; Baldwin 2005), or used annotated instances of MWEs to train models to find other instances (Diab and Bhutada 2009; Li and Sporleder 2010; Schneider et al. 2014a). But what if we do not have such linguistic insights for a given language? In this thesis, we propose a method which discovers the common patterns of MWEs in a given language without using any existing annotation of MWE instances or language-specific knowledge. Instead, our proposed model learns the behaviour of MWE patterns from a resource-rich language, which can be used to predict the (possibly substantially different) MWE patterns in a resource-poor language.

1.3 Contributions

This section outlines the contributions of this thesis.

- We investigate three approaches to predict the overall compositionality degree of MWEs, and one unsupervised approach to detect the individual non-compositional components of MWEs. These approaches do not make any language specific assumptions, and can be applied to any type of MWE in a wide range of languages. To evaluate our work, we use three datasets to examine two types of MWEs across two languages (English and German).

- Using a multilingual dictionary with coverage over more than one thousand language varieties, we explore a translation-based string similarity approach to predict the compositionality degree of MWEs. In this approach, we translate the MWE and its components to multiple languages, and use string similarity between the translation of the MWE and its components to measure the
compositionality. Based on a training set, we find the best target languages for a specific type of MWE in a specific language. Our best results are shown to be competitive with state-of-the-art results. The results also show that this approach complements state-of-the-art methods.

- Based on our observation of the utility of cross-lingual information, we measure the distributional similarity between the translations of MWEs and their components in multiple languages. In order to measure the distributional similarity, we do not perform any language-specific pre-processing. The strength of this approach is particularly for MWEs which are hard to identify. Our results also show that the translation-based string similarity approach complements this approach. Together, these approaches achieve state-of-the-art results for compositionality prediction.

- We report on the first application of word embeddings in predicting the compositionality degree of MWEs. We compare this approach with a traditional distributional similarity approach. Similar to our translation-based distributional similarity approach, we show that this approach complements the translation-based string similarity approach, and vice versa. We also show that this approach is insensitive to the hyper parameter settings.

- We investigate using Wiktionary to detect non-compositional components of MWEs. In this unsupervised approach, we propose to use definitions and translations provided in Wiktionary, and explore the use of different methods for combining the multiple senses of MWEs. Our results reveal the superiority of
this approach to supervised state-of-the-art approaches.

- To demonstrate the importance of including MWE compositionality in NLP applications, we present the first attempt to integrate MWE compositionality into machine translation evaluation. The main intuition underlying this work is that we should reward partial mismatches between machine translation outputs and the reference translations, based on compositionality. Our experiments show a marginal improvement when compositionality scores are considered.

- In addition to proposing general language-independent approaches to measure the compositionality of MWEs, we explore a language-independent approach to determine the MWE inventory of a language. Our proposed approach has no explicit prior knowledge of MWE patterns, or annotated MWE data of the examined language. This model is trained on a treebank with MWE relations of a source language, and can be applied to identify the MWE construction types in another language.

1.4 Thesis Outline

Chapter 2: This chapter provides a detailed definition of MWEs and explains their specific characteristics. Then, we introduce a number of famous MWE types and discuss their individual characteristics. We also review the computational linguistic literature on MWEs, and discuss previously proposed approaches to predicting the compositionality of MWEs along with their strengths and shortcomings. We finish this chapter by reviewing previously proposed methods to integrating MWEs and
their compositionality degree into NLP applications.

Chapter 3: This chapter describes the three datasets which we use to evaluate our proposed methods, and the five primary resources used in this research. Additionally, we explain the evaluation metrics and machine learning tools we use throughout this thesis.

Chapter 4: This chapter details our three proposed approaches to predicting the compositionality of MWEs: (1) translation-based string similarity (Salehi and Cook 2013), (2) translation-based distributional similarity (Salehi et al. 2014b), and (3) word-embeddings (Salehi et al. 2015a). These approaches do not make any assumptions about the language or the types of MWEs. We report empirical results and discuss the strengths and shortcomings of each approach.

Chapter 5: In this chapter, we propose an unsupervised approach to detecting the non-compositional components of MWEs (Salehi et al. 2014a). The proposed approach makes use of the definitions, synonyms and translations provided in Wiktionary. We additionally investigate different approaches to handle the polysemy of MWEs. Finally, we report empirical results and compare this approach with the state-of-the-art methods, in addition to the approaches proposed in Chapter 4.

Chapter 6: This chapter presents the first attempt to integrate compositionality scores into machine translation evaluation (Salehi et al. 2015b). We explain the intuition behind this work and introduce TESLA, the machine translation evaluation system we use in our experiments. Then, we will explain our approach to integrating compositionality scores into TESLA, and compare the proposed method with the original TESLA as well as other popular metrics in this area. Finally, we provide error
analysis and discussion of the results.

**Chapter 7:** This chapter provides insights into existing annotations of MWE patterns across a number of languages in the Universal Dependency formalism (Salehi et al. 2016). We describe the first attempt to determine the MWE inventory of a language using the MWE annotations of another language. We also examine combinations of languages to train our model.

**Chapter 8:** This chapter summarizes the contributions and outcomes of this thesis. We will also present possible future research directions, and finish the chapter by discussing untouched, but interesting topics in this research area.
Chapter 2

Literature Review

This chapter reviews the literature on multiword expressions (MWEs). First, we provide a detailed overview of multiword expressions along with a description of their linguistic features. The most famous types of MWEs are introduced in Section 2.2. In Section 2.3, various computational studies of MWEs are discussed. Next, previous studies on measuring the compositionality of MWEs are surveyed in Section 2.4. Finally, we review the potential impact of MWE identification and compositionality prediction on NLP applications in Section 2.5.

2.1 What is a multiword expression?

In the literature, a multiword expression is defined more or less similarly, in being (1) a combination of words, which (2) exhibits characteristics not predictable from the components. Some of the previous definitions are provided below:

Def 1: “MWEs can be broadly classified into lexicalized phrases and institutionalized
phrases” (Bauer 1983; Sag et al. 2002)

Def 2: “a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components” (Choueka 1988)

Def 3: “arbitrary and recurrent word combination” (Smadja 1993)

Def 4: “idiosyncratic interpretation that cross word boundaries” (Sag et al. 2002)

Def 5: “a sequence of words that acts as a single unit at some level of linguistic analysis” (Calzolari et al. 2002)

Def 6: “Lexical items that (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity.” (Baldwin and Kim 2009)

Def 7: “lexicalized combination of two or more words that are exceptional enough to be considered as single unit in the lexicon” (Schneider et al. 2014a)

Def 8: “a group of tokens in a sentence that cohere more strongly than ordinary syntactic combinations: that is, they are idiosyncratic in form, function, or frequency.” (Schneider et al. 2014b)

In this thesis, we adopt the definition provided by Baldwin and Kim (2009), which, we believe covers the other definitions as well. The first requirement for an MWE is that it contains more than one lexeme. The individual lexemes (components) can be easily detectable if there is a space between the components (such as kick the bucket and spelling bee) or a hyphen (such as long-term and knowledge-based). Some
MWEs can be found in different forms such as *trade off*, *trade-off* and *tradeoff*. In Persian, a half space (smaller than a space) is used between the components of some MWEs such as *پایان نامه* ("thesis", literally payan ("end") + naameh ("letter")) instead of a space, *پایان نامه*, or the joint form of *پایان نامه*. There are also MWEs which look like single words in that there is no space between the components (e.g., *tradeoff*). Such MWEs are especially frequent in Germanic languages (German: *Erdbeere* (*Erde* + *Beere*, "strawberry"), Swedish: *markköp* (*mark* + *köp*, "land purchase")) (Bauer 2001). The process of identifying the individual components of such MWEs is called decompounding (Koehn and Knight 2003a). In non-segmenting languages such as Japanese and Chinese, there is no space used to separate words of a sentence nor the components of MWEs.

The second characteristic of MWEs is idiosyncrasy, meaning that some properties of the components are not reflected in the MWEs or vice versa. Idiosyncrasies can be lexical, syntactic, semantic, pragmatic and/or statistical in nature, which we explain in more detail in the remainder of this section. It is important to mention that idiosyncrasy in MWEs is a continuum. For example, for semantic idiosyncrasy, or the degree that the MWE is idiomatic, *kick the bucket* is more idiosyncratic than *make a decision*, which in turn is more idiosyncratic than *swimming pool*. A related notion to idiosyncrasy is compositionality, which is the degree that the properties of an MWE are the result of combining the properties of its components. In terms of semantic compositionality, *swimming pool* is more compositional than *make a decision*, and both are more compositional than *kick the bucket* (See Figure 2.1).

In the following, each of lexical, syntactic, semantic, pragmatic and statistical
idiosyncrasy with respect to MWEs is explained in more detail.

2.1.1 Lexical idiosyncrasy

An MWE has lexical idiosyncrasy when there is no entry for one or more of the components in the lexicon of the language (Bauer 1983; Baldwin and Kim 2009). Examples are ad hoc, a priori, al fresco and et al. All these examples are originally from Latin, and although the MWE has a meaning in English, the individual components do not. Despite the non-existence of the individual components in English, the same components can be observed in other English MWEs such as et in et al. and et cetera. There are also examples of meaningless components with no origins, such as reduplication\(^1\) of mish in mish mash, vai in chai vai (“tea”) in Hindi, and miz in

\(^1\)Reduplication is the process of repeating a word, or with small change in the vowels or first consonants.
chiz miz (“things”) in Persian.

MWEs with lexical idiosyncrasy will inevitably have syntactic and semantic idiosyncrasy, because there is no syntactic or semantic property attached to the components in that language.

### 2.1.2 Syntactic idiosyncrasy

The MWEs with syntactic idiomaticity have syntactic characteristics which are different from the components (Bauer 1983; Sag et al. 2002; Baldwin and Kim 2009). Some examples of MWEs with syntactic idiosyncrasy are *by and large* (made up of a preposition and an adjective, but used as an adverb), and *wine and dine* (made up of two intransitive verbs but used as a transitive verb compound).

Some MWEs tend to have limited syntactic flexibility as semantic idiosyncrasy increases (Nunberg et al. 1994; Fazly and Stevenson 2006; Fazly et al. 2009). For example, the semantically idiosyncratic *shoot the breeze* is not syntactically as flexible as a compositional verb+object combination; for instance it cannot be passivized (i.e., *#the breeze was shot*).

### 2.1.3 Semantic idiosyncrasy

Semantic idiosyncrasy occurs when the meaning of the MWE is not predictable from its components (Bauer 1983; Sag et al. 2002; Baldwin and Kim 2009). As mentioned earlier, semantic idiosyncrasy is also referred to as semantic compositionality in literature, which shows the amount that the meaning of components is reflected in the meaning of MWE. The degree of semantic idiosyncrasy, as with
other idiosyncratic behavior of MWEs, can vary from fully idiomatic to fully compositional. Idioms (introduced in Section 2.2.4) are the extreme case of semantic idiosyncrasy for which the meanings are totally different from the meaning of the components. The meaning of an idiom is nearly impossible to guess for a language learner, and excessive usage of idioms is indicative of being a native speaker (Grant 2003). Some MWEs are figurative and while the individual parts do not convey the meaning, the meaning can be conveyed through visual metaphors (Nunberg et al. 1994; Grant 2003), such as *add fuel to the fire* and *kill two birds with one stone*.

There are MWEs for which only a subset of the components exhibit the meaning of the MWE such as *eat up*, which is related to *eating* but the sense of *eating thoroughly* is exhibited by adding the particle *up*. There are also MWEs in which the components partially convey the meaning. For example, the term *give in give orders* is more figurative and is considered to be somewhat semantically idiosyncratic.

On the other hand, some MWEs are more compositional (less idiomatic) such as *bus driver* (driver that drives a bus), *swimming pool* (pool used for swimming), *apple pie* (pie made from apples) and *morning juice* (juice for mornings) (Barker and Szpakowicz 1998; Kim and Baldwin 2005). Despite being more compositional, these MWEs still exhibit extra semantic information implicitly. For example both *apple juice* and *morning juice* are semantically compositional, however there is an implicit information which says *apple juice* is the juice of apples, while *morning juice*, is juice made for consumption in the morning.
2.1.4 Pragmatic idiosyncrasy

Multiword expressions with pragmatic idiosyncrasy are combinations of words which are associated with a situation, event, or specific context (Jackendoff 1996; Sag et al. 2002; Baldwin and Kim 2009). Example of MWEs with pragmatic idiosyncrasy are *all aboard* (the term used to announce the departure of the train, ship or flight) and *gin and tonic* (the specific term used for a kind of beverage).

2.1.5 Statistical idiosyncrasy

Multiword expressions with statistical idiosyncrasy are those with surprising statistical behavior in terms of frequency. While such behavior is observable for a group of MWEs, it is hard to find a concrete definition for it in linguistics, thus several definitions are provided in literature (Petrović et al. 2010). The surprising statistical behavior can be defined as (1) when the combination of words occur together more often than expected by chance, or (2) when a combination of words is more frequent than the other variants of the phrase by replacing the words with their synonyms. (Cruse 1986; Pearce 2001; Sag et al. 2002)

An example of statistical idiosyncrasy, adopted from Cruse (1986), is shown in Table 2.1. The example shows the correct (shown with ‘+’), neutral (shown with ‘?’) and incorrect (shown with ‘-’) usage of six semantically different nouns with four synonymous adjectives (*flawless, immaculate, impeccable* and *spotless*). The correct and incorrect usages are mainly observed by counting the number of times each combination is observed within a corpus. The results show that some combinations are more frequent than the other variants, or in other words, they have statistical
This definition is closely related to the notion of institutionalisation, which relates to the preferred way of using a combination of words to refer to a concept or an object (Bauer 1983; Nunberg et al. 1994; Sag et al. 2002). For example, strong coffee is preferred to powerful coffee. For a new learner of language, it might not be clear why strong is preferred to powerful, while for a native speaker, strong coffee sounds right without being able to explain why it feels right. The variants of MWEs which are less preferred are called anti-collocations (Pearce 2001). Interestingly, the preferred combination can be utilized to distinguish between the various dialects of a language, such as mail man vs. post man, and phone booth vs. phone box: the first combinations are preferred in American English, while the latter are more common in British English.

Some binomials contain statistical idiosyncrasy, where they refer to a single concept and there is a preference on the order of the components. Examples are black and white, salt and pepper and gin and tonic, for which white and black, pepper and salt and tonic and gin do not sound correct to a native speaker and do not preserve the initial meaning (Benor and Levy 2006). The reversed order, however, might be

<table>
<thead>
<tr>
<th></th>
<th>flawless</th>
<th>immaculate</th>
<th>impeccable</th>
<th>spotless</th>
</tr>
</thead>
<tbody>
<tr>
<td>condition</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>credentials</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>hair</td>
<td>–</td>
<td>+</td>
<td>?</td>
<td>–</td>
</tr>
<tr>
<td>house</td>
<td>?</td>
<td>+</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>logic</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>timing</td>
<td>?</td>
<td>+</td>
<td>+</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 2.1: Examples of statistical idiosyncrasy (adopted from Cruse (1986)). The correct, neutral and incorrect usages are shown with ‘+’, ‘?’ and ‘-’.

Idiosyncrasy.
Chapter 2: Literature Review

2.2 Types of MWEs

In this section, we introduce some of the more widely researched MWE types and their characteristics (examples are provided in Table 2.2). The purpose of this section is to firstly show the difficulties of dealing with MWEs, as they occur across a wide range of types with varying characteristics. Secondly, in understanding the better-known types of MWEs, the reader will have a better appreciation for previously proposed type-specific approaches to determining MWE compositionality. Thirdly, we will introduce a cross-lingual approach to investigate the patterns of existing MWE types in a target language in Chapter 7, which requires understanding of the better-known MWE types.

2.2.1 Words-With-Spaces

The simplest type of MWEs to identify is a sequence of words with a predetermined order, known as a word-with-spaces (Sag et al. 2002). MWEs such as in short, ad hoc

<table>
<thead>
<tr>
<th>MWE type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words-with-spaces</td>
<td>ad hoc, in short, by and large</td>
</tr>
<tr>
<td>Noun compound</td>
<td>swimming pool, ivory tower, guilt trip</td>
</tr>
<tr>
<td>Verb particle construction</td>
<td>look up, figure out, clean up</td>
</tr>
<tr>
<td>Light verb construction</td>
<td>give a demo, make a mistake, take a nap</td>
</tr>
<tr>
<td>Idiom</td>
<td>kick the bucket, spill the beans, cat got your tongue</td>
</tr>
</tbody>
</table>

Table 2.2: English Examples of MWE types

preferred in other languages, such as black and white in English, versus “white and black” in Japanese (shirokuro) and Spanish (blanco y negro) (Baldwin and Kim 2009).
and by and large are examples of this category. Note that the sequence is syntactically fixed, e.g., #in shorter or #in very short are not valid syntactic variants of in short. The MWEs in this category mostly have lexical and/or statistical idiosyncrasy.

### 2.2.2 Noun Compounds

Noun compounds (NCs) are one of the most common MWE types in many languages (Tanaka and Baldwin 2003). They are usually composed of two elements with a head noun such as swimming pool and ivory tower, but can also be longer than two elements, such as daylight saving time (Sag et al. 2002; Huddleston and Pullum 2002; Kim and Baldwin 2005).

In terms of semantic idiosyncrasy, NCs vary from compositional to idiomatic, which makes them interesting to study. It is important to mention that NCs are not fully compositional, meaning that the components hold an implicit meaning together, such as swimming pool which is the pool used for swimming. The more compositional NCs are very productive such as orange juice, for which the first component can be replaced with similar words, like apple.

Noun compounds, in contrast with words-with-spaces, are more flexible; they can be pluralized as in [swimming pool]s. But they are less flexible than verb-based MWEs, which are introduced next.

### 2.2.3 Verb-based MWEs

Verb-based MWEs behave syntactically or semantically as unique verbs, motivating Blaheta and Johnson (2001) to refer to them as multiword verbs (MWVs). Two
important groups of verb-based multiword expressions are verb particle constructions (VPCs) and light verb constructions (LVCs).

**Verb Particle Constructions**

Verb particle constructions (VPCs) consist of a verb (as the head component) and a particle, which is usually a preposition (such as *hand in, look up, battle on*), an adjective (such as *cut short*) or a verb (such as *let go*) (Baldwin 2005; Villavicencio and Copestake 2002; Huddleston and Pullum 2002).

Much of the work on VPCs has focused on prepositional VPCs (or VPCs with a preposition as a particle), particularly because prepositional VPCs are very productive and more frequent than VPCs with other particles in English, e.g., the verb *turn* can form a VPC with different prepositions: *turn up, turn out, turn in, turn down, turn off* and *turn around*.

VPCs are mostly either transitive (e.g., *look up, hand in*) or intransitive (e.g., *break up, pitch in*). Transitive VPCs can be non-contiguous, i.e., a noun phrase can be placed between the verb and the particle, such as:

(2.1) Think carefully before you **hand** the paper **in**.

(2.2) **Look** it **up** in a dictionary.

Intransitive VPCs are generally contiguous (Huddleston and Pullum 2002), although there are some exceptions such as inserting one of a small set of adverbs (*back, right, straight, way* and *well*) between the verb and its particle such as *jump well up* or *come right back* (Baldwin 2005).
On the semantic side, VPCs are highly polysemous and sometimes can occur in both transitive and intransitive senses (Villavicencio and Copestake 2002), such as add up in:

(2.3) Add up the numbers! (Transitive)

(2.4) Save money, it all adds up! (Intransitive)

Polysemy and the non-contiguous behavior of VPCs make them difficult to extract or identify, as discussed in Section 2.3.1. For example, consider put on in the following examples:

(2.5) **Put** your clothes **on**!

(2.6) **Put** your clothes **on** the table!

*Put on* is a VPC in (2.5), but not in (2.6). Huddleston and Pullum (2002) discuss a number of diagnostics to identify English VPCs:

- The verb and the preposition can be separated (*split* form) by a noun phrase, as the object (e.g., Example 2.5).
- If the object is a pronoun, the VPC should be used in split form.
- Manner adverbs cannot be used between the verb and the particle.

Recently, PARSEME\(^2\) has published an annotation guideline to distinguish VPCs from verb-preposition combinations.\(^3\) In this guideline, a verb+preposition is VPC if:

\(^2\)An interdisciplinary scientific network in Europe, devoted to the role of MWEs in parsing.

\(^3\)http://typo.uni-konstanz.de/parseme/images/shared-task/guidelines/PARSEME-ST-annotation-guidelines-VPC-IPrepV.pdf
• Changing the sentence containing the VPC to a question formation leaves the preposition intact.

   – I went to the shop. Where did you go? (non-VPC)
   – He turned the TV off. What did he turn off? (VPC)

• A noun can be derived from it such as work out and workout.

• There is a synonym verb for it, such as to try out and to run away are synonyms with to test and to escape, respectively.

Light Verb Constructions

A light verb construction (LVC) is a combination of a semantically light verb and another component (a.k.a. co-verbal element) which can be a noun, verb, adjective or a prepositional phrase (Wierzbicka 1982; Sag et al. 2002; Huddleston and Pullum 2002; Butt 2003; Stevenson et al. 2004). “Semantically light verb” means that the meaning of the LVC is largely determined by the noun component, and the verb part loses its original meaning (e.g., compare make in make an offer (“to offer”), with make a cake which is not an LVC and more transparently conveys the meaning of both make and cake). The set of light verbs in English are make (e.g., make an offer, make a decision), take (e.g., take a nap, take a photo), have (e.g., have a rest, have a thought), give (e.g., give a hug, give a demo), get (e.g., get support, get a ride) and do (e.g., do a favor, do a review). LVCs also occur frequently in languages other than English, such as Japanese (Miyamoto 2000), Italian (Alba-Salas 2002), Persian (Karimi-Doostan 1997), Hindi (Mohanan 1994) and Korean (Ahn 1991). Different
types of co-verbal elements are possible in different languages, such as adjectives and adverbs in Persian (Megerdoomian 2004), or verbs in Japanese (Uchiyama et al. 2005).

In terms of language modeling, the choice of light verb is largely predictable for a given noun component: e.g., give a demo or do a demo instead of #take a demo (Baldwin and Kim 2009).

Semantically, LVCs are known to be semi-compositional, since the meaning is determined primarily by the co-verbal element rather than the light verb (Wierzbicka 1982). Such LVCs might be: (1) morphologically related to a verb (take a walk and walk), (2) etymologically related to a verb (give a speech and speak), or (3) not related to a verb (give a demo) (Fazly 2007). In the case of LVCs related to a verb, there is a slight difference between the meaning of the LVC and the relative verb, such as the slight semantic difference between take a walk vs. walk, and make a decision vs. decide. Light verbs may also exhibit some degree of semantic compositionality such as in give a speech but they are not as literal as in give a book, where the possession of an item is transferred (Fazly et al. 2005).

Syntactic flexibility of LVCs makes them hard to extract and identify, as many LVCs are syntactically as variable as literal phrases. Consider make a decision as an example of a semi-compositional LVC and make a cake as a literal phrase. Here are some examples that show the syntactic flexibility of an LVC:

(2.7) **make** the wrong **decision**.

(2.8) how many **decisions** did you **make**?!

---

4See Section 2.3.1
(2.9) the decision is made.

Note that by replacing decision with cake, we completely replace the LVC with a literal combination of verb+noun with the same syntactic flexibility. While many LVCs are syntactically flexible, the noun component is less flexible morphologically and allows only singular forms with an indefinite determiner (a or an) (Stevenson et al. 2004). There are, however, exceptions such as make amends and take charge (Baldwin and Kim 2009).

2.2.4 Idioms

Idioms (such as let the cat out of the bag, shoot the breeze, spill the beans, lose one’s cool, and kick the bucket) are a subset of multiword expressions which are well known for semantic idiosyncrasy, in addition to lexical and syntactic idiosyncrasy. Idioms are highly lexically fixed, meaning that the component words cannot be replaced by their synonyms (or closely related words) while preserving the idiomatic meaning. For example #shoot the wind or #hit the breeze do not maintain the idiomatic meaning of shoot the breeze.

Idioms are generally syntactically fixed. For example #the breeze was shot is usually not acceptable as a variation of the idiom shoot the breeze. The more idiomatic an expression is, the less flexible the expression is, and therefore the easier it is to automatically extract it from a corpus (Fazly et al. 2009).

Among English idioms, much of the work has been focused on verb+noun idioms, in which the noun is in the direct object position (Nunberg et al. 1994; Sag et al. 2002; Fazly et al. 2009; Baldwin and Kim 2009). This type of idiom is particularly
interesting because of cross-lingual occurrences, in addition to high semantic and lexical variability (Baldwin and Kim 2009).

2.2.5 Other Types

Beside all these well-known types of multiword expressions, there exist some less commonly studied MWE types, such as similes (e.g. *as brave as a lion, as busy as a bee*) (Moirón 2005). Additionally, different languages may have other types of MWEs with different characteristics, which need further studies. For example, Japanese has compound verbs, which are similar to LVCs with a verb as their second component (Uchiyama et al. 2005). Chapter 7 will discuss our proposed method to identify possible patterns of MWEs in different languages.

2.3 Computational studies of multiword expressions

Previous computational studies on MWEs are mainly focused on identification, extraction, or the semantics of MWEs. The remainder of this section discusses each of these challenges in addition to the most common proposed approaches.

It is worth mentioning that most recent work on MWEs can be divided into two categories: language/construction-specific and general-purpose. Much prior work on MWEs has been tailored to specific kinds of MWEs in particular languages (e.g. Fazly et al. (2009) on English verb+noun combinations, Tsvetkov and Wintner (2010) on Hebrew MWEs, and Salehi et al. (2012) on Persian LVCs). There has, however, been
recent interest in approaches to MWEs that are more broadly applicable to a wider range of languages and MWE types (Brooke et al. 2014; Schneider et al. 2014a). This thesis is in this same direction, investigating language independent approaches.

### 2.3.1 Identification and Extraction of MWEs

Identification of MWEs is a token-based task that locates each token usage (i.e., instance) of a given MWE in text. On the other hand, extraction (a.k.a discovery) of MWEs refers to deciding which combinations of words can be used as an MWE. For example, *red carpet* can be used both as an MWE or a literal combination. In an identification task, we decide which usages are MWEs and which are not, while in an extraction task, we conclude that *red carpet* can be used as an MWE. That is, extraction is used for lexicon development. These two tasks are highly related to each other, in that identification generally requires knowledge of potential MWEs, while extraction methods should generally be able to identify at least one usage of a word combination as an MWE.

Several approaches have been proposed to extract MWEs from a corpus. Some of these approaches make use of lexical information in addition to the output of POS taggers and parsers (Baldwin and Villavicencio 2002; McCarthy et al. 2003; Lapata and Lascarides 2003; Baldwin 2005; Kim and Baldwin 2010). For example, by finding prepositional particles using a POS tagger, and the nearest verb to the left, a VPC can be extracted/identified (Baldwin and Villavicencio 2002; Baldwin 2005). Such approaches, however, heavily depend on the accuracy of POS tagger or parser, and are not readily able to distinguish between the literal and MWE usages
of a combination.

The second group of extraction approaches use lexical association measures such as point-wise mutual information to measure the statistical idiosyncrasy and degree of association between words (Smadja 1993; Evert 2005; Pecina 2008), or the degree of lexical substitutability (Lin 1999; Pearce 2001) (see Section 2.4.1).

There are also several studies on the identification of MWE instances. Probably the first approach that comes to mind is to use word sense disambiguation (WSD) techniques, which uses the context of a word to distinguish what sense is used (Agirre and Edmonds 2007; Navigli 2009). In the task of MWE identification, WSD techniques are used to distinguish between idiomatic and literal usages of a combination of words (Katz and Giesbrecht 2006; Hashimoto and Kawahara 2008; Li and Sporleder 2010; Muzny and Zettlemoyer 2013). For example, the literal usage of red carpet tends to be used with words such as floor, home and furniture, while the idiomatic usage is used with terms such as celebrity, movie and ceremony.

Another approach is to identify MWE instances by considering the syntactic configuration (a.k.a. canonical forms). In these approaches, the canonical forms act like a template, and if a combination does not match that template, it is a good sign that it is not an MWE instance. For example, #kick a bucket, #the bucket was kicked and #kick the buckets are not canonical forms of the MWE kick the bucket. The main challenge of these approaches is obtaining the knowledge for the templates. Some approaches use hand-built rules (Li et al. 2003; Hashimoto et al. 2006), while others use supervised and unsupervised techniques to learn the canonical configurations (Fazly 2007; Fazly et al. 2009).
2.3.2 Semantic related studies

The semantic idiosyncrasy/compositionality of MWEs has been the focus of many studies. Predicting the compositionality degree of MWEs is a common research task, and will be discussed in detail in Section 2.4.

Semantic interpretation is another style of approach related to the semantics of compositional MWEs (Lapata 2002; Kim and Baldwin 2005). As mentioned in Section 2.1.3, MWEs such as NCs carry an implicit meaning in addition to the meaning of the components. Table 2.3 shows various interpretations of English NCs adopted from Kim and Baldwin (2005). Semantic interpretation has been widely studied for English NCs (Lapata 2002; Kim and Baldwin 2005; Girju 2007). In addition, the NLP community has had several competitions on identifying the relations between the components of English noun compounds (Girju et al. 2007; Hendrickx et al. 2009; Hendrickx et al. 2013).

The approaches to semantic interpretation mainly use resources such as WordNet (Kim and Baldwin 2005), paraphrasing (e.g., by considering bus driver vs. drive the bus) (Lapata 2002) or cross-lingual evidence (Girju 2007).

While semantic interpretation is more concerned with compositional MWEs, there has been some research on the semantics of non-compositional MWEs, in particular, figurative MWEs and similes such as sound like a prophet (Veale and Hao 2007; Li et al. 2012; Qadir et al. 2016). The goal of these methods is to determine the sentiment of a sentence containing similes, or to automatically associate similes such as sound like a prophet with characteristics such as wise, insightful, prescient and enlightened. The proposed approaches mainly use the semantics of the components,
Chapter 2: Literature Review

32

Table 2.3: The semantic relation between the components of English noun compounds (N1 is the modifier and N2 is the head), adopted from Kim and Baldwin (2005).

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>N2 is performed by N1</td>
<td>student protest, band concert</td>
</tr>
<tr>
<td>BENEFICIARY</td>
<td>N1 benefits from N2</td>
<td>student price, charitable compound</td>
</tr>
<tr>
<td>CAUSE</td>
<td>N1 causes N2</td>
<td>printer tray, flood water</td>
</tr>
<tr>
<td>CONTAINER</td>
<td>N1 contains N2</td>
<td>exam anxiety</td>
</tr>
<tr>
<td>CONTENT</td>
<td>N1 is contained in N2</td>
<td>paper tray, eviction notice</td>
</tr>
<tr>
<td>DESTINATION</td>
<td>N1 is destination of N2</td>
<td>game bus, exit route</td>
</tr>
<tr>
<td>EQUATIVE</td>
<td>N1 is also head</td>
<td>composer arranger, player coach</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>N1 is used in N2</td>
<td>electron microscope, diesel engine</td>
</tr>
<tr>
<td>LOCATED</td>
<td>N1 is located at N2</td>
<td>building site, home town</td>
</tr>
<tr>
<td>LOCATION</td>
<td>N1 is the location of N2</td>
<td>lab printer, desert storm</td>
</tr>
<tr>
<td>MATERIAL</td>
<td>N2 is made of N1</td>
<td>carbon deposit, gingerbread man</td>
</tr>
<tr>
<td>OBJECT</td>
<td>N1 is acted on by N2</td>
<td>engine repair, horse doctor</td>
</tr>
<tr>
<td>POSSESSOR</td>
<td>N1 has N2</td>
<td>student loan, company car</td>
</tr>
<tr>
<td>PRODUCT</td>
<td>N1 is a product of N2</td>
<td>automobile factory, light bulb</td>
</tr>
<tr>
<td>PROPERTY</td>
<td>N2 is N1</td>
<td>elephant seal</td>
</tr>
<tr>
<td>PURPOSE</td>
<td>N2 is meant for N1</td>
<td>concert hall, soup pot</td>
</tr>
<tr>
<td>RESULT</td>
<td>N1 is a result of N2</td>
<td>storm cloud, cold virus</td>
</tr>
<tr>
<td>SOURCE</td>
<td>N1 is the source of N2</td>
<td>chest pain, north wind</td>
</tr>
<tr>
<td>TIME</td>
<td>N1 is the time of N2</td>
<td>winter semester, morning class</td>
</tr>
<tr>
<td>TOPIC</td>
<td>N2 is concerned with N1</td>
<td>computer expert, safety standard</td>
</tr>
</tbody>
</table>

as well as dependency relations and paraphrasing techniques.

2.4 Compositionality of MWEs

Computational research on MWEs has mainly focused on the acquisition of various properties of MWEs such as lexical, semantic idiosyncrasy, syntactic or lexical fixedness (Krenn and Evert 2001; Fazly and Stevenson 2006; Cook et al. 2007; Reddy et al. 2011). Semantic idiosyncrasy has been of particular interest to NLP researchers with research on binary compositional/non-compositional MWE classification (Lin 1999;
Table 2.4: Compositionality degree of English noun compounds with their components and compound mean.

<table>
<thead>
<tr>
<th>Example</th>
<th>Component 1</th>
<th>Component 2</th>
<th>Noun Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>snail mail</em></td>
<td>0.60</td>
<td>4.59</td>
<td>1.31</td>
</tr>
<tr>
<td><em>guilt trip</em></td>
<td>4.71</td>
<td>0.86</td>
<td>2.19</td>
</tr>
<tr>
<td><em>radio station</em></td>
<td>4.66</td>
<td>4.34</td>
<td>4.47</td>
</tr>
</tbody>
</table>

Table 2.5: Examples of English verb particle constructions with binary annotations for each component.

<table>
<thead>
<tr>
<th>Example</th>
<th>Verb</th>
<th>Particle</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>come back</em></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><em>bring up</em></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><em>figure out</em></td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Baldwin *et al.* 2003; Bannard 2006), or a three-way compositional/semi-compositional/non-compositional distinction (Fazly and Stevenson 2007). There has also been research to suggest that MWEs span the entire continuum from full compositionality to full non-compositionality (McCarthy *et al.* 2003; Reddy *et al.* 2011). Table 2.4 shows some examples from Reddy *et al.* (2011), in which the annotators score the compositionality degree of the noun compounds and their components between zero and five. Table 2.5 shows the binary classification from Bannard (2006).

In the literature, compositionality has been viewed as either compositionality of the whole MWE as one unit (McCarthy *et al.* 2003; Venkatapathy and Joshi 2005; Katz and Giesbrecht 2006; Biemann and Giesbrecht 2011; Farahmand *et al.* 2015), or compositionality relative to each component (Reddy *et al.* 2011; Hermann *et al.* 2012; Schulte im Walde *et al.* 2013). There have also been studies which focus only on one component of the MWE. For example, Korkontzelos and Manandhar (2009) induce the most probable sense of the MWE first. Then, they measure the semantic similarity
between the MWE and its semantic head. This approach of considering only the head component has been shown to be quite accurate for English verb particle constructions (Bannard et al. 2003). However, it might not be always the case. For example, as shown in Reddy et al. (2011), the compositionality of the first noun (the modifier) has more impact than the second noun (the head) for English noun compounds.

Previous studies which have considered MWE compositionality have focused on either the identification of non-compositional MWE token instances (Kim and Baldwin 2007; Fazly et al. 2009; Fothergill and Baldwin 2011; Muzny and Zettlemoyer 2013), or the prediction of the compositionality of MWE types, independent of usage (Schone and Jurafsky 2001; Bannard et al. 2003; Reddy et al. 2011). The identification of non-compositional MWE tokens is an important task when a word combination such as kick the bucket or saw logs is ambiguous between a compositional (generally non-MWE) and non-compositional MWE usage (“die” and “sleep”, respectively). Approaches have ranged from the unsupervised learning of type-level preferences (Fazly et al. 2009) to supervised methods specific to particular MWE constructions (Kim and Baldwin 2007) or applicable across multiple constructions using features similar to those used in all-words word sense disambiguation (Fothergill and Baldwin 2011; Muzny and Zettlemoyer 2013).

In another point of view, a lot of work has been on specific types of MWE in specific languages. In English, studies have been done specifically on VPCs (McCarthy et al. 2003; Bannard et al. 2003), verb+noun MWEs (Venkatapathy and Joshi 2005; McCarthy et al. 2007; Fazly et al. 2009), noun compounds (Reddy et al. 2011), and adjective+noun compounds (Vecchi et al. 2011). There have also been studies focus-
Chapter 2: Literature Review

Chapter 2: Literature Review

35

ing on a specific language other than English, such as Arabic (Saif et al. 2013) and
German (Schulte im Walde et al. 2013). This thesis investigates language independent
approaches applicable to any type of MWE in any language.

Most previous work on measuring the compositionality degree of MWEs uses
approaches based on lexical substitutability, syntactic fixedness, semantic similarity
or cross-lingual information. The rest of this section reviews each of these approaches
in turn.

2.4.1 Lexical substitutability

One of the first studies on MWE compositionality was Lin (1999), who observed
that the distribution of non-compositional MWEs (e.g. shoot the breeze) differs sig-
nificantly from the distribution of expressions formed by substituting one of the com-
ponents with a semantically similar word (e.g. shoot the wind).

In order to extract candidates, the MINIPAR (Lin 1993) parser was used to ex-
tract the dependency triples of two words (head (H) and modifier (M)) and the
grammatical relation (R) between them. Mutual information (MI) is computed for
each triple of < H, R, M >, as follows:

\[
\text{mutualInfo}(H, R, M) = \log \frac{P(H, R, M)}{P(M|R)P(H|R)P(R)}
\]

(2.1)

where, \(|H, R, M|\) shows the frequency of the triple, and * is a wild card. For each
collocation, the mutual information of the collocation and its variants are measured
and compared. Adopted from Lin (1999), Table 2.6 shows the mutual information of
Table 2.6: Mutual information of *red carpet* and *economic impact* and their variants

<table>
<thead>
<tr>
<th>head&amp;modifier</th>
<th>freq</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>red tape</td>
<td>259</td>
<td>5.87</td>
</tr>
<tr>
<td>yellow tape</td>
<td>12</td>
<td>3.75</td>
</tr>
<tr>
<td>orange tape</td>
<td>2</td>
<td>2.64</td>
</tr>
<tr>
<td>black tape</td>
<td>9</td>
<td>1.07</td>
</tr>
<tr>
<td>economic impact</td>
<td>171</td>
<td>1.85</td>
</tr>
<tr>
<td>financial impact</td>
<td>127</td>
<td>1.72</td>
</tr>
<tr>
<td>political impact</td>
<td>46</td>
<td>0.50</td>
</tr>
<tr>
<td>social impact</td>
<td>15</td>
<td>0.94</td>
</tr>
<tr>
<td>budgetary impact</td>
<td>8</td>
<td>3.20</td>
</tr>
<tr>
<td>ecological impact</td>
<td>4</td>
<td>2.59</td>
</tr>
<tr>
<td>economic effect</td>
<td>84</td>
<td>0.70</td>
</tr>
<tr>
<td>economic implication</td>
<td>17</td>
<td>0.80</td>
</tr>
<tr>
<td>economic consequence</td>
<td>59</td>
<td>1.88</td>
</tr>
<tr>
<td>economic significance</td>
<td>10</td>
<td>0.84</td>
</tr>
<tr>
<td>economic fallout</td>
<td>7</td>
<td>1.66</td>
</tr>
<tr>
<td>economic repercussion</td>
<td>7</td>
<td>1.84</td>
</tr>
<tr>
<td>economic potential</td>
<td>27</td>
<td>1.24</td>
</tr>
<tr>
<td>economic ramification</td>
<td>8</td>
<td>2.19</td>
</tr>
<tr>
<td>economic risk</td>
<td>17</td>
<td>-0.33</td>
</tr>
</tbody>
</table>

two collocations, *red carpet* and *economic impact*, and their variants. The variants of the compositional MWE, *economic impact* have similar MI, while variants of *red carpet* are very infrequent with very different MI to the collocation. A collocation is called non-compositional if for none of the substituted variations of that collocation, the collocation’s and the variant’s mutual information have an overlap in the 95% confidence interval (using Z-score).

In order to automatically extract semantically similar substitutions for components, Lin (1999) used a variant of point-wise mutual information measure introduced in Lin (1998). For a more accurate substitution with synonyms, Pearce (2001) suggested using synsets from WordNet (Miller 1995). Similar to Lin (1999), the substitutability of a collocation was determined by using the frequency of that collocation, and its variants.
Wurmbrand (2000) showed that literal particles in German particle verbs can be replaced with semantically similar particles. Cook and Stevenson (2006) went one step further and extended this idea to the prediction of the particle semantics in English VPCs. Their experiments are focused on up, as this particle is highly frequent and can bear a wide range of meanings (vertical up such as jump up, goal oriented up such as kiss up, completive up such as clean up, and reflexive up such as curl up).

They hypothesize that the patterns of how a specific verb component is combined with different particles as a VPC can indicate the semantics of a particular particle (in their study up) when combined with that specific verb.

Bannard et al. (2003) follows Lin (1999) and proposed four different approaches to determining the compositionality of English VPCs. In their first approach, following Lin (1998), they made a thesaurus to substitute the components with semantically similar words. They investigated making a thesaurus automatically using latent semantic indexing (LSI) (Deerwester et al. 1990) in the second approach. They argue that the substitutability approach tends to fall down in cases of high statistical idiosyncrasy (or “institutionalization”): consider frying pan which is compositional but not productive, with less familiar phrases produced through word-substitution such as sauteing pan or frying plate. They propose a third approach to overcome the sensitivity of the substitution approach to institutionalization. In this approach, they use distributional similarity to measure the semantic similarity between the MWEs and their word-substituted versions. However, as the problem resists, they finally propose a fourth approach, in which they use measure the semantic similarity between each VPC and its components, which will be discussed in Section 2.4.3.
2.4.2 Syntactic Fixedness

Some research has investigated the syntactic properties of MWEs to detect their compositionality (Bolinger 1976; Cook and Stevenson 2006; Fazly et al. 2009). The assumption behind these methods is that non-compositional MWEs are more syntactically fixed than compositional MWEs. For example, *make a decision* can be passivised, but *shoot the breeze* cannot, which shows that the former is more semantically compositional.

The patterns investigated by Fazly et al. (2009) are passivization, determiner type and pluralization in English Verb-Noun MWEs. The probability of each pattern is calculated as follows:

\[
P(pt) = \frac{\sum_{v_i \in V} \sum_{n_j \in N} f(v_i, n_j, pt)}{\sum_{v_i \in V} \sum_{n_j \in N} \sum_{pt_k \in p} f(v_i, n_j, pt_k)} = \frac{f(\ast, \ast, pt)}{f(\ast, \ast, \ast)}
\]

where, \( V \) and \( N \) are the set of transitive verbs and nouns appearing as direct objects in the corpus, respectively, and \( p \) is the set of selected patterns.

Then, the maximum likelihood of each pattern \((pt)\) of a given Verb(\(v\))-Noun(\(n\)) is calculated as follows:

\[
P(pt|v, n) = \frac{f(v, n, pt)}{\sum_{pt_k \in p} f(v, n, pt_k)} = \frac{f(v, n, pt)}{f(v, n, \ast)}
\]

Finally, they use Kullback Leibler (KL-) divergence to measure the syntactic fixedness, where higher values indicate a higher degree of syntactic fixedness of the given Verb(\(v\))-Noun(\(n\)):
\[ \text{Fixedness}_{\text{sym}}(v, n) = D(P(pt|v, n)||P(pt)) \]
\[ = \sum_{pt_k \in P} P(pt_k|v, n) \log \frac{P(pt_k|v, n)}{P(pt_k)} \]

Similar to English Verb-Noun MWEs, Bolinger (1976) notes the syntactic fixedness of non-compositional English VPC, in that non-compositional particles in English VPCs tend to be used in joined forms. Cook and Stevenson (2006) applied this finding and used fixedness as one feature to determine the semantics of particles in English VPCs.

One serious problem with syntax-based methods is their lack of generalization: as discussed in Section 2.2, each type of MWE has its own characteristics, and these characteristics differ from one language to another. Moreover, some MWEs (such as noun compounds) are not generally flexible syntactically, no matter whether they are compositional or non-compositional (Reddy et al. 2011).

### 2.4.3 Semantic Similarity

Considering that compositionality degree is the relative semantic similarity between the MWE and its components, a great body of recent work on MWEs has focused on techniques to measure this semantic similarity (Schone and Jurafsky 2001; Bannard et al. 2003; McCarthy et al. 2003; Bannard 2006; Piao et al. 2006; Giesbrecht 2009; Reddy et al. 2011; Schulte im Walde et al. 2013; Muzny and Zettlemoyer 2013). To measure semantic similarity, various approaches have been proposed using either lexical resources or distributional semantics. This section explains these two approaches in more detail.
Lexical Resources

In this approach, semantic similarity between an MWE and its components is measured using a man-made lexical resource.

Piao et al. (2006) use the Lancaster English semantic lexicon. This lexicon provides 21 semantic tags, which are further divided into 232 sub-categories (Piao et al. 2005). For example, the provided semantic tags for mass are $N5$, $N3.5$, $S9$, $S5$ and $B2$. These tags show the semantic categories of QUANTITIES, MEASUREMENT: WEIGHT, RELIGION AND SUPERNATURAL, GROUPS AND AFFILIATION and HEALTH AND DISEASE. In order to detect the compositionality of a MWE, they proposed to compare its semantic tags with the semantic tags of its components.

Muzny and Zettlemoyer (2013) use definitions provided by Wiktionary to: (1) detect which MWE sense is idiomatic; and (2) to identify the idiomatic usage of MWE in a context. Their approach is based on the Lesk (1986) algorithm to disambiguate word sense based on the overlap between the definitions of that word in dictionary and its neighboring words in a sentence. More information on Wiktionary is provided in Section 3.2.3 and our approach on using Wiktionary is explained in Chapter 5.

In contrast with statistical approaches, which can be affected by noise, lexical resource based approaches have the advantage of being based on expert knowledge. However, such approaches suffer from not being language independent due to relying on a language-specific resource. Also, if such resources are not updated regularly, due to dynamic behavior of human language, they might cover outdated senses of words, while they lack new or domain-specific words of the language.
Distributional Similarity

The first part of this section describes distributional similarity (DS). The second part explains how distributional similarity can be used to predict the compositionality of MWEs.

**What is Distributional Similarity**  Distributional similarity is based on the following famous insights on semantics in Linguistics:

*Similar words tend to have similar contexts* (Harris 1954)

and

*You shall know a word by the company it keeps* (Firth 1957:p.11).

According to the above insights, (1) the meaning of a word is predictable from its neighboring words, and (2) semantically similar words share similar neighboring words (i.e., contexts). For example, the unknown *xyz* in *He ate all xyz on his plate*, seems be the name of some kind of food according to its neighboring words. In the absence of lexical resources, we can use these insights to model the meaning of words. In this model, the words are represented geometrically as vectors in a high dimensional semantic space, where the dimensions are based on statistical analysis of the neighboring (context) words that co-occur with the target words. Less distance between the word vectors implies greater semantic similarity between those words. Each of the mentioned terms and parameters will be fully described in the remainder of this section. Figure 2.2 demonstrates the hypothetical example of *soup, sun* and *stew* semantic vectors in a 2D semantic space. The example shows that *stew* and *soup* are semantically similar while they are different from *sun*. 
The strength of distributional similarity is in its parameters, which can be adjusted based on the task. We introduce these parameters in the following.

**Context type:** The context of a token can simply be considered to be the neighbouring words of its token-level occurrences. However, to include more information about the co-occurring words and to reduce the effect of polysemy, we can also consider other information from the co-occurring words such as part-of-speech tags, dependency relations, and distance from the examined token. We can also decide whether to ignore some of context words such as the stopwords.\(^5\)

**Context window size:** We can also tune the number of neighboring words to consider as the context of each token occurrence of the target word. The context size can be a sentence, a paragraph or even a document. It can also be based on a context window, which can be symmetric or asymmetric; a symmetric window with size 3 is

---

5Stopwords are frequent words which carry little or no standalone semantic content, such as determiners.
made up of the 3 left and 3 right terms of the examined token, while an asymmetric window considers only the left or the right terms. For example, the neighboring words of *Shiraz* with a symmetric context window size of 3, positional, ignoring word-order, direction and all punctuations, are underlined below:

The city of *Shiraz* is most likely more than 4000 years old *Shiraz* is known as the city of poets literature wine and flowers Many *Iranians* consider *Shiraz* to be the city of gardens due to the many gardens and fruit trees that can be seen in the city

The values of semantic vectors: The values of semantic vectors show the degree of association between the target word and context words. This association can be measured using raw co-occurrence frequency or a binary co-occurrence value, or can be weighted with point wise mutual information, tf-idf, log likelihood ratio, t-score, chi-square, etc. Such weights assign values to the context words based on how much they indicate the semantic of the target word (Manning and Schütze (1999:Ch.5), Curran (2003))

Similarity of the context vectors: The semantic similarity between two context vectors can be measured using metrics such as cosine similarity, Hamming distance (the number of positions at which the two vectors are different), Manhattan distance (the sum of the absolute differences of the Cartesian coordinates of two points), and Minkowski distance (a generalized version of Manhattan and Euclidean distance).

The most common similarity measure in distributional similarity, which we also use in our experiments, is cosine similarity. Given the semantic vectors of *A* and *B* the cosine similarity is measured as follows:
\[
\text{similarity} = \cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

**Dimensionality reduction:** There are several approaches to reduce the number of dimensions in order to:

1. represent data in a lower space while preserving the necessary information.
2. reduce data sparseness by removing less-informative context words.
3. improve the model by eliminating noisy context words.

One fast and easy approach is to consider only the 1000 most frequent words in a corpus, and consider them to be content-bearing words (Schütze 1997). Schütze (1992) used singular value decomposition (SVD) to reduce the size of context vectors in a way such that the structural similarity between the columns is preserved. Latent semantic analysis (LSA) is another SVD-based approach to measuring the semantic similarity between words. This technique (also known as latent semantic indexing (LSI)) was introduced by Deerwester et al. (1990). After reducing the dimensions, the ultimate vectorial representations can be used to measure the similarity between the words, similarly to the method discussed for distributional similarity.

**Measuring compositionality using distributional similarity**

Given that the meaning of a word is predictable from its context of use, we can use it to predict the compositionality degree of MWEs (Schone and Jurafsky 2001; Bannard et al. 2003; McCarthy et al. 2003; Reddy et al. 2011). In order to predict the
compositionality of a given MWE using distributional similarity, the semantic vector of the MWE is compared with that of its components. A MWE is compositional iff its semantic vector is similar to its components’ semantic vectors. Table 2.7 shows an example of a vector space model for silver screen and its components, using a small subset of an English Wikipedia dump as the background corpus.

Based on the semantic vectors the similarity between silver screen and screen is greater than the similarity between silver screen and silver. This suggests that silver screen is a semi-compositional noun compound, which contains the meaning of screen more than the meaning of silver.

**Table 2.7: An example of semantic vectors for silver screen and its components.**

<table>
<thead>
<tr>
<th></th>
<th>movie</th>
<th>gold</th>
<th>games</th>
<th>film</th>
<th>medal</th>
</tr>
</thead>
<tbody>
<tr>
<td>silver</td>
<td>2</td>
<td>249</td>
<td>67</td>
<td>44</td>
<td>349</td>
</tr>
<tr>
<td>screen</td>
<td>230</td>
<td>2</td>
<td>8</td>
<td>103</td>
<td>1</td>
</tr>
<tr>
<td>silver screen</td>
<td>44</td>
<td>0</td>
<td>1</td>
<td>32</td>
<td>0</td>
</tr>
</tbody>
</table>

**Previous studies on Using Distributional Similarity**

Distributional similarity (DS) is a general-purpose type-level approach, which is applicable to any type of MWE in any language, as long as the tokens can be identified in a large enough corpus.

One of the early studies on using DS to measure the compositionality of MWEs was by Schone and Jurafsky (2001). Their original objective was to identify MWEs with the help of several association measures. However, they hypothesized that non-compositionality is a factor which can find MWEs better than association measures,
and used LSA to measure the compositionality degree.

In the end, they were not satisfied with the results and concluded that LSA is not a suitable approach for this task. However, the way they evaluated their work was slightly problematic. Two idiomatic dictionaries were used for evaluation, and any MWE which was not in these dictionaries was considered to be a compositional MWE. This way of evaluating is problematic because firstly, some idioms are domain specific and might not exist in every dictionary. Secondly, a lot of MWEs are not completely compositional or non-compositional, and people may have different ideas about the degree of their compositionality, which will impact on the dictionary coverage. Thirdly, as mentioned in Section 1.1, MWEs can be compositional but be lexicalised in dictionaries because of statistical or syntactic idiosyncrasy. Due to this slightly problematic evaluation, some correctly identified idioms might not be found in dictionaries, which leads to an underestimate of the utility of the approach.

Bannard et al. (2003) used LSA to measure the semantic similarity of a VPC with each of its components. In contrast to Schone and Jurafsky (2001), they first annotated the dataset using a number of native English speakers to avoid their problematic evaluation. They compared LSA with the lexical substitutability approach of Lin (1999) and showed LSA to be superior. Unlike Schone and Jurafsky (2001), by changing the evaluation setup, they showed that LSA can be a good technique for measuring the degree of compositionality. However, the experimental dataset used in Bannard et al. (2003) is small and therefore might not be generalizable. They extended their dataset to 160 VPCs in Bannard (2006), which will be used in our experiments in Chapter 4 and Chapter 5. With a similar approach, Katz and Giesbrecht
(2006) applied LSA to identify non-compositional German Preposition-Noun-Verb MWEs by comparing the representation vector of an MWE with its compositional meaning (the sum of the component vectors).

McCarthy et al. (2003) use DS in a different manner to predict the compositionality of VPCs. They find the \( N \) most semantically similar terms to the VPC and each of the components, using distributional similarity. Then, they used a number of measures to model the similarity based on the overlap between the semantically similar terms to the VPC and its components.

Reddy et al. (2011) use distributional similarity to measure how literal each component is within the MWE. In this study, the semantic similarity between the MWE and each of the components (\( s_1 \) and \( s_2 \)) is measured and the compositionality score of the whole MWE (\( s_3 \)) is computed using one of the following:

- **ADD**: \( s_3 = a.s_1 + b.s_2 \)
- **MULT**: \( s_3 = a.s_1.s_2 \)
- **COMB**: \( s_3 = a.s_1 + b.s_2 + c.s_1.s_2 \)
- **WORD1**: \( s_3 = a.s_1 \)
- **WORD2**: \( s_3 = a.s_2 \)

where, \( a, b \) and \( c \) are coefficients and their values are set according to a training set. Among the above operators, Reddy et al. (2011) showed that **ADD** performs the best in predicting the compositionality degree of MWEs (English noun compounds in their study).
Following studies on modeling the meaning of phrases based on the vector representations of individual words (Widdows 2008; Mitchell and Lapata 2008; Mitchell and Lapata 2010), Reddy et al. (2011) investigate composing a vector based on the components vectors using vector addition/element-wise multiplication. The composed vector is then compared with the vector representation of the MWE: the higher the similarity, the higher the compositionality score. They conclude that the second approach, using vector addition, performs slightly better than the first approach.

**Shortcomings of Distributional Similarity**

While distributional similarity based approaches have been shown to perform quite well despite being unsupervised, there are several shortcomings, which we discuss below.

First, some combinations of words can be both non-compositional and literal combinations (such as *red carpet*), meaning token-based identification is needed to find the required MWEs (Korkontzelos and Manandhar 2009). Failure to do so will result in noisy vectors which represent the meaning of both literal and idiomatic combinations, or vectors which are biased towards the more popular sense.

Second, identifying token instances of MWEs is not always easy, especially when the component words do not occur sequentially. For example consider *put on* in the following example:

(2.10) **put** your jacket **on**.

(2.11) **put** your jacket **on** the chair.
In (2.10), *put on* is an MWE while in (2.11), it is a simple verb with a prepositional phrase and not an instance of an MWE. Furthermore, if we adopt a conservative identification method, the number of token occurrences will be limited and the distributional scores may not be reliable.

And third, things are complicated by the polysemy or homonymy of components. For example *light* in *traffic light* can occur with different meanings in different contexts, such as *quantum theory, optics, lamps* and *spiritual theory* (Reddy *et al.* 2011). It is important to determine which sense of the component is used in measuring semantic similarity.

Fourth, for morphologically-rich languages, it can be difficult to predict the different word forms a given MWE type will occur across, posing a challenge for language-independent preprocessing. This becomes a serious challenge when the goal is to propose a general approach applicable to a wide range of languages, as well as when proposing cross-lingual approaches.

And fifth, low frequency expressions are usually discarded, as distributional semantic models generally do not capture their meaning well (Schone and Jurafsky 2001; Cook *et al.* 2007). Unfortunately, large corpora are not available for all languages which, for the same reason, weakens distributional semantic models.

**Word embeddings**

There has been a recent surge of interest in learning distributed representations of word meaning, in the form of word embeddings. Such methods are originally derived from language modeling techniques, and map the words to low-dimension vectors of
real numbers. Different methods are proposed to generate these mappings such as
methods based on neural network (Bengio et al. 2003), principle component analysis
(Lebret and Collobert 2014), matrix factorization (Pennington et al. 2014), and log-
linear models (Mikolov et al. 2013a; Mikolov et al. 2013b).

One of the most famous approaches was proposed by Mikolov et al. (2013a) and
Mikolov et al. (2013b), using the WORD2VEC package.\footnote{https://code.google.com/p/word2vec/} WORD2VEC uses a log-linear
model inspired by the original approach of Mikolov et al. (2010), in two forms: (1) a
continuous bag-of-words (“CBOW”) model, whereby all words in a context window
are averaged in a single projection layer; and (2) a continuous skip-gram model,
whereby a given word in context is projected onto a projection layer, and used to

---

**Figure 2.3: CBOW vs. skip-gram model**

![Diagram of CBOW vs. skip-gram model]
predict its immediate context (preceding and following words) (Figure 2.3).

(Collobert and Weston 2008) proposed the first attempt to utilize word embeddings in an application. Recently, word embeddings have been applied to produce improved methods for a wide range of NLP tasks, including identifying various syntactic and semantic relations (Mikolov et al. 2013a), dependency parsing (Bansal et al. 2014), named-entity recognition (Passos et al. 2014), and translation (Zou et al. 2013). We will investigate using word embeddings to predict the compositionality of multiword expressions in Section 4.3.

\subsection*{2.4.4 Cross-lingual Approaches}

Methods have also been proposed that use parallel corpora to detect the degree of compositionality (Melamed 1997; Moirón and Tiedemann 2006; Caseli et al. 2010; Salehi et al. 2012).

By parallel, we mean each sentence in the source language is translated to a sentence in the target language. The idea behind using parallel corpora is that they have been found to be more informative than monolingual corpora (Dagan et al. 1991), especially when dealing with ambiguity or polysemy. Using the IBM alignment models (Brown et al. 1993), the words of a sentence in a source language can be aligned to the corresponding words in the target language using translation models. The word-alignments between the source and target language sentences have been used to analyze the alignment patterns for MWEs (e.g. if the MWE is always aligned as a single “phrase”, then it is a strong indicator of non-compositionality).

Caseli et al. (2010) consider non-compositional MWEs to be those candidates that
align to the same target language unit, without decomposition into word alignments (Figure 2.4). Melamed (1997) suggests using mutual information to investigate how well the translation model predicts the distribution of words in the target text given the distribution of words in the source text. Moirón and Tiedemann (2006) show that entropy is a good indicator of compositionality, because word alignment models are often confused by non-compositional MWEs. In other words, low entropy shows highly predictable alignments which corresponds to more compositional expressions, and on the other hand, high entropy corresponds to non-compositionality. However, this assumption is not always valid, especially when dealing with high-frequency non-compositional MWEs. Salehi et al. (2012) tried to solve this problem with high frequency MWEs by using word alignment in both directions.\footnote{The IBM models (Brown et al. 1993), e.g., are not bidirectional, which means that the alignments are affected by the alignment direction.} They computed backward and forward entropy to try to remedy the problem with especially high-frequency phrases. However, their assumptions were not easily generalisable across languages,
e.g., they assume that the relative frequency of a specific type of MWE (light verb constructions) in Persian is much greater than in English.

Although methods using bilingual corpora are intuitively appealing, they have a number of drawbacks. The first and most important problem is data: they need large-scale parallel bilingual corpora, which are available for relatively few language pairs. Second, since they use statistical measures, they are not suitable for measuring the compositionality of MWEs with low frequency. Third, the selection of the target language can be quite important. Most approaches that use bilingual corpora assume that idiomatic MWEs are not translated word-for-word in the target language. Therefore, if the target language is very similar to the source language, some idiomatic MWEs might be translated word-by-word, which contradicts the basic assumption. And finally, most experiments have been carried out on English paired with other European languages, and it is not clear whether the results translate across to other language pairs.

A number of studies use monolingual features alongside cross-lingual features, when a large bilingual corpus is not available. Tsvetkov and Wintner (2010) used monolingual corpus alongside a small English-Hebrew bilingual corpora to extract Hebrew MWEs. They use both resources to overcome the difficulties of a morphologically rich language as well as the problems of using a small bilingual corpus. Girju (2007) attempted to interpret the meaning of noun compounds using both monolingual features extracted from English and cross lingual features gathered from five Romance languages. She observed that adding cross-lingual features enhances the identification results.
To overcome the problem of finding an appropriate target language, Pichotta and DeNero (2013) proposed a token-based method for identifying English phrasal verbs using parallel corpora for 50 languages. They showed that the combination of information from multiple languages can enhance the identification performance. This finding lends weight to our hypothesis in Chapter 4 that using translation data and distributional similarity from each of a range of target languages, can improve compositionality prediction. Having said that, the general applicability of their proposed method is questionable – there are many parallel corpora involving English, but for other languages, this tends not to be the case. In Section 4.2, we propose an approach which uses multiple monolingual corpora in 52 languages, which are freely available from Wikipedia dumps (see Section 3.2.2).

2.5 Applications

According to Sag et al. (2002), MWEs present a significant problem to NLP applications. An explicit handling of MWEs has been shown to be useful in NLP applications including information retrieval (Acosta et al. 2011) and machine translation (Tsvetkov and Wintner 2010; Weller et al. 2014; Carpuat and Diab 2010; Venkatapathy and Joshi 2006). Other studies have shown the positive influence of considering MWEs for semantic tagging (Piao et al. 2003) and deep parsing (Baldwin et al. 2004b).

Acosta et al. (2011) showed that by considering non-compositional MWEs as a single unit, the effectiveness of document ranking in an information retrieval system improves. For example, while searching for documents related to *ivory tower*, we are
almost certainly not interested in documents relating to elephant tusks.

A large portion of recent studies on considering MWEs in NLP applications, investigate how to integrate MWEs into machine translation systems. MWEs have been shown to be hard to align in statistical machine translation (Lambert and Banchs 2005). Tanaka and Baldwin (2003) claim that two problematic aspects of translating noun compounds are idiomaticity and overgeneration (i.e., lexical idiosyncrasy). Carpuat and Diab (2010) showed that even a large bilingual corpus cannot capture all the necessary information to translate MWEs, and that in adding the facility to model the compositionality of MWEs into their system, they could improve translation quality.

Lambert and Banchs (2005) identified the MWEs in source and target sentences, and their components were grouped as one unique token before feeding the bilingual corpora to the aligner. They showed that alignment improves through these techniques, however, the BLEU score, which reassures the translation quality, decreases.

Ren et al. (2009) proposed three approaches to enhance translation quality. First, they added bilingual MWEs (MWEs in the source language with their translation, which is also an MWE) to the training corpus. In the second approach, they added a binary feature to the Moses statistical machine translation system (Koehn et al. 2007). This binary feature indicates whether a bilingual MWE exists in the given sentence or not. In the third approach, bilingual MWEs are integrated into the constructed phrase table. They conclude that while all three approaches improve the translation quality, the second approach achieves the best result.

According to Ghoneim and Diab (2013), while machine translation systems cap-
ture n-grams, they do not explicitly capture MWEs. They propose three approaches to integrating MWEs into a phrase-based statistical machine translation system: static, dynamic and zone integration. In the static approach, the MWEs are identified before feeding the text into the machine translation system. The components are combined into a single token, so that the machine translation system will consider the MWEs as one single token. The dynamic approach is based on the Ren et al. (2009) method, in which they integrate MWEs to Moses by adding a binary feature that identifies the existence or nonexistence of MWEs. However, instead of using a binary feature, Ghoneim and Diab (2013)’s method counts the number of MWEs in a given sentence as one of the features. As a result, the machine translation system will try to produce the phrases in a way that MWEs are not broken and considered as a single unit at decoding time. In the zone approach, MWEs zone are shown with XML tags as below example:

invading <zone> iraqs kurdistan </zone> is <zone> no longer </zone> an easy task.

The decoder\textsuperscript{8} is forced to translate the words inside the zone first, before translating the rest of the words. Marking such zones is easily available in Koehn and Haddow (2009). According to the results, Ghoneim and Diab (2013) show that the more flexible a MWE is, the more dynamic the integration should be.

Weller et al. (2014) study the effect of splitting German compounds, based on compositionality degree, on machine translation. They investigated both aggressive splitting and splitting only compositional compounds. They observe that while both

\textsuperscript{8}The decoder is a component of the statistical machine translation model which looks for the best possible translation for the input by consulting the trained translation tables.
splitting approaches outperform the baseline for German noun compounds, none of the two approaches is superior to the other. However, they notice slight improvements in BLEU scores when the compositionality degree is considered for particle verbs. They also experiment on German noun compounds and conclude that noun compounds are easy targets. They anticipate that more complex structures will show different results. They also suggest that in machine translation, considering only monolingual compositionality of MWEs is not enough, and the translation behavior of the compounds should also be considered.

In Chapter 6, we will present the first attempt to integrate predicted compositionality scores of multiword expressions into automatic machine translation evaluation.

2.6 Summary

In this chapter, we provided a detailed overview of multiword expressions definition along with a description of their characteristics. We described multiple types of multiword expressions, and their lexical, syntactic and semantic characteristics. Then, we reviewed previous studies on determining the compositionality of MWEs. Finally, we discussed how integrating MWEs and their compositionality degree affects natural language processing applications. This thesis aims to propose approaches to MWEs that are broadly applicable to a wide range of languages and MWE types. Chapters 4 and 5 propose general approaches to determine and predict the compositionality of MWEs. Chapter 6 will discuss how to integrate compositionality degree of MWEs into a machine translation evaluation system. Finally, Chapter 7 will discuss a general method to identify possible MWE patterns in a given language, based on
other languages.
Chapter 3

Resources

This chapter presents the datasets, resources, and evaluation metrics used in this thesis.

3.1 Primary Datasets

We evaluate our proposed approaches on predicting the compositionality degree of MWEs in Chapter 4 with three pre-existing datasets containing the following MWE construction types: English noun compounds, English verb particle constructions and German noun compounds. These datasets were purposely chosen to include two different types of MWE (noun compounds and verb particle constructions), and two different languages (English and German). We also use the two English datasets in Chapter 5, where we propose a method to detect non-compositional components in MWEs.

The remainder of this section describes the three datasets.
Table 3.1: Examples from English noun compound (ENC) dataset, with their components and compound mean ± deviation scores

<table>
<thead>
<tr>
<th>Example</th>
<th>Component 1</th>
<th>Component 2</th>
<th>Noun Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>snail mail</td>
<td>0.60±0.80</td>
<td>4.59±1.10</td>
<td>1.31±1.02</td>
</tr>
<tr>
<td>guilt trip</td>
<td>4.71±0.59</td>
<td>0.86±0.94</td>
<td>2.19±1.16</td>
</tr>
<tr>
<td>radio station</td>
<td>4.66±0.96</td>
<td>4.34±0.80</td>
<td>4.47±0.72</td>
</tr>
</tbody>
</table>

3.1.1 English Noun Compounds (ENCs)

Our first dataset was constructed by Reddy et al. (2011) and contains 90 English (binary) noun compounds (ENCs), where the overall NC and each component word has been annotated for compositionality on an integer scale from 0 (non-compositional) to 5 (compositional). In order to avoid issues with polysemy, the annotators were presented with each NC in a sentential context. The average of the scores given by annotators was considered as the compositionality score for each ENC or its components. The authors tried to make the dataset balanced in its composition of compositional and non-compositional NCs, by manually adding noun compounds from Wikipedia, which they inspect are non-compositional. Based on a threshold of 2.5, the dataset consists of 43 (48%) compositional NCs, 46 (51%) NCs with a compositional usage of the first component, and 54 (60%) NCs with a compositional usage of the second component. Overall, the average of ENC compositionality scores is 2.66 ± 1.48. Table 3.1 shows three examples from the ENC dataset.

3.1.2 English Verb Particle Constructions (EVPCs)

In the second dataset, 160 English verb particle constructions (VPCs) were annotated for compositionality relative to each of the two component words (the verb
Table 3.2: Examples from the English verb particle construction (EVPC) dataset, with the annotations for each component

<table>
<thead>
<tr>
<th>Example</th>
<th>Verb</th>
<th>Particle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>come back</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>bring up</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>figure out</td>
<td>19</td>
<td>8</td>
</tr>
</tbody>
</table>

and the particle) (Bannard 2006). Each annotator was asked to annotate each of the verb and particle as yes, no or don’t know. Annotators choose yes if they believe the VPC instance implies its verb/particle component, and no otherwise. They also have the option of don’t know to choose if they cannot decide. When a compositionality degree is required, we compute the proportion of yes tags to get the continuous compositionality score. For example, the compositionality score of come in come back (Table 3.2) would be \(\frac{25}{25+2}\) or 0.93. The overall average of verb and particle continuous compositionality scores are 0.70 ± 0.23 and 0.46 ± 0.27, respectively. Based on the majority annotation, among the 160 VPCs, 122 (76%) are verb-compositional and 76 (48%) are particle-compositional. Table 3.2 shows three examples from the EVPC dataset.

This dataset, unlike the ENC dataset, does not include annotations for the compositionality of the whole VPC, and is also less balanced, containing more VPCs which are verb-compositional than verb-non-compositional. Since we do not have judgements for the compositionality of the full VPC in the EVPC dataset (we instead have separate judgements for the verb and particle) we use the verb compositionality score as the compositionality score for the full VPC, in line with Bannard et al. (2003) observation.
Table 3.3: Examples from the German noun compound (GNC) dataset, with their components and compound mean±deviation scores

### 3.1.3 German Noun Compounds (GNCs)

Our final dataset is made up of 246 German noun compounds (von der Heide and Borgwaldt 2009; Schulte im Walde et al. 2013). Multiple annotators were asked to rate the compositionality of each German noun compound on an integer scale of 1 (non-compositional) to 7 (compositional). The overall compositionality score is then calculated as the mean across the annotators. The average of GNC compositionality scores is 4.83 ± 1.53.

Note that the component words are provided as part of the dataset, and that there is no need to perform decompounding, such as *bullauge* (*bulle + auge*) and *erdbeere* (*erde + beere*). Examples from the GNC dataset are shown in Table 3.3.

### 3.2 Secondary Datasets

This section details the publicly available resources used as inputs in our proposed approaches. In Chapter 4, PanLex is used as a translation database, and Wikipedia is used as a corpus in multiple languages to measure distributional similarity. Wiktionary is used in Chapter 5 to detect non-compositional components based on the provided definitions and translations. It is a monolingual dictionary in multiple languages which sometimes provides translation of words or MWEs in
other languages. And finally, the Universal Dependency Treebank, in Chapter 7, contains part-of-speech tag and dependency annotations in a universal formalism across multiple languages.

### 3.2.1 PanLex

Translations of MWEs and their components are used in Sections 4.1 and 4.2 to estimate the compositionality of MWEs. At the time of proposing the translation-based approaches in 2012 and 2013, there were several resources available to translate words into various languages such as Babelnet (Navigli and Ponzetto 2010), Wiktionary, PanLex (Baldwin et al. 2010) and Google Translate. As we were ideally after broad coverage over multiple languages and MWEs/component words in a given language, we excluded Babelnet and Wiktionary from our research. Babelnet covered only six languages at the time of proposing our approaches, and Wiktionary, because of its crowd-sourced nature, was patchy in its cross-linguistic coverage. This left translation resources such as PanLex and Google Translate. However, after manually analyzing the two resources for a range of MWEs, we decided not to use Google Translate for two reasons: (1) we consider the MWE out of context (i.e., we are working at the type level and do not consider the usage of the MWE in a particular sentence), and Google Translate tended to generate compositional translations of MWEs out of context; and (2) Google Translate provided only one translation for each component word/MWE. This left PanLex.

---

1. [http://lcl.uniroma1.it/babelnet/](http://lcl.uniroma1.it/babelnet/)
PanLex is an online translation database that is freely available. It contains lemmatized words and MWEs in a large variety of languages, with lemma-based (and less frequently sense-based) links between them. The database covers more than 1353 languages, and is made up of 12M lemmas and expressions. The translations are sourced from hand-built electronic dictionaries, making it more accurate than translation dictionaries generated automatically, e.g. through word alignment. Usually there are several direct translations for a word/MWE from one language to another, as in translations which were extracted from electronic dictionaries. If there is no direct translation for a word/MWE in the database, we can translate indirectly via one or more pivot languages (i.e., indirect translation (Soderland et al. 2010)). For example, the English *ivory tower* has direct translations in only 13 languages in PanLex, including French (*tour divoire*) but not Esperanto. There is, however, a translation of the French *tour divoire* into Esperanto (*ebura turo*), allowing us to infer an indirect translation between *ivory tower* and *ebura turo*, as illustrated in Figure 3.1. Since there might exist more than one translation for each word/MWE in a target language, we limit the number of translations to 10 for each target language, in order to reduce the complexity of our work.

For each MWE dataset, we translate the MWE and component words from the source language into 54 languages (see Appendix A for the list of the languages). We deliberately selected target languages with enough textual content in Wikipedia, in order to be able to compare a string similarity approach with a distributional similarity approach (Section 4.1 and Section 4.2, respectively).
3.2.2 Wikipedia

The proposed approach in Section 4.2 requires monolingual corpora in multiple languages, which we gathered from Wikipedia. Wikipedia is a free online encyclopedia, which is constantly updated by users around the world. According to Wikipedia, Wikipedia is:

a free-access, free-content Internet encyclopedia, supported and hosted by the non-profit Wikimedia Foundation. Those who can access the site can edit most of its articles, with the expectation that they follow the website’s policies. Wikipedia is ranked among the ten most popular websites and constitutes the Internet’s largest and most popular general reference work.

Wikipedia is considered a proper resource for our proposed method in Section 4.2, due to the large body of text, and since Wikipedia covers a large number of language varieties. English Wikipedia is the largest, containing over 4.9 million articles. All articles are freely available as Wikipedia dumps (or Wikidumps).

We collected monolingual corpora for 52 languages (51 target languages + 1 source language) from XML dumps of Wikipedia. These languages are based on the 54 target languages mentioned in Section 3.2.1 and listed in Appendix A, excluding Spanish because we happened not to have a dump of Spanish Wikipedia, and also Chinese and

---

4 https://www.wikipedia.org/
6 288 language varieties at the time of writing this thesis
7 https://dumps.wikimedia.org/
Chapter 3: Resources

Figure 3.2: The XML format of the English Wikipedia dump (https://en.wikipedia.org/wiki/Help:Export)
Japanese because of the need for a language-specific word tokeniser such as Stanford word segmenter for Chinese (Tseng et al. 2005). The raw corpora (Figure 3.2) were preprocessed using the WP2TXT toolbox\(^8\) to eliminate XML tags, HTML tags and hyperlinks. The corpora vary in size from roughly 750M tokens for English, to roughly 640K tokens for Marathi.

### 3.2.3 Wiktionary

The proposed approach in Chapter 5 requires a monolingual dictionary, which is ideally available in multiple languages and provides translations in other languages. We use Wiktionary (a lexical blend of *wiki* and *dictionary*), which is an online dictionary of words and expressions and available in more than 158 languages. Wiktionary, similarly to Wikipedia, is run by Wikimedia Foundation and is constantly growing with volunteer editors all around the world. According to Wiktionary, Wiktionary is:\(^9\)

1. A collaborative project run by the Wikimedia Foundation to produce a free and complete dictionary in every language.
2. The dictionaries, collectively, produced by that project.
3. A particular version of this dictionary project, written in a certain language, such as the English-language Wiktionary or simply the English Wiktionary.

A snapshot of a Wiktionary page is shown in Figure 3.3. Entries in Wiktionary usually contain one or more definitions, pronunciation, etymology, part of speech tags, some examples and translations.

---

\(^8\) [http://wp2txt.rubyforge.org/](http://wp2txt.rubyforge.org/)

natural language processing

Figure 3.3: A snapshot of Wiktionary entry for natural language processing
3.2.4 Universal Dependency Treebank

The Universal Dependency treebank \(^{10}\) (“UD”) is the first attempt to build a universal annotation scheme for dependency parsing, which is applicable across languages (Nivre et al. 2016). The annotation scheme includes a universal part-of-speech tag set (Petrov et al. 2012), and universal dependency relation set (De Marneffe and Manning 2008). One of the ultimate goals of the project is to build general-purpose, language-independent NLP pipelines.

At the time of writing this thesis, the most recent version was released in Nov 2015. We will use this version in our proposed approach to learn the MWE inventory of a surprise language introduced in Chapter 7. MWEs are labeled as either name (for named entities), compound (for binary compound expressions) or mwe (for fixed expressions (Sag et al. 2002)). In our proposed method in Chapter 7, we focus specifically on mwe and compound, as named entity recognition is a specialised sub-task of MWE identification with its own dedicated literature (Maynard et al. 2003; Huang et al. 2003; Steinberger and Pouliquen 2007), and there is every expectation that all languages contain named entities! According to the UD documentation mwe are used for fixed expressions, also known as words-with-spaces (Section 2.2.1). The compound should be used for any of the following compounding relations:

1. noun compounds (such as phone book), verb and adjective compounds in Japanese and Persian

2. numbers, such as four thousand

\(^{10}\)https://universaldependencies.github.io/docs/
3. particles of phrasal verbs such as *put up*

This relation should be used for noun compounds, verb and adjective compounds in Japanese and Persian, numbers (e.g., *four thousand*) and VPCs (Section 2.2) in labelling MWEs. There seem to be major inconsistencies in how they have been interpreted for particular languages: some languages do not use these relations at all, while others only annotate a subset of MWE types with these relations.

### 3.3 Classification and regression

This section outlines the classifiers and regressors we use in our experiments.
3.3.1 Support Vector Machine

In Chapters 4 and 5, we optionally use supervised classifiers to determine whether a given MWE is compositional or non-compositional, in form of support vector machines.

Supervised models are trained to recognize patterns in data based on a training set, based on which they can be used to predict the label of test instances. The training instances consist of a class label (e.g., compositional or non-compositional) and features, also known as attributes or observed variables. For example, the features in Section 4.1.2 are the string similarity score from each target language. Support vector machines (SVMs) are trained via optimization, in finding the optimal hyperplane (with slope \( w \) and intercept \( b \)) that provides the largest separation (margin) between two classes (Figure 3.4). Given \( N \) training instances of \( x_i \) with labels of \( y_i \in \{1, -1\}, i = 1\ldots N \), the soft margin SVM solves the following optimization problem:

\[
\min_{w,b,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} \xi_i
\]

subject to

\[
y_i(w.\phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.
\]

where, \( \phi \) maps the train instances into a higher dimensional space, and the outcome will be a linear hyperplane to separate the two classes, while solving the trade-off between the size of the margin and the slack variable\(^{11} \) of \( \xi \), with the defined cost of

\(^{11}\text{introduced to cope with infeasible constraints}\)
In our experiments, we use the LIBSVM package with radial basis function (RBF) kernel.\footnote{https://www.csie.ntu.edu.tw/~cjlin/libsvm/}

For more information on SVM classifiers, see Cortes and Vapnik (1995) and Hsu et al. (2003).

### 3.3.2 Support Vector Regression

In Chapter 4, we use support vector regressor to predict the compositionality degree of a given MWE based on the observed values of variables.

Support vector regression (SVR) is based on the SVM model, except that instead of predicting a class for a test sample, SVR predicts a score based on the observed values. For example, instead of determining whether an MWE is compositional or not, the regressor predicts the degree of compositionality. In SVR, $y_i \in R$, which slightly changes the optimization problem:

$$
\min_{w, b, \xi, \xi^*} \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*)
$$

subject to

$$
\begin{cases}
  y_i - w.\phi(x_i) - b \leq \xi_i \\
  w.\phi(x_i) + b - y_i \leq \xi_i^* \\
  \xi_i, \xi_i^* \geq 0
\end{cases}
$$

### 3.4 Evaluation metrics

Finally, we outline the metrics used to evaluate our proposed approaches.
3.4.1 Cross Validation

All experiments in Chapter 4 are carried out using 10 iterations of 10-fold cross validation, known as 10-times-10-fold cross validation. The dataset is randomly partitioned into 10 equal folds. Each fold will be considered as the test set once, with the rest of the folds as the training set. In order to have a more reliable evaluation, we repeat the same process 10 times, each time shuffling the data and partitioning it into 10 folds.

3.4.2 Correlation evaluation metric

Since we measure the compositionality degree of MWEs in Chapter 4, we need an evaluation metric to measure the linear correlation between our system’s scores and human annotations. Pearson’s correlation ($r$) is the metric we use through this thesis to measure such correlation. Pearson’s correlation coefficient between two variables $X$ and $Y$ ranges between 1 and $-1$, with 1 being perfect positive correlation, $-1$ being perfect negative correlation, and 0 indicating no correlation at all (i.e., random distribution between the sets of scores). The formula for Pearson’s correlation is as follows:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where, $X$ and $Y$ dataset $n$ values of $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$, respectively, and $\bar{x}$ and $\bar{y}$ are the $X$ and $Y$ sample mean.

13If the dataset size is not divisible by 10, we will have 10 roughly equal folds.
3.4.3 Classification evaluation metrics

In some of our experiments in Chapter 4 and the experiments in Chapter 5, our proposed methods determine whether a given instance belongs to one class or to another. Such binary classification problems are evaluated using precision, recall, f-score and accuracy metrics. In the following, we detail each metric, which are widely used in information retrieval and pattern recognition.

Before explaining each of these metrics, we introduce the terms true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Suppose we have 2 classes, namely positive and negative. TP instances are the instances in the positive class, which our classifier correctly classified as positive. FP are instances of the negative class, for which our classifier was wrong and classified as positive. With the same logic, FN instances are those positive instances that the classifier mistakenly labeled as negative, and TN instances, are those negative instances that the classifier labeled correctly (see Figure 3.5).

**Precision**  Precision (also known as positive predictive value) measures the proportion of positively classified instances that are correct. It does not capture any information about instances that are classified negatively.

\[
Precision = \frac{TP}{TP + FP}
\]

**Recall**  Recall (also known as sensitivity) measures the proportion of instances of a given class that are correctly classified as such. Recall does not capture any informa-
Figure 3.5: True positive, False positive, False negative and True negative.

Recall = \( \frac{TP}{TP + FN} \)

**F-score**  Precision and recall separately measure the correctness and coverage of the classifier, which we often want to combine into a single figure of merit. F-score aggregates precision and recall into a single value, in the form of the harmonic mean of the two values. F-score is measured as follows:

\[
F-score = 2 \times \frac{Precision \times Recall}{Precision + Recall}
\]

**Accuracy**  While precision, recall and F-score evaluate classifiers on positive class, accuracy measures the proportion of correctly predictions, as follows:
\[
\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
\]

**F-score or Accuracy?**

In cases where the classes are imbalanced, meaning that one class is significantly larger than the other class, F-score is more reliable than accuracy because a model can label every instance as the majority class and achieve a high accuracy, but the F-score of the smaller class will be undefined or considered zero. Also, when classification of one class is more important than the other, F-score is a better metric. For example, in Chapter 5, identifying instances which are non-compositional (positive class) is more important than the other class, because such MWEs should be treated differently to compositional MWEs in NLP applications. Therefore, F-score over the non-compositional class is more informative.

**3.4.4 Distribution Distance**

This section introduces two metrics we use in Chapter 7 to measure the similarity/distance between two probability distribution: Jensen-Shannon divergence and Mean squared error.

**Jensen-Shannon divergence**

Jensen-Shannon divergence (JSD) (also known as “information radius” or “total divergence to the average”), is a metric to measure the similarity between two probability distributions. JSD is a symmetrized and smoothed version of the Kullback-
Leibler divergence and is defined as follows:

\[ JS(P_1, P_2) = \frac{1}{2}(KL(P_1||\frac{P_1 + P_2}{2}) + KL(P_2||\frac{P_1 + P_2}{2})) = JS(P_2, P_1) \]

where, \( P_1 \) and \( P_2 \) are probability distributions and KL represents the Kullback-Leibler divergence defined as follows:

\[ KL(P_1||P_2) = \sum_i P_1(x_i) \log \frac{P_1(x_i)}{P_2(x_i)} \]

where, \( x \) is a discrete random variable.

We use JSD to measure the similarity between the distribution of two sets of MWE patterns in Chapter 7. JSD is equal to zero iff two distributions are the same.

**Mean squared error**

Mean squared error (MSE) measures the average of the squared errors between two discrete distributions. Errors are measured as the element-wise difference between two distributions (e.g., the predicted value and the observed value). MSE is measured as follows:

\[ MSE(p, q) = \frac{1}{n} \sum_{i=1}^{n} (p_i - q_i)^2 \]

MSE is another metric to measure the similarity between two distributions, and equals zero iff the two distributions are the same.
3.5 Summary

In this chapter, we described three datasets we will use to evaluate our proposed methods in chapters 4 and 5. Then, the SVM classifier and regressor, as well as the evaluation metrics used in this thesis were reviewed. Finally, two evaluation metrics to measure the similarity between two probability distribution were introduced, which will be used in Chapter 7.
Chapter 4

Predicting the Compositionality Degree of MWEs

As discussed in Section 2.4, recent studies on predicting the compositionality of MWEs are mostly language-dependent or type-specific. Either they consider characteristics of a specific type of MWE in a specific language or they require a resource which is not available in every language.

In this chapter we present three general approaches to predict the compositionality of multiword expressions. Such methods are easily applicable to any type of multiword expression, in a wide range of languages. Our three approaches are based on (1) translation string similarity, (2) translation distributional similarity, and (3) word embeddings.
4.1 Translation-based String Similarity

This section investigates the possibility of using MWEs’ translations in multiple languages to measure their degree of semantic compositionality by investigating how literal the meaning of each component is within the MWE. We first translate the MWE and its components, and then compare the translations of the MWE with the translations of its components using string similarity measures. The greater the string similarity, the more compositional we assume the MWE will be.

Our hypothesis is that compositional MWEs are more likely to be word-for-word translations in a given language than non-compositional MWEs. Hence, if we can locate the translations of the components in the translation of the MWE, we can deduce that it is compositional. Our second hypothesis is that the more languages we use as the basis for determining translation similarity between the MWE and its component words, the more accurately we will be able to estimate compositionality. The logic behind this hypothesis is that using multiple target languages will smooth out the effects of noisy and loan translations. Thus, rather than using just one translation language, we experiment with as many languages as possible.

As an example of our method, consider the English-to-Persian translation of *kick the bucket* as a non-compositional MWE and *make a decision* as a semi-compositional MWE (Table 4.1). By locating the translation of *decision* (*tasnim*) in the translation of *make a decision* (*tasnim gereftan*), we can deduce that it is semi-compositional. However, we cannot locate any of the component translations in the translation of *kick the bucket*. Therefore, we conclude that it is non-compositional. Note that in

---

1 Note that the Persian words are transliterated into English for ease of understanding.
Chapter 4: Predicting the Compositionality Degree of MWEs

Table 4.1: Examples of English MWEs and their components with translations into Persian. Direct matches between the translation of a MWE and its components are shown in **bold**; partial matches are undere line

<table>
<thead>
<tr>
<th>English</th>
<th>Persian Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>kick the bucket</td>
<td>mord</td>
</tr>
<tr>
<td>kick</td>
<td>zad</td>
</tr>
<tr>
<td>the bucket</td>
<td>satl</td>
</tr>
<tr>
<td>make a decision</td>
<td><strong>tasnim</strong> gereft</td>
</tr>
<tr>
<td>make</td>
<td>sakht</td>
</tr>
<tr>
<td>a decision</td>
<td>yek</td>
</tr>
<tr>
<td>public service</td>
<td>khadamaat <strong>omumi</strong></td>
</tr>
<tr>
<td>public</td>
<td><strong>omumi</strong></td>
</tr>
<tr>
<td>service</td>
<td>khedmat</td>
</tr>
</tbody>
</table>

In this example, the match is word-level, but that due to the effects of morphophonology, the more likely situation is that the components don’t match exactly (as we observe in the case of khadamaat and khedmat for the public service example, also shown in Table 4.1), which motivates our use of string similarity measures which can capture partial matches.

We will compare our work with the state-of-the-art approach which uses distributional similarity (Section 2.4.3). This distributional approach requires MWE-specific identification, which is difficult for some MWE types such as English verb particles (See Section 2.4.3). However, the proposed string similarity based approach does not require identification of MWEs in a large corpus and is less computationally expensive than distributional similarity based methods.

In the following sections we explain how the MWEs and components are translated into other languages (hereafter, “target languages”). Then, we introduce four
methods to measure the string similarity, and explain how the best target languages for each type of MWEs are selected. Then, we discuss the experiments over three datasets. Later, we propose an unsupervised approach based on string similarity. Finally, we show the findings of our error analysis.

### 4.1.1 Translations

The MWEs and their components are translated using PanLex (discussed in Section 3.2.1). In the case that no translation (direct or indirect) can be found for a given source language term into a particular target language, the compositionality score for that target language is set to the average across all target languages for which scores can be calculated for the given term. If no translations are available for any target language (e.g., the term is not in PanLex) the compositionality score for each target language is set to the average score for that target language across all other source language terms.

### 4.1.2 String Similarity

This section provides a simple, language-independent method for predicting the degree of compositionality of MWEs. We compare the translations of a MWE with the translations of its components, using a range of different languages and string similarity measures.

In the following, we introduce the measures we use to calculate the similarity between the MWE and its components. In each case, we normalize the output value to the range $[0, 1]$, where 1 indicates identical strings and 0 indicates completely different
strings. We will indicate the translation of the MWE in a particular language $t$ as $MWE^t$, and the translation of a given component in language $t$ as $component^t$.

**Longest common substring (LCS):** The LCS measure finds the longest common substring between two strings (Wagner and Fischer 1974). For example, the LCS between $\text{ABABC}$ and $\text{BABCAB}$ is $\text{BABC}$. We calculate a normalized similarity value based on the length of the LCS as follows:

$$\frac{\text{LongestCommonSubstring}(MWE^t, component^t)}{\min(\text{len}(MWE^t), \text{len}(component^t))}$$

**Levenshtein (LEV1):** The Levenshtein distance calculates the number of basic edit operations required to transpose one word into the other (Wagner and Fischer 1974). Edits consist of single-letter insertions, deletions or substitutions. We normalize LEV1 as follows:

$$1 - \frac{\text{LEV1}(MWE^t, component^t)}{\max(\text{len}(MWE^t), \text{len}(component^t))}$$

**Levenshtein with substitution penalty (LEV2):** One well-documented feature of Levenshtein distance (Baldwin 2009) is that substitutions are in fact the combination of an addition and a deletion, and as such can be considered to be two edits. Based on this observation, we experiment with a variant of LEV1 with this penalty applied for substitutions. We normalize LEV2 as follows:

$$1 - \frac{\text{LEV2}(MWE^t, component^t)}{\text{len}(MWE^t) + \text{len}(component^t)}$$
Chapter 4: Predicting the Compositionality Degree of MWEs

Smith Waterman (SW)  This method is based on the Needleman-Wunsch algorithm, and was developed to locally-align two protein sequences (Smith and Waterman 1981). It finds the optimal similar regions by maximizing the number of matches and minimizing the number of gaps necessary to align the two sequences. For example, the optimal local sequence for the two sequences below is AT−−ATCC, in which “−” indicates a gap:

Seq1: ATGCATCCATGAC
Seq2: TCTATATCCGT

As the example shows, it looks for the longest common substring but has an in-built mechanism for including gaps in the alignment (with penalty). This characteristic of SW might be helpful in our task, because there may be morphophonological variations between the MWE and component translations (as seen in the public service example of Table 4.1).

Let \( A = a_1a_2...a_m \) and \( B = b_1b_2...b_n \) and \( s(a_i, b_j) \) be 1 if \( a_i = b_j \) and 0, otherwise. Smith Waterman is a dynamic programming algorithm, in which matrix \( H \) is built as follows:

\[
H(i, 0) = 0, 0 \leq i \leq m
\]

\[
H(j, 0) = 0, 0 \leq j \leq n
\]

---

2The Needleman-Wunsch (NW) algorithm was designed to align two sequences of amino-acids (Needleman and Wunsch 1970). The algorithm looks for the sequence alignment which maximizes the similarity. As with the LEV score, NW minimizes edit distance, but also takes into account character-to-character similarity based on the relative distance between characters on the keyboard. We exclude this score, because it is highly similar to the LEV scores, and we did not obtain encouraging results using NW in our preliminary experiments.
Chapter 4: Predicting the Compositionality Degree of MWEs

\[
H(i, j) = \max \begin{cases} 
0 \\
H(i - 1, j - 1) + s(a_i, b_j) & \text{Match/Mismatch} \\
\max_{k \geq 1} H(i - k, j) + W_k & \text{Deletion} \\
\max_{l \geq 1} H(i, j - l) + W_l & \text{Insertion} 
\end{cases}, 1 \leq i \leq m, 1 \leq j \leq n
\]

where, \(W_k\) and \(W_l\) are deletion and insertion penalty, set to \(-1\).

Backtracking starts at the highest cell value and proceeds depending on the direction the matrix \(H\) is constructed, shown in Matrix \(D\).

We normalize SW similarly to LCS:

\[
\frac{\text{len(alignedSequence)}}{\min(\text{len}(\text{MWE}^t), \text{len}(\text{component}^t))}
\]

Where, \(\text{len(alignedSequence)}\) equals the number of diagonal jumps in matrix \(D\).
Chapter 4: Predicting the Compositionality Degree of MWEs

4.1.3 Computational Model

Given the scores calculated by the aforementioned string similarity measures between the translations for a given component word and the MWE, we need some way of combining scores across component words.\(^3\) First, we measure the compositionality of each component ($s_1$ and $s_2$) within the MWE (Figure 4.1):

\[
\begin{align*}
    s_1 &= f_1(sim_1(component_1, MWE), \ldots, sim_1(component_1, MWE)) \\
    s_2 &= f_1(sim_1(component_2, MWE), \ldots, sim_1(component_2, MWE))
\end{align*}
\]

where \(sim\) is a string similarity measure, \(sim_i\) indicates that the calculation is based on translations in language \(i\), and \(f_1\) is a score combination function.

\(^3\)Note that in all experiments we only combine scores given by the same string similarity measure.
component\textsubscript{1} scores for each language \hspace{1cm} component\textsubscript{2} scores for each language

\[ CS_{\text{method}} = f_1 \]

\[ f_2(s_1, s_2) = \alpha s_1 + (1 - \alpha) s_2 \]

compositionality score (s\textsubscript{3})

Figure 4.2: Schematic outline of our proposed method to combine the scores from multiple languages

Then, we compute the overall compositionality of the MWE (s\textsubscript{3}) from s\textsubscript{1} and s\textsubscript{2} using \( f_2 \):

\[ s_3 = f_2(s_1, s_2) \]

Since we often have multiple translations for a given component word/MWE in PanLex, we exhaustively compute the similarity between each MWE translation and component translation, and use the highest similarity as the result of \( sim_i \). Note that word order is not an issue in our method, as we calculate the similarity independently for each MWE component.

In this research, we consider simple functions for \( f_1 \) such as Mean, Median, Product, Min and Max. Here, following Reddy \textit{et al.} (2011), we experimented with
weighted mean for $f_2$ (Figure 4.2):

$$f_2(s_1, s_2) = \alpha s_1 + (1 - \alpha) s_2$$

### 4.1.4 Language Selection

Our method is based on the translation of a MWE into many languages. In the first stage, we chose 54 languages (Appendix A) for which relatively large corpora were available for our next proposed approach (Section 4.2).

All experiments are carried out using 10 iterations of 10-fold cross validation, randomly partitioning the data independently on each of the 10 iterations, and averaging across all 100 test partitions in our presented results (Section 3.4.1). Then, we select the 10 most correlated languages with human judgements for each cross-validation training partition, based on the Pearson’s correlation (introduced in Section 3.4.2) between the given scores in that language and human judgements.\(^4\)

### 4.1.5 Experiment (I): ENCs

**Compositionality of the components**

In this section, we investigate the relation between the compositionality score of the components and the compositionality score of the whole noun compound. Based on the ENC dataset, there is a high correlation between the noun compound compositionality score and the first component score ($r=0.784$). The correlation shows a positive, roughly linear, relation between the two compositionality scores.

\(^4\)Note that for VPCs, we calculate the compositionality of only the verb part, because we do not have the human judgements for the whole VPC (see Section 3.1.2).
other words, the higher the compositionality score for the first component, the higher the compositionality score is for the whole noun compound. This is the same for the second component, although the correlation is slightly lower ($r=0.717$).

Figures 4.3 and 4.4 show the relation between the compositionality of English noun compounds and each of the components. These correlation scores confirm that by estimating the scores for each component within the noun compound, we can estimate the compositionality score of the whole noun compound.

**Results**

As mentioned in Section 4.1.4, we perform 10-fold cross-validation to select the 10 best languages on the training data for each fold. The selected languages for a given fold are then used to compute $s_1$, $s_2$ and $s_3$ for each instance in the test
Chapter 4: Predicting the Compositionality Degree of MWEs

Figure 4.4: Relation between the compositionality of the second component and the whole noun compound over the ENC dataset.

set. Using Pearson’s correlation, the computed $s_1$, $s_2$ and $s_3$ scores are compared with human judgements for the first components, the second components and the noun compounds, respectively. The results are shown in Table 4.2. Among the five functions we experimented with for $f_1$, Mean performs much more consistently than the others. Median is less prone to noise, and therefore performs better than Prod, Max and Min, but it is still worse than Mean.

For the most part, LCS and SW perform better than the other measures. There is little to separate these two methods, partly because they both look for a sequence of similar characters, unlike LEV1 and LEV2 which do not consider contiguity of match.

Our best result using the 10 selected languages on the ENC dataset is $r = 0.644$, as compared to the best single-language correlation of $r = 0.543$ for Portuguese. The results support our hypothesis that using multiple target languages rather than one,
Chapter 4: Predicting the Compositionality Degree of MWEs

Table 4.2: Correlation on the ENC dataset. N1, N2 and NC, are the first component of the noun compound, its second component, and the noun compound itself, respectively. Correlation using only distributional similarity (DS) is \( r = 0.700 \).

<table>
<thead>
<tr>
<th>( f_1 )</th>
<th>( CS_{method} )</th>
<th>N1</th>
<th>N2</th>
<th>NC</th>
<th>NC+DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>( CS_{SW}(Mean) )</td>
<td>0.532</td>
<td>0.415</td>
<td>0.644</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>( CS_{LCS}(Mean) )</td>
<td>0.523</td>
<td>0.436</td>
<td>0.644</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV1}(Mean) )</td>
<td>0.398</td>
<td>0.210</td>
<td>0.502</td>
<td>0.713</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV2}(Mean) )</td>
<td>0.453</td>
<td>0.290</td>
<td>0.566</td>
<td>0.727</td>
</tr>
<tr>
<td>Product</td>
<td>( CS_{SW}(Product) )</td>
<td>0.465</td>
<td>0.256</td>
<td>0.526</td>
<td>0.691</td>
</tr>
<tr>
<td></td>
<td>( CS_{LCS}(Product) )</td>
<td>0.440</td>
<td>0.264</td>
<td>0.497</td>
<td>0.685</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV1}(Product) )</td>
<td>0.333</td>
<td>0.106</td>
<td>0.303</td>
<td>0.706</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV2}(Product) )</td>
<td>0.327</td>
<td>0.149</td>
<td>0.327</td>
<td>0.689</td>
</tr>
<tr>
<td>Median</td>
<td>( CS_{SW}(Median) )</td>
<td>0.447</td>
<td>0.327</td>
<td>0.570</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>( CS_{LCS}(Median) )</td>
<td>0.432</td>
<td>0.339</td>
<td>0.562</td>
<td>0.692</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV1}(Median) )</td>
<td>0.308</td>
<td>0.069</td>
<td>0.386</td>
<td>0.675</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV2}(Median) )</td>
<td>0.421</td>
<td>0.154</td>
<td>0.491</td>
<td>0.707</td>
</tr>
<tr>
<td>Min</td>
<td>( CS_{SW}(Min) )</td>
<td>0.415</td>
<td>0.203</td>
<td>0.398</td>
<td>0.633</td>
</tr>
<tr>
<td></td>
<td>( CS_{LCS}(Min) )</td>
<td>0.330</td>
<td>0.242</td>
<td>0.334</td>
<td>0.602</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV1}(Min) )</td>
<td>0.371</td>
<td>0.311</td>
<td>0.390</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV2}(Min) )</td>
<td>0.361</td>
<td>0.325</td>
<td>0.391</td>
<td>0.601</td>
</tr>
<tr>
<td>Max</td>
<td>( CS_{SW}(Max) )</td>
<td>0.335</td>
<td>0.344</td>
<td>0.442</td>
<td>0.701</td>
</tr>
<tr>
<td></td>
<td>( CS_{LCS}(Max) )</td>
<td>0.325</td>
<td>0.347</td>
<td>0.442</td>
<td>0.708</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV1}(Max) )</td>
<td>0.304</td>
<td>0.363</td>
<td>0.492</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>( CS_{LEV2}(Max) )</td>
<td>0.338</td>
<td>0.361</td>
<td>0.493</td>
<td>0.708</td>
</tr>
</tbody>
</table>

Our baseline is a reimplementation of the method of Reddy et al. (2011), which is based on distributional similarity (DS) and was the previous state-of-the-art for the ENC dataset. They proposed to compare the semantic vector of the MWE with that of its components using cosine similarity (see Section 2.4.3 for more details). The computed correlation is \( r = 0.700 \). We experimented with combining our method \( (CS_{LCS}(Mean)) \) with that method of Reddy et al. (2011), based on simple averaging, as detailed in Table 4.2 (reported as NC+DS). The results are higher than both the indi-
Table 4.3: The 10 best languages for the ENC dataset using LCS.

<table>
<thead>
<tr>
<th>Language</th>
<th>Frequency</th>
<th>Family</th>
<th>Coverage(%)</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>100</td>
<td>Slavic</td>
<td>91</td>
<td>0.531</td>
</tr>
<tr>
<td>Portuguese</td>
<td>100</td>
<td>Romance</td>
<td>81</td>
<td>0.543</td>
</tr>
<tr>
<td>Thai</td>
<td>100</td>
<td>Kam-thai</td>
<td>79</td>
<td>0.515</td>
</tr>
<tr>
<td>Norwegian</td>
<td>100</td>
<td>Germanic</td>
<td>65</td>
<td>0.487</td>
</tr>
<tr>
<td>French</td>
<td>95</td>
<td>Romance</td>
<td>82</td>
<td>0.492</td>
</tr>
<tr>
<td>Chinese</td>
<td>95</td>
<td>Chinese</td>
<td>94</td>
<td>0.492</td>
</tr>
<tr>
<td>Dutch</td>
<td>94</td>
<td>Germanic</td>
<td>77</td>
<td>0.487</td>
</tr>
<tr>
<td>Romanian</td>
<td>89</td>
<td>Romance</td>
<td>61</td>
<td>0.476</td>
</tr>
<tr>
<td>Hindi</td>
<td>74</td>
<td>Indic</td>
<td>63</td>
<td>0.472</td>
</tr>
<tr>
<td>Arabic</td>
<td>39</td>
<td>Semitic</td>
<td>76</td>
<td>0.446</td>
</tr>
</tbody>
</table>

Table 4.3: The 10 best languages for the ENC dataset using LCS.

Individual string similarity methods and the state-of-the-art for the ENC dataset, demonstrating the complementarity between our proposed method and methods based on distributional similarity.

**Best languages**

The 10 best languages are selected based only on the training set for each fold. The languages selected for each fold will later be used to predict the compositionality of the items in the testing portion for that fold. In Table 4.3, we show how often each language was selected in the top-10 languages over the combined 100 (10×10) folds of cross validation, based on LCS.

The frequencies show that the selected languages were mostly consistent over the folds. The languages are a mixture of Romance, Germanic and languages from other families (based on the language classification of Voegelin and Voegelin (1977)).

The question that arises here is why these languages are consistently selected as the best languages. Figure 4.5 shows the relation between the number of missing
Figure 4.5: Relation between the coverage and the correlation score of target languages for the ENC dataset.

translations in a language and the correlation of the scores given by that language and the human annotations. According to the figure, the languages with higher translation coverage produce scores that correlate better with human annotation. In Table 4.3, we also show the translation coverage and the correlation with human annotation of each language. Using only the 10 languages with the highest coverage (shown in Table 4.10, Section 4.1.9), however, does not perform as well as the selected languages in Table 4.3, with the NC score of 0.642 and NC+DS correlation score of 0.740.
Chapter 4: Predicting the Compositionality Degree of MWEs

Figure 4.6: Relation between different values of $\alpha$ and the correlation score for the compositionality of English noun compounds

**Alpha parameter**

As mentioned before, we use a weighted mean to combine the scores from each component to calculate the compositionality score of the whole noun compound.

$$f_2(s_1, s_2) = \alpha s_1 + (1 - \alpha) s_2$$

The parameter $\alpha$ was 0.7 in all of the above experiments, and chosen according to the findings of Reddy *et al.* (2011). However, in order to make sure that this value is also the best value for our proposed method when applied to English noun
compounds, we experimented with different values of $\alpha$. The result is shown in Figure 4.6. The curves show that 0.7 gives the highest correlation score for all string similarity measures except LEV1. Also, it is shown that the score given by the first component ($\alpha = 1$) is more important than the score given by the second component ($\alpha = 0$), because the first component compositionality is more strongly correlated with the compositionality of noun compound ($r = 0.789$) than the second component ($r = 0.717$). Therefore, this also confirms that the first component should get a higher weight in the weighted mean.

4.1.6 Experiment (II): EVPCs

In Table 4.4, we compare the four string similarity measures and five proposed $f_1$ functions for English verb particle constructions (EVPCs). Note that the VPC correlation scores are the same as the verb correlation scores. This is because the EVPC dataset does not have annotations for the compositionality of the whole EVPC. Therefore, in accordance with Bannard et al. (2003) as discussed in Section 3.1.2, we consider the verb scores to be equivalent to the overall EVPC scores. As with the English noun compounds, Mean is the best among the five $f_1$ functions. Among the four string similarity measures, LCS performs better than the others.

Compositionality prediction has been shown to be difficult for MWE types including English VPCs (McCarthy et al. 2003; Baldwin 2005). As such, the fact that our method is as competitive as this is highly encouraging, especially when you consider that it can equally be applied to different types of MWEs in other languages.

In Table 4.5, we compare our results ($CS_{LCS(Mean)}$) with those of Bannard et al.
Table 4.4: Correlation on EVPC dataset, based on the best-10 languages for the verb and particle individually. Correlation using only distributional similarity (DS) is \( r = 0.177 \) (2003), who interpreted the dataset as a binary classification task. The dataset used in their study is a subset of the EVPC dataset, containing 40 VPCs, of which 29 (72%) were verb compositional and 23 (57%) were particle compositional. By applying a threshold of 0.5 over the output of our regression model, we binarize the VPCs into the compositional and non-compositional classes. According to the results shown in Table 4.4, LCS is the best similarity measure for this task. Our proposed method has higher results than the best results of Bannard et al. (2003), in part due to their

\[ f_1(sim) \]

<table>
<thead>
<tr>
<th></th>
<th>Verb</th>
<th>Particle</th>
<th>VPC</th>
<th>VPC+DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( CS_{SW} ) (Mean)</td>
<td>0.349</td>
<td>0.504</td>
<td>0.349</td>
<td>0.318</td>
</tr>
<tr>
<td>( CS_{LCS} ) (Mean)</td>
<td><strong>0.385</strong></td>
<td><strong>0.509</strong></td>
<td><strong>0.385</strong></td>
<td><strong>0.360</strong></td>
</tr>
<tr>
<td>( CS_{LEV1} ) (Mean)</td>
<td>0.328</td>
<td>0.477</td>
<td>0.328</td>
<td>0.348</td>
</tr>
<tr>
<td>( CS_{LEV2} ) (Mean)</td>
<td>0.327</td>
<td>0.449</td>
<td>0.327</td>
<td>0.333</td>
</tr>
<tr>
<td>Product</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( CS_{SW} ) (Product)</td>
<td>0.327</td>
<td>0.318</td>
<td>0.327</td>
<td>0.348</td>
</tr>
<tr>
<td>( CS_{LCS} ) (Product)</td>
<td>0.348</td>
<td>0.332</td>
<td>0.348</td>
<td>0.369</td>
</tr>
<tr>
<td>( CS_{LEV1} ) (Product)</td>
<td>0.327</td>
<td>0.268</td>
<td>0.327</td>
<td>0.368</td>
</tr>
<tr>
<td>( CS_{LEV2} ) (Product)</td>
<td>0.350</td>
<td>0.292</td>
<td>0.350</td>
<td><strong>0.383</strong></td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( CS_{SW} ) (Median)</td>
<td>0.300</td>
<td>0.430</td>
<td>0.303</td>
<td>0.306</td>
</tr>
<tr>
<td>( CS_{LCS} ) (Median)</td>
<td>0.348</td>
<td>0.425</td>
<td>0.348</td>
<td>0.358</td>
</tr>
<tr>
<td>( CS_{LEV1} ) (Median)</td>
<td>0.291</td>
<td>0.396</td>
<td>0.291</td>
<td>0.327</td>
</tr>
<tr>
<td>( CS_{LEV2} ) (Median)</td>
<td>0.287</td>
<td>0.403</td>
<td>0.287</td>
<td>0.319</td>
</tr>
<tr>
<td>Min</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( CS_{SW} ) (Min)</td>
<td>0.247</td>
<td>0.201</td>
<td>0.247</td>
<td>0.279</td>
</tr>
<tr>
<td>( CS_{LCS} ) (Min)</td>
<td>0.325</td>
<td>0.249</td>
<td>0.325</td>
<td>0.349</td>
</tr>
<tr>
<td>( CS_{LEV1} ) (Min)</td>
<td>0.313</td>
<td>0.292</td>
<td>0.313</td>
<td>0.346</td>
</tr>
<tr>
<td>( CS_{LEV2} ) (Min)</td>
<td>0.304</td>
<td>0.287</td>
<td>0.304</td>
<td>0.334</td>
</tr>
<tr>
<td>Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( CS_{SW} ) (Max)</td>
<td>0.087</td>
<td>0.160</td>
<td>0.087</td>
<td>0.203</td>
</tr>
<tr>
<td>( CS_{LCS} ) (Max)</td>
<td>0.157</td>
<td>0.213</td>
<td>0.157</td>
<td>0.224</td>
</tr>
<tr>
<td>( CS_{LEV1} ) (Max)</td>
<td>0.112</td>
<td>0.462</td>
<td>0.112</td>
<td>0.204</td>
</tr>
<tr>
<td>( CS_{LEV2} ) (Max)</td>
<td>0.122</td>
<td>0.407</td>
<td>0.122</td>
<td>0.214</td>
</tr>
</tbody>
</table>

5 This dataset was the only available dataset of VPCa at the time of Bannard et al. (2003). They increased the number of VPCs to 160 and published the dataset in Bannard (2006).
Table 4.5: Results for the classification task.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score ($\beta = 1$)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bannard et al. (2003)</td>
<td>60.8</td>
<td>66.6</td>
<td>63.6</td>
<td>60.0</td>
</tr>
<tr>
<td>$CS_{\text{LCS(Mean)}}$</td>
<td>86.2</td>
<td>71.8</td>
<td>77.4</td>
<td>69.3</td>
</tr>
</tbody>
</table>

reliance on VPC identification, and the low recall on the task which they report. Our proposed method does not rely on a corpus or MWE identification.

**Best languages**

In this section we show the 10 best selected target languages when experimenting with EVPCs. In Table 4.6, we show how often each language was selected in the top-10 languages over the combined 100 ($10 \times 10$) folds of nested 10-fold cross validation, based on LCS.

There is no overlap whatsoever between the verb and the particle top-10 languages. The lists of the selected best languages also differ from the best languages for ENC (Table 4.3). This shows that our proposed method is selecting the best target languages according to the type of the word/MWE, and also that no single language is the best language for all three different cases (ENC, VPC:verb and VPC:particle). This is an advantage over previous studies which select a fixed target language for this task.

Figures 4.7 and 4.8 show the relation between the coverage of the target languages and their correlation with annotations of the EVPC dataset. As for the ENC dataset, languages with higher coverage tend to have a higher correlation. However, using the top-10 languages with the highest coverage (Table 4.10) instead of the best selected
Table 4.6: The 10 best languages for the verb/particle component of the EVPC dataset using LCS.

languages results in a lower overall correlation score (0.323 for VPC correlation score and 0.318 for VPC+DS, Section 4.1.9).

4.1.7 Experiment (III): GNCs

Compositionality of the components

In this section, we examine the correlation between the compositionality score of each component and the compositionality of the noun compound over the German
Figure 4.7: Correlation between the coverage and the correlation score of target languages for the verb component of the EVPC dataset

Figure 4.8: Relation between the coverage and the correlation score of target languages for the second component of the EVPC dataset
noun compound (GNC) dataset. The results show that there is a positive correlation ($r = 0.671$) between the noun compound and the first component, although it is not as large as for English noun compounds. However, unlike the English noun compounds, the correlation of the second component and the noun compound is very low ($r = 0.202$). Figures 4.9 and 4.10 illustrate the scatter plots for a better understanding of the relations. The results show that by knowing the compositionality of the components within the German noun compound, we can estimate the compositionality of the whole noun compound. However, the first component has more influence on the compositionality of German noun compounds.
Chapter 4: Predicting the Compositionality Degree of MWEs

Second component compositionality score
Noun compound compositionality score
\[ r = 0.202 \]

Figure 4.10: Relation between the compositionality of the second component and the whole noun compound over the GNC dataset

Results

In this section, we evaluate the four string similarity measures and five proposed \( f_1 \) functions on German noun compounds. The German compound noun dataset is deliberately chosen to show how our proposed approach works on non-English MWEs. Note that German noun compounds need to be decompounded to extract the components such as Bullauge (Bulle + Auge) and Erdbeere (Erde + Beere). Our dataset contains the MWEs along with their decompounded components.

In Table 4.7, we show the results of applying the four different string similarity measures and five \( f_1 \) functions. As for the ENC and EVPC datasets, Mean is the best \( f_1 \) function. LEV2, SW and LCS are performing better than LEV1. Unlike the other two datasets, LEV2 is the best performing method and SW is slightly better.
### Table 4.7: Correlation on GNCs dataset.

<table>
<thead>
<tr>
<th>Sim</th>
<th>N1</th>
<th>N2</th>
<th>NC</th>
<th>NC+DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CS_{SW}$(Mean)</td>
<td>0.368</td>
<td>0.351</td>
<td>0.379</td>
<td>0.357</td>
</tr>
<tr>
<td>$CS_{LCS}$(Mean)</td>
<td>0.358</td>
<td>0.357</td>
<td>0.372</td>
<td>0.353</td>
</tr>
<tr>
<td>$CS_{LEV1}$(Mean)</td>
<td>0.2660</td>
<td>0.310</td>
<td>0.318</td>
<td>0.318</td>
</tr>
<tr>
<td>$CS_{LEV2}$(Mean)</td>
<td>0.3296</td>
<td>0.341</td>
<td>0.389</td>
<td>0.361</td>
</tr>
<tr>
<td>Product</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CS_{SW}$(Product)</td>
<td>0.215</td>
<td>0.181</td>
<td>0.140</td>
<td>0.176</td>
</tr>
<tr>
<td>$CS_{LCS}$(Product)</td>
<td>0.187</td>
<td>0.175</td>
<td>0.126</td>
<td>0.167</td>
</tr>
<tr>
<td>$CS_{LEV1}$(Product)</td>
<td>0.117</td>
<td>0.107</td>
<td>0.016</td>
<td>0.137</td>
</tr>
<tr>
<td>$CS_{LEV2}$(Product)</td>
<td>0.119</td>
<td>0.172</td>
<td>0.060</td>
<td>0.146</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CS_{SW}$(Median)</td>
<td>0.309</td>
<td>0.288</td>
<td>0.313</td>
<td>0.314</td>
</tr>
<tr>
<td>$CS_{LCS}$(Median)</td>
<td>0.291</td>
<td>0.309</td>
<td>0.301</td>
<td>0.305</td>
</tr>
<tr>
<td>$CS_{LEV1}$(Median)</td>
<td>0.264</td>
<td>0.213</td>
<td>0.311</td>
<td>0.313</td>
</tr>
<tr>
<td>$CS_{LEV2}$(Median)</td>
<td>0.303</td>
<td>0.225</td>
<td>0.322</td>
<td>0.314</td>
</tr>
<tr>
<td>Min</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CS_{SW}$(Min)</td>
<td>0.195</td>
<td>0.232</td>
<td>0.218</td>
<td>0.269</td>
</tr>
<tr>
<td>$CS_{LCS}$(Min)</td>
<td>0.195</td>
<td>0.229</td>
<td>0.214</td>
<td>0.267</td>
</tr>
<tr>
<td>$CS_{LEV1}$(Min)</td>
<td>0.135</td>
<td>0.179</td>
<td>0.144</td>
<td>0.200</td>
</tr>
<tr>
<td>$CS_{LEV2}$(Min)</td>
<td>0.154</td>
<td>0.238</td>
<td>0.190</td>
<td>0.243</td>
</tr>
<tr>
<td>Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CS_{SW}$(Max)</td>
<td>0.254</td>
<td>0.094</td>
<td>0.180</td>
<td>0.201</td>
</tr>
<tr>
<td>$CS_{LCS}$(Max)</td>
<td>0.256</td>
<td>0.111</td>
<td>0.187</td>
<td>0.206</td>
</tr>
<tr>
<td>$CS_{LEV1}$(Max)</td>
<td>0.166</td>
<td>0.238</td>
<td>0.143</td>
<td>0.170</td>
</tr>
<tr>
<td>$CS_{LEV2}$(Max)</td>
<td>0.212</td>
<td>0.239</td>
<td>0.176</td>
<td>0.197</td>
</tr>
</tbody>
</table>

As with the previous experiments, distributional similarity approach is considered as the baseline, which requires identifying the instances in a large corpus (See Sections 2.4.3 and 4.2 for more information). German is a morphologically-rich language, with marking of number and case on nouns. Given that we do not perform any lemmatisation or other language-specific preprocessing, we inevitably achieve low recall for the identification of noun compound tokens, although the precision should be higher than LCS.
Table 4.8: The 10 best languages for German noun compounds

<table>
<thead>
<tr>
<th>Language</th>
<th>Frequency</th>
<th>Family</th>
<th>Coverage(%)</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polish</td>
<td>100</td>
<td>Slavic</td>
<td>64.2</td>
<td>0.105</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>100</td>
<td>Baltic</td>
<td>54.5</td>
<td>0.069</td>
</tr>
<tr>
<td>Finnish</td>
<td>80</td>
<td>Uralic</td>
<td>56.9</td>
<td>0.185</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>78</td>
<td>Slavic</td>
<td>49.2</td>
<td>0.134</td>
</tr>
<tr>
<td>Czech</td>
<td>45</td>
<td>Slavic</td>
<td>71.5</td>
<td>0.229</td>
</tr>
<tr>
<td>Arabic</td>
<td>35</td>
<td>Semitic</td>
<td>65.9</td>
<td>0.235</td>
</tr>
<tr>
<td>Telugu</td>
<td>35</td>
<td>Dravidian</td>
<td>66.7</td>
<td>0.130</td>
</tr>
<tr>
<td>Romanian</td>
<td>22</td>
<td>Romance</td>
<td>64.6</td>
<td>0.176</td>
</tr>
<tr>
<td>English</td>
<td>21</td>
<td>Germanic</td>
<td>65.5</td>
<td>0.147</td>
</tr>
<tr>
<td>Slovenian</td>
<td>20</td>
<td>Slavic</td>
<td>70.3</td>
<td>0.194</td>
</tr>
</tbody>
</table>

be nearly 100%. Partly because of the resultant sparseness of semantic vectors in the distributional similarity method, the result for DS is low ($r = 0.141$). Simple string similarity achieves better results for this dataset, and the results actually drop slightly when combined with distributional similarity.

Our best result on the GNC dataset using string similarity of $r = 0.389$ is competitive with the state-of-the-art results ($r = 0.45$) using a window-based distributional similarity approach over monolingual German data (Schulte im Walde et al. 2013). Note, however, that the state-of-the-art method uses part-of-speech information and lemmatisation, where ours does not, in keeping with the language-independent philosophy of this research.
Best languages

In this section, we show the 10 best languages selected as the target languages when experimenting with German noun compounds. In Table 4.8, we show how often each language is selected along with the correlation and translation coverage of each language in PanLex. Unlike the other two datasets, less consistency in the language selection over the combined 100 folds is observable according to the number of times each language is chosen (Frequency) as one of the 10 best languages.

6 Due to morphological richness
7 This score is reported by Schulte im Walde et al. (2013) when predicting the compound ratings by adding the the modifier and head predictions.
Figure 4.11 shows that the languages with a smaller number of missing translations tend to correlate better with human annotations. However, only using the top-10 languages with the highest coverage (Table 4.10, Section 4.1.9) does not result in a better performance (NC correlation = 0.343 and NC+DS = 0.332, Section 4.1.9). According to the figure, the highest correlation of a single language is around $r = 0.32$. However, by selecting the best target languages according to the training set, and combining their scores, we observe a higher correlation ($r = 0.39$). This again confirms our hypothesis that multiple target languages perform better than only one target language.

**Alpha parameter**

In this section, we optimise the value of $\alpha$ in the following equation:

$$f_2(s_1, s_2) = \alpha s_1 + (1 - \alpha) s_2$$

Figure 4.12 shows the correlation scores for different $\alpha$ values and different string similarity measures. According to the diagrams, $\alpha = 0.7$ is once again optimal. Based on the plots, we conclude that, as for the ENC dataset, the first component compositionality score ($\alpha = 1$) is more correlated with the compositionality of the whole GNC than the second component ($\alpha = 0$).

### 4.1.8 Missing Translations

According to the coverage statistics in Table 4.8, and the number of missing translation for each instance in Figure 4.11, we observe that unlike the other two datasets, the GNC dataset does not have high coverage in PanLex for any of the
selected target languages. This means that the approach for dealing with missing translations is vitally important. So far, we have used the mean score for missing translations. In this section for the GNC dataset, we do not consider the missing scores at all. We consider the scores we get from each language and ignore languages with missing translations for a given GNC. The results are shown in Table 4.9. The results are more consistent with previous results for the ENC and EVPC datasets; LCS and SW are ahead of LEV1 and LEV2, with LCS the best and LEV1 the worst.
Chapter 4: Predicting the Compositionality Degree of MWEs

<table>
<thead>
<tr>
<th>sim()</th>
<th>N1</th>
<th>N2</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW</td>
<td>0.384</td>
<td>0.298</td>
<td>0.404</td>
</tr>
<tr>
<td>LCS</td>
<td>0.392</td>
<td>0.324</td>
<td>0.423</td>
</tr>
<tr>
<td>LEV1</td>
<td>0.227</td>
<td>0.285</td>
<td>0.236</td>
</tr>
<tr>
<td>LEV2</td>
<td>0.299</td>
<td>0.293</td>
<td>0.321</td>
</tr>
</tbody>
</table>

Table 4.9: The results after ignoring missing translations on GNC dataset

<table>
<thead>
<tr>
<th>ENC</th>
<th>EVPC:verb</th>
<th>GNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>German</td>
<td>English</td>
</tr>
<tr>
<td>Finnish</td>
<td>Finnish</td>
<td>Japanese</td>
</tr>
<tr>
<td>French</td>
<td>French</td>
<td>French</td>
</tr>
<tr>
<td>Italian</td>
<td>Italian</td>
<td>Italian</td>
</tr>
<tr>
<td>Russian</td>
<td>Japanese</td>
<td>Russian</td>
</tr>
<tr>
<td>Spanish</td>
<td>Hungarian</td>
<td>Hungarian</td>
</tr>
<tr>
<td>Portuguese</td>
<td>Dutch</td>
<td>Dutch</td>
</tr>
<tr>
<td>Japanese</td>
<td>Polish</td>
<td>Turkish</td>
</tr>
<tr>
<td>Chinese</td>
<td>Chinese</td>
<td>Chinese</td>
</tr>
<tr>
<td>Czech</td>
<td>Czech</td>
<td>Czech</td>
</tr>
</tbody>
</table>

Table 4.10: The 10 languages with the highest translation coverage for ENC, EVPC and GNC (Common languages with the supervised selected languages are shown in bold).

4.1.9 Unsupervised Approach

The proposed translation-based string similarity approach has been supervised so far, in that the best target languages are selected based on training data. As mentioned in Section 4.1.4, the best target languages are those whose scores have the highest correlation with annotations. According to our experiments, we showed that there is a correlation between being a good language and its coverage in Panlex. In other words, the languages for which most of the MWEs have a translation to, result in higher correlation. In this section, an unsupervised approach is presented in which
Table 4.11: Correlation on ENC, EVPC and GNC datasets using supervised and unsupervised string similarity approach.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>SS</th>
<th>SS+DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENC</td>
<td>Supervised</td>
<td>0.644</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td>Unsupervised</td>
<td>0.642</td>
<td>0.740</td>
</tr>
<tr>
<td>EVPC</td>
<td>Supervised</td>
<td>0.385</td>
<td>0.360</td>
</tr>
<tr>
<td></td>
<td>Unsupervised</td>
<td>0.323</td>
<td>0.318</td>
</tr>
<tr>
<td>GNC</td>
<td>Supervised</td>
<td>0.372</td>
<td>0.353</td>
</tr>
<tr>
<td></td>
<td>Unsupervised</td>
<td>0.343</td>
<td>0.332</td>
</tr>
</tbody>
</table>

The results of our unsupervised approach along with the supervised approach are shown in Table 4.11 for the three datasets. According to the results, despite the lower correlation scores for the proposed unsupervised method, this method is comparable to the supervised approach. Therefore, in the case of not having a train set for a group of MWEs (no matter in what language or what type of MWE), we suggest using the target languages for which the majority of those MWEs have a translation to. We will use this unsupervised approach in Section 4.3.

4.1.10 Error Analysis

We analysed items in the ENC dataset which have a high difference (> 2.5) between the human annotation and our scores (based on LCS and Mean). This results in four ENCs: *cutting edge*, *melting pot*, *gold mine* and *ivory tower* which are non-
compositionally according to the ENC dataset. After investigating their translations, we came to the conclusion that the first three MWEs have word-for-word translations in most languages. Hence, they disagree with our original hypothesis that the presence of word-for-word translations is a strong indicator of compositionality. The word-for-word translations might be because of the fact that they have both compositional and non-compositional senses, or because they are calques (loan translations).\(^8\) However, we have tried to avoid such problems with calques by using translations into several languages.

For *ivory tower* ("a state of mind that is discussed as if it were a place")\(^9\) we noticed that we have a direct translation in 13 languages. Other languages have indirect translations. By checking the direct translations, we noticed that, in French, the MWE is translated to *tour* and *tour d’ivoire*. A noisy (wrong) translation of *tour* "tower" resulted in wrong indirect translations for *ivory tower* and an inflated estimate of compositionality.

We repeat the same error analysis for the EVPC dataset. The items with a high difference between the human annotation and our scores are: *carry out*, *drop out*, *get in*, *carry away*, *wear down* and *turn on*. All of theses items are annotated as non-compositional. These VPCs also have a compositional sense beside the non-compositional meaning. Also, as with the ENC dataset, we have problems of calques. For example *drop out* when translated to German (*ausfallen*) includes the word *fallen*, which is one of the translations of *drop*.

---

\(^8\)Loan translation is “the adoption by one language of a phrase or compound word whose components are literal translations of the components of a corresponding phrase or compound in a foreign language” (http://www.thefreedictionary.com/loan+translation)

\(^9\)This definition is from Wordnet 3.1.
4.1.11 Summary of the Translation-based String Similarity

In this study, we proposed a method to predict MWE compositionality based on the translation of the MWE and its component words into multiple languages. We used string similarity measures between the translations of the MWE and each of its components to predict the relative degree of compositionality. Among the four similarity measures that we experimented with, LCS and SW were found to be superior to edit distance-based methods. Our best results were found to be competitive with state-of-the-art results using vector-based approaches, and were also shown to complement state-of-the-art methods.

We compared our work with a distributional similarity approach, which was also the state-of-the-art. This distributional approach requires MWE-specific identification, which is difficult for some MWE types such as English verb particles. Our string similarity based approach does not require identification of MWEs in a large corpus and is less computationally expensive than distributional similarity based methods. By combining our method with distributional similarity, we showed that our proposed method complements the methods based on distributional similarity.

Finally, we proposed an unsupervised approach and showed that it is comparable to the supervised approach. This shows that the unsupervised approach can be used when an annotated dataset is not available.

In the next section, we will propose a method to measure the distributional similarity between the translations of MWEs and their components using multiple target languages. We will also combine the translation string similarity approach and translation distributional similarity approach, and show that they complement each other.
Chapter 4: Predicting the Compositionality Degree of MWEs

4.2 Translation Distributional Similarity Approach

Distributional similarity (DS) is a measure to approximately compute the similarity between words, which is based on the distributional hypothesis: similar words have similar distributional properties in a corpus. DS has also been used in predicting the compositionality of MWEs by measuring the semantic similarity between the MWE and its components (Schone and Jurafsky 2001; Baldwin et al. 2003; Reddy et al. 2011) (See Section 2.4.3 for more details on DS). Despite the effectiveness of DS, we discussed several drawbacks in Section 2.4.3, such as identification problems of non-contiguous MWEs, or predicting the different word forms of MWEs in morphologically-rich languages. These drawbacks make DS not easily applicable to every language.

In this section, we propose to use DS to measure the semantic similarity between the translations of MWEs and their components in multiple languages rather than only the source language. There are two hypotheses behind this work: First, for those hard to identify types of MWEs (such as English verb particle constructions), measuring the distributional similarity of the translations in other languages improves the compositionality prediction. Second, using multiple languages rather than only one leads to a better prediction of compositionality. We evaluate our work on three different datasets: English noun compounds, English verb particle constructions and German noun compounds. For a more general approach, we do not perform any language-specific pre-processing.

In this section, first, details of our chosen parameters and corpus preparation step for measuring the distributional similarity are given. We then present our findings and
Context window Correlation ($r$)

<table>
<thead>
<tr>
<th>Sentence</th>
<th>0.425</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window N=3</td>
<td>0.175</td>
</tr>
<tr>
<td>Window N=3 considering words positions</td>
<td>0.031</td>
</tr>
</tbody>
</table>

Table 4.12: Results of distributional similarities using 10 best languages on ENC dataset (N is window size)

results over English noun compound, English verb particle construction and German noun compound datasets introduced in Section 3.1. We further show that string similarity complements the distributional similarity approach.

### 4.2.1 Calculating Distributional Similarity

In order to be consistent across all languages and be as language-independent as possible, we calculate distributional similarity\(^\text{10}\) in the following manner for a given language.

Tokenisation is based on whitespace delimiters and punctuation; no lemmatisation or case-folding is carried out. Token instances of a given MWE or component word are identified by full-token $n$-gram matching over the token stream. We assume that all full stops and equivalent characters for other orthographies are sentence boundaries, and chunk the corpora into (pseudo-)sentences on the basis of them. For each language, we identify the 51st–1051st most frequent words, and consider them to be content-bearing words, in the manner of Schütze (1997). This is based on the assumption that the top-50 most frequent words are stop words, and not a good choice of word for calculating distributional similarity over. That is not to say that we can’t

\(^{10}\text{See Section 2.4.3 for an introduction of distributional similarity.}\)
calculate the distributional similarity for stop words, however they are simply not used as the dimensions in our calculation of distributional similarity.

We form a vector of content-bearing words across all token occurrences of the target word. Our preliminary results on selecting the best context window size are shown in Table 4.12. According to this table, for predicting the compositionality using the best 10 languages, the sentence context window results in a higher correlation. We use sentence boundaries as the context window in the rest of our experiments. According to Weeds (2003) and Padó and Lapata (2007), using dependency relations with the neighbouring words of the target word can better predict the meaning of the target word. However, in line with our assumption of no language-specific preprocessing, we just use word co-occurrence. Finally, distributional similarity is calculated over these context vectors using cosine similarity.

4.2.2 Calculating Compositionality

The procedure of calculating the compositionality is similar to what we used in Section 4.1.3: after translating the MWE and its components into multiple languages and measuring the distributional similarity between the translations of the MWE and its components (Figure 4.13), we find the best languages according to the training set. Then, we combine the scores from those best languages and finally calculate a combined compositionality score from the individual distributional similarities between each component word and the MWE. Based on our findings in Section 4.1.3, we combine the component scores using the weighted mean (as shown in Figure 4.14):

\[
\text{Compositionality} = \alpha s_1 + (1 - \alpha) s_2
\]
Chapter 4: Predicting the Compositionality Degree of MWEs

Figure 4.13: Outline of our approach to computing the distributional similarity (DS) of translations of an MWE with each of its component words, for a given target language. $score_1$ and $score_2$ are the similarity for the first and second components, respectively.

where $s_1$ and $s_2$ are the scores for the first and the second component, respectively.

We use different $\alpha$ settings for each dataset, while the values are the same as in Section 4.1.

We experiment with a range of methods for calculating compositionality, as follows:

$CS_{\text{L1}}$: calculate distributional similarity using only distributional similarity in the source language corpus.

$CS_{\text{L2N}}$: exclude the source language, and compute the mean of the distributional similarity scores for the best-$N$ target languages. The value of $N$ is selected
Chapter 4: Predicting the Compositionality Degree of MWEs

Figure 4.14: Outline of the method for combining distributional similarity scores from multiple languages, across the components of the MWE. $CS_{method}$ refers to any of the methods described in Section 4.2.2 for calculating compositionality.

$score_1$ for each language

$CS_{method}$

$s_1$

$score_2$ for each language

$CS_{method}$

$s_2$

$\alpha s_1 + (1 - \alpha)s_2$

compositionality score

according to training data, as detailed in Section 4.2.4.

$CS_{L1+L2N}$: calculate distributional similarity over both the source language ($CS_{L1}$) and the mean of the best-$N$ languages ($CS_{L2N}$), and combine via the arithmetic mean.\footnote{We also experimented with taking the mean over all the languages — target and source — but found it best to combine the scores for the target languages first, to give more weight to the source language.} By comparing this with $CS_{L1}$, we can examine the hypothesis that using multiple target languages is better than just using the source language.

$CS_{SVR(L1+L2)}$: train a support vector regressor (SVR: Smola and Schölkopf (2004)) over the distributional similarities for all 52 languages (source and target lan-
guages). Unlike what we show in Figure 4.14, in this approach we use all of the languages instead of only the $N$ best languages.

$CS_{LCS(mean)}$: calculate string similarity using the LCS-based method (Section 4.1).

$CS_{LCS(Mean)+L1}$: calculate the mean of the string similarity ($CS_{LCS(mean)}$) and distributional similarity in the source language.

$CS_{all}$: calculate the mean of the string similarity ($CS_{LCS(mean)}$) and distributional similarity scores ($CS_{L1}$ and $CS_{L2N}$).

$$CS_{all} = \frac{CS_{LCS(mean)} + CS_{L1} + CS_{L2N}}{3}$$

4.2.3 Results

In this section, we evaluate our work on the three introduced datasets in Section 3.1: (1) English noun compounds (ENCs), which are easy to identify in a corpus, as they are not as flexible as other types of MWEs (Section 2.2.2); (2) English verb particles (EVPCs), which are hard to identify as they can be used in a split form (see Section 2.2); and (3) German noun compounds, which are significant in being a different language than that of our other two datasets, and also in that German has relatively rich morphology, which we expect to impact on the identification of both the MWE and the component words (a shortcoming of distributional similarity as discussed in Section 2.4.3).

All experiments are carried out using 10-times-10-fold cross-validation (Section 3.4.1). In the case of $CS_{L2N}$ and other methods that make use of it (i.e. $CS_{L1+L2N}$ and $CS_{all}$),
Figure 4.15: Histograms displaying how many times a given $N$ is selected as the best number of languages over ENC, EVPC and GNC dataset.

the languages selected for a given training fold are then used to compute the compositionality scores for the instances in the test set.

4.2.4 Selecting Target Languages

We experiment with two approaches for combining the compositionality scores from multiple target languages.

First, in $CS_{L2N}$ (and then $CS_{L1+L2N}$ and $CS_{all}$ that build off it), we use training data to rank the target languages according to Pearson’s correlation between the predicted compositionality scores and the gold-standard compositionality judgements. Based on this ranking, we take the best-$N$ languages, and combine the individual compositionality scores by taking the arithmetic mean. We select $N$ by determining
### Table 4.13: The 5 best languages along with the language family and the number of times each language is selected in our 10×10 cross-validation folds setup, for ENC, EVPC and GNC.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Language</th>
<th>Frequency</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENC</td>
<td>Italian</td>
<td>100</td>
<td>Romance</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>99</td>
<td>Romance</td>
</tr>
<tr>
<td></td>
<td>German</td>
<td>86</td>
<td>Germanic</td>
</tr>
<tr>
<td></td>
<td>Vietnamese</td>
<td>83</td>
<td>Viet-Muong</td>
</tr>
<tr>
<td></td>
<td>Portuguese</td>
<td>62</td>
<td>Romance</td>
</tr>
<tr>
<td>EVPC</td>
<td>Bulgarian</td>
<td>100</td>
<td>Slavic</td>
</tr>
<tr>
<td></td>
<td>Breton</td>
<td>100</td>
<td>Celtic</td>
</tr>
<tr>
<td></td>
<td>Occitan</td>
<td>100</td>
<td>Romance</td>
</tr>
<tr>
<td></td>
<td>Indonesian</td>
<td>100</td>
<td>Indonesian</td>
</tr>
<tr>
<td></td>
<td>Slovenian</td>
<td>100</td>
<td>Slavic</td>
</tr>
<tr>
<td>GNC</td>
<td>Polish</td>
<td>100</td>
<td>Slavic</td>
</tr>
<tr>
<td></td>
<td>Lithuanian</td>
<td>99</td>
<td>Baltic</td>
</tr>
<tr>
<td></td>
<td>Finnish</td>
<td>74</td>
<td>Uralic</td>
</tr>
<tr>
<td></td>
<td>Bulgarian</td>
<td>72</td>
<td>Slavic</td>
</tr>
<tr>
<td></td>
<td>Czech</td>
<td>40</td>
<td>Slavic</td>
</tr>
</tbody>
</table>

The value that optimises the correlation over the training data. In other words, the selection of $N$, and accordingly the best-$N$ languages, are based on nested cross-validation over training data, independently of the test data for that iteration of cross-validation.

Figures 4.15(a), 4.15(b) and 4.15(c) are histograms of the number of times each $N$ (i.e., the number of best languages) is selected over 100 folds on ENC, EVPC and GNC, respectively. From the histograms, $N = 6$, $N = 15$ and $N = 2$ are the most commonly selected settings for ENC, EVPC and GNC, respectively. That is, multiple languages are generally used, but more languages are used for English VPCs than either of the compound noun datasets. The 5 most-selected languages for ENC, EVPC and GNC are shown in Table 4.13. As we can see, there are some languages
<table>
<thead>
<tr>
<th>Method</th>
<th>Summary of the Method</th>
<th>ENC</th>
<th>EVPC</th>
<th>GNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CS_{L1}$</td>
<td>Source language</td>
<td>0.700</td>
<td>0.177</td>
<td>0.141</td>
</tr>
<tr>
<td>$CS_{L2N}$</td>
<td>Best-N target languages</td>
<td>0.434</td>
<td>0.398</td>
<td>0.113</td>
</tr>
<tr>
<td>$CS_{L1+L2N}$</td>
<td>Source + best-N target languages</td>
<td>0.725</td>
<td>0.312</td>
<td>0.178</td>
</tr>
<tr>
<td>$CS_{SVR(L1+L2)}$</td>
<td>SVR (Source + all 51 target languages)</td>
<td><strong>0.744</strong></td>
<td><strong>0.389</strong></td>
<td><strong>0.085</strong></td>
</tr>
<tr>
<td>$CS_{LCS(mean)}$</td>
<td>String Similarity using LCS (Section 4.1)</td>
<td>0.644</td>
<td>0.385</td>
<td><strong>0.372</strong></td>
</tr>
<tr>
<td>$CS_{LCS(Mean)+L1}$</td>
<td>$CS_{LCS(mean)} + CS_{L1}$</td>
<td>0.739</td>
<td>0.360</td>
<td>0.353</td>
</tr>
<tr>
<td>$CS_{all}$</td>
<td>$CS_{L1} + CS_{L2N} + CS_{LCS(mean)}$</td>
<td>0.732</td>
<td><strong>0.417</strong></td>
<td>0.364</td>
</tr>
</tbody>
</table>

Table 4.14: Pearson’s correlation on the ENC, EVPC and GNC datasets

which are always selected for a given dataset, but equally the commonly-selected languages vary considerably between datasets.

In our second approach ($CS_{SVR(L1+L2)}$), we combine the compositionality scores from the source language and all 51 target languages into a feature vector, and train an SVR model over the data using LIBSVM (Section 3.3.1).

### 4.2.5 Experiment (I): ENCs

English noun compounds are relatively easy to identify in a corpus,\footnote{Although see Lapata and Lascarides (2003) for discussion of the difficulty of reliably identifying low-frequency English noun compounds.} because the components occur sequentially, and the only morphological variation is in noun number (singular vs. plural). In other words, the precision for our token matching method is expected to be very high, and the recall is also acceptably high. Partly as a result of the ease of identification, as shown in Table 4.14, we get a high correlation of $r = 0.700$ for $CS_{L1}$ (using only source language data). Using only target languages ($CS_{L2N}$), the results drop to $r = 0.434$, but when we combine the two ($CS_{L1+L2N}$), the correlation is higher than using only source or target language data, at $r = 0.725$. When we
combine all languages using SVR, the results rise slightly higher again to \( r = 0.744 \), which is slightly above the correlation of the string similarity approach \((CS_{LCS(\text{mean})})\), which we introduced in Section 4.1 and its combination with distributional similarity \((CS_{LCS(\text{Mean})} + L1)\). These last two results support our hypothesis that using the distributional similarity and string similarity of translations in multiple languages can improve the prediction of compositionality. The results for string similarity on its own \((CS_{LCS(\text{mean})}, r = 0.644)\) are slightly lower than those using only source language distributional similarity, but when combined with \(CS_{L1+L2N}\) (i.e. \(CS_{\text{all}}\)) there is a slight rise in correlation (from \( r = 0.725 \) to \( r = 0.732 \)).

4.2.6 Experiment (II): EVPCs

One area of particular interest with this dataset is that English VPCs can occur in “split” form (e.g. put your jacket on), which will complicate identification, and the verb component will often be inflected and thus not match under our identification strategy (for both VPCs and the component verbs). Also there is sometimes an ambiguity in split form, e.g. consider put on in put your jacket on the chair. In the previous example put on is an MWE while in the second example, put on is a simple verb with prepositional phrase and not an instance of an MWE. Moreover, if we adopt a conservative identification method, the number of token occurrences will be limited and the distributional scores may not be reliable.

As mentioned before, English VPCs are hard to identify. VPC components may not occur sequentially, and even when they do occur sequentially, they may not be a VPC. As such, our simplistic identification method has low precision and recall (hand
### Table 4.15: Results (%) for the binary compositionality prediction task on the EVPC dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score ($\beta = 1$)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bannard et al. (2003)</td>
<td>60.8</td>
<td>66.6</td>
<td>63.6</td>
<td>60.0</td>
</tr>
<tr>
<td>$CS_{LCS(mean)}$ (Section 4.1.2)</td>
<td><strong>86.2</strong></td>
<td>71.8</td>
<td><strong>77.4</strong></td>
<td>69.3</td>
</tr>
<tr>
<td>$CS_{all}$</td>
<td>79.5</td>
<td><strong>89.3</strong></td>
<td><strong>82.0</strong></td>
<td><strong>74.5</strong></td>
</tr>
</tbody>
</table>

There is no question that this is a contributor to the low correlation for the source language method ($CS_{L1}$; $r = 0.177$). When we use target languages instead of the source language ($CS_{L2N}$), the correlation jumps substantially to $r = 0.398$.

When we combine English and the target languages ($CS_{L1+L2N}$), the results are actually lower than just using the target languages, because of the high weight on the source language, which is not desirable for VPCs, based on the source language results. Even for $CS_{SVR(L1+L2)}$, the results ($r = 0.389$) are slightly below the target language-only results. This suggests that when predicting the compositionality of MWEs which are hard to identify in the source language, it may actually be better to use target languages only. The results for string similarity ($CS_{LCS(mean)}$: $r = 0.385$) are similar to those for $CS_{L2N}$. However, as with the ENC dataset, when we combine string similarity and distributional similarity ($CS_{all}$), the results improve, and we achieve the best correlation for the dataset.

In Table 4.15, we present classification-based evaluation over a subset of EVPC, binarising the compositionality judgements in the manner of Bannard et al. (2003). Our $CS_{all}$ method achieves the highest results in terms of overall F-score and accuracy.
4.2.7 Experiment (III): GNCs

German is a morphologically-rich language, with marking of number and case on nouns. Given that we do not perform any lemmatization or other language-specific preprocessing, we inevitably achieve low recall for the identification of noun compound tokens, although the precision should be nearly 100%. Partly because of the resultant sparseness in the distributional similarity method, the results for CS\textsubscript{L1} are low ($r = 0.141$), although they are lower again when using target languages ($r = 0.113$). However, when we combine the source and target languages (CS\textsubscript{L1+L2N}) the results improve to $r = 0.178$, which again confirms that combining distributional similarity information from multiple languages helps. The results for CS\textsubscript{SVR(L1+L2)}, on the other hand, are very low ($r = 0.085$). Ultimately, simple string similarity achieves the best results for the dataset ($r = 0.372$), and this result actually drops slightly when combined with the distributional similarities (CS\textsubscript{LCS(Mean)+L1} and CS\textsubscript{all}).

To better understand the reason for the lacklustre results using SVR, we carried out error analysis and found that, unlike the other two datasets, about half of the target languages return scores which correlate negatively with the human judgements. Further analysis reveals that 32 (63%) target languages for ENC, 25 (49%) target languages for EVPC, and only 5 (10%) target languages for GNC have a correlation of $r \geq 0.1$ with gold-standard compositionality judgements. On the other hand, 8 (16%) target languages for ENC, 2 (4%) target languages for EVPC, and no target languages for GNC have a correlation of $r \leq -0.1$. When we filter languages with very low/negative correlation from the data, the score for SVR improves appreciably. For example, over the best-3 languages overall, we get a correlation score of $r = 0.179$, \ldots
which is slightly higher than $CS_{L1+L2N}$.

We further investigated the reason for getting very low and sometimes negative correlations with many of our target languages. We noted that about 24% of the German noun compounds in the dataset do not have entries in PanLex. This contrasts with ENC where only one instance does not have an entry in PanLex, and EVPC where all VPCs have translations in at least one language in PanLex. We experimented with using string similarity scores in the case of such missing translations. The results for $CS_{SVR(L1+L2)}$ rose to $r = 0.269$, although this is still below the correlation for just using string similarity.

### 4.2.8 Summary of Translation Distributional Similarity Approach

In this section, we proposed a method to predict the compositionality of MWEs based on monolingual distributional similarity between the MWE and each of its component words, under translation into multiple target languages. We showed that using translation and multiple target languages enhances compositionality modeling.

Our experiments reveal that in case of hard to identify MWEs, such as English verb particle constructions, measuring distributional similarity of translations results in a better compositionality prediction compared to using only the source language.

Our further investigation shows strong complementarity between the distributional approach and the string similarity approach which was introduced in Section 4.1. This complementary is due to using two different resources, with different strengths and shortcomings: a multilingual dictionary (PanLex), and multiple mono-
lingual corpora (Wikipedia).

4.3 Word Embeddings Approach

As mentioned in Section 2.4.3, there has been a recent surge of interest in learning and using word embeddings in a wide range of NLP applications. This section discusses the first attempt, at the time of completing this research, to bring together the work on word embedding-style distributional analysis with compositionality prediction of MWEs. Since word embeddings have been shown to improve various NLP tasks, our first research question here is:

1: Are word embeddings superior to conventional count-based models of distributional similarity, in the context of compositionality prediction.

We observed that the word embedding approaches require parameter settings, such as the size of context window or the vector size, in order to train the model. This brought up the second research question:

2: How sensitive to parameter optimisation are different word embedding approaches?

Recent work on multi-prototype word embeddings proposed to model polysemy of the words lead us to the third research question:

3: Are multi-prototype word embeddings empirically superior to single-prototype word embeddings?
In this work, we estimate the compositionality of an MWE based on the similarity between the expression and its component words in vector space. We use three different vector-space models: (1) a simple count-based model of distributional similarity (as introduced in Section 4.2.1 but only in the source language); (2) word embeddings based on word2vec (Mikolov et al. 2013a); and (3) a multi-sense skip-gram model that, unlike the previous two models, is able to learn multiple embeddings per target word (or MWE) (Neelakantan et al. 2014). For all three models, we first greedily pre-tokenise the corpus to represent each MWE as a single token, similarly to Baldwin et al. (2003). Same as Section 4.2.1, we apply the constraint that no language-specific pre-processing can be applied to the training corpus, in order to make the method maximally language independent. As such, we cannot perform any form of lemmatisation, and MWE identification takes the form of simple string match for concatenated instances of the component words, naively assuming that all occurrences of that word combination are MWEs. We detail each of the word-embedding methods below.

### 4.3.1 word2vec

Our first word embeddings based method is based on the recurrent neural network language model (RNNLM) approach to learning word embeddings of Mikolov et al. (2013a), using the word2vec package. As mentioned in Section 2.4.3, word2vec can be used in two forms: (1) a continuous bag-of-words (“CBOW”) model and (2) a continuous skip-gram model (“C-Skip”). word2vec generates a vector of fixed dimensionality $d$ for each pre-tokenised word/MWE type with frequency above
a certain threshold in the training corpus.

To measure the similarity of the MWE vector and the component word vectors, we considered two different approaches.

The first approach (\(comp_1\)) is based on Reddy et al. (2011) and Schulte im Walde et al. (2013). The similarity between the MWE and each of its components is measured, and the overall compositionality of the MWE is computed by combining the similarity scores for the two components as in Section 4.1.3.

\[
comp_1(MWE) = \alpha \text{sim}(MWE, C_1) + (1 - \alpha)\text{sim}(MWE, C_2)
\]

where \(MWE\) is the vector associated with the MWE, \(C_i\) is the vector associated with the \(i\)th component word of the MWE, \(\text{sim}\) is a vector similarity function, and \(\alpha \in [0,1]\) is a weight parameter.

We also experimented with the approach from Mitchell and Lapata (2010), where \(MWE\) is compared directly with a composed vector of the component words, based on vector addition:\[13\]

\[
comp_2(MWE) = \text{sim}(MWE, C_1 + C_2)
\]

where \(MWE\) is the vector associated with the MWE, \(C_i\) is the vector associated with the \(i\)th component word of the MWE, \(\text{sim}\) is a vector similarity function (here, cosine similarity).

\[13\] We also experimented with vector multiplication, but found it to perform poorly compared to the other approaches.
4.3.2 Multi-Sense Skip-gram Model

One potential shortcoming of word2vec is that it generates a single word embedding for each word, irrespective of the relative polysemy of the word. Neelakantan et al. (2014) proposed a method motivated by word2vec, which efficiently learns multiple embeddings per word/MWE. We refer to this approach as the multi-sense skip-gram (MSSG) model. We once again compose the resultant vectors with $comp_1$ and $comp_2$, but modify the formulation slightly to handle the variable number of vectors for each word/MWE, by searching over the cross-product of vectors in each $sim$ calculation and taking the maximum in each case. We initially set the number of embeddings to 2 in our MSSG experiments — in keeping with the findings in Neelakantan et al. (2014) — but come back to examine the impact of the number of embeddings on compositionality prediction in Section 4.3.7.

4.3.3 Results

For all experiments, we train our models over raw text Wikipedia corpora for either English or German, depending on the language of the dataset (Section 3.1). Word-tokenisation was performed based on simple whitespace delimitation, after which we greedily identified all string occurrences of the MWEs in each of our datasets and combined them into a single token.\footnote{For English, a single model was trained over a corpus containing both ENC and EVPC tokens.} The evaluation metric for all experiments is Pearson correlation, which measures the linear correlation between annotations and our systems’ outputs. This metric is selected since the task of predicting the degree of compositionality is a regression task rather than classification (see Section 3.4.2).
Chapter 4: Predicting the Compositionality Degree of MWEs

Table 4.16: Parameter settings of the examined methods (\(d\) =vector dimensionality and \(w\) =window size). The \texttt{word2vec}-1 and \texttt{MSSG}-1 parameters were set based on the default recommendations of Mikolov et al. (2013a) and Neelakantan et al. (2014). The settings were changed in other examined methods, accordingly.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{word2vec}-1</td>
<td>(d = 500, \text{C-Skip})</td>
</tr>
<tr>
<td>\texttt{word2vec}-2</td>
<td>(d = 500, \text{CBOW})</td>
</tr>
<tr>
<td>\texttt{word2vec}-3</td>
<td>(d = 1000, \text{C-Skip})</td>
</tr>
<tr>
<td>\texttt{word2vec}-4</td>
<td>(d = 1000, \text{CBOW})</td>
</tr>
<tr>
<td>\texttt{MSSG}-1</td>
<td>(d = 300, w = 5)</td>
</tr>
<tr>
<td>\texttt{MSSG}-2</td>
<td>(d = 600, w = 5)</td>
</tr>
<tr>
<td>\texttt{MSSG}-3</td>
<td>(d = 600, w = 10)</td>
</tr>
</tbody>
</table>

The word embedding approaches are unable to generate vector representations for tokens which occur with frequency below a fixed cutoff. For a frequency threshold of 15, the total numbers of \texttt{ENCs}, \texttt{EVPCs} and \texttt{GNCs} for which we were unable to generate word embeddings were 3, 0 and 25, respectively. In \texttt{GNC} case, our simple tokenisation strategy of full-token \(n\)-gram matching to identify MWEs, and lack of lemmatization in line with no language-specific preprocessing resulted in 25 instances with frequency lower than the threshold.

In order to generate a compositionality prediction back-off for the small numbers of MWEs in this category, we assign a default value, which is the mean of computed compositionality scores for other instances.\(^{15}\) This specific approach was adopted in order to make the results comparable to previous/state-of-the-art results.

As a baseline, we use the unsupervised translation string similarity approach of Section 4.1.9. We further include a linear combination of the string similarity method with each of the various approaches based on word embeddings.

\(^{15}\)We also experimented with using the string similarity approach as a back-off, which resulted in marginally lower results than what is reported in Table 4.17.
### Table 4.17: Pearson’s correlation ($r$) for the different methods (Table 4.16) over the three datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>$comp_1$</th>
<th>$comp_1$+SS</th>
<th>$comp_2$</th>
<th>$comp_2$+SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENC</td>
<td>WORD2VEC-1</td>
<td>0.628</td>
<td>0.731</td>
<td>0.632</td>
<td>0.729</td>
</tr>
<tr>
<td>ENC</td>
<td>WORD2VEC-2</td>
<td>0.696</td>
<td>0.762</td>
<td>0.710</td>
<td>0.765</td>
</tr>
<tr>
<td>ENC</td>
<td>WORD2VEC-3</td>
<td>0.636</td>
<td>0.733</td>
<td>0.648</td>
<td>0.735</td>
</tr>
<tr>
<td>ENC</td>
<td>WORD2VEC-4</td>
<td>0.717</td>
<td>0.766</td>
<td>0.736</td>
<td><strong>0.771</strong></td>
</tr>
<tr>
<td>ENC</td>
<td>MSSG-1</td>
<td>0.640</td>
<td>0.733</td>
<td>0.624</td>
<td>0.729</td>
</tr>
<tr>
<td>ENC</td>
<td>MSSG-2</td>
<td>0.615</td>
<td>0.725</td>
<td>0.594</td>
<td>0.717</td>
</tr>
<tr>
<td>ENC</td>
<td>MSSG-3</td>
<td>0.614</td>
<td>0.719</td>
<td>0.631</td>
<td>0.725</td>
</tr>
<tr>
<td>ENC</td>
<td></td>
<td>Distributional similarity</td>
<td>.714</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENC</td>
<td></td>
<td>String similarity</td>
<td>.642</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENC</td>
<td></td>
<td>State-of-the-art (SS+DS in Section 4.1)</td>
<td>.744</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVPC</td>
<td>WORD2VEC-1</td>
<td>0.289</td>
<td>0.373</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EVPC</td>
<td>WORD2VEC-2</td>
<td>0.293</td>
<td>0.364</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EVPC</td>
<td>WORD2VEC-3</td>
<td>0.289</td>
<td>0.375</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EVPC</td>
<td>WORD2VEC-4</td>
<td>0.289</td>
<td>0.363</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EVPC</td>
<td>MSSG-1</td>
<td>0.309</td>
<td>0.403</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EVPC</td>
<td>MSSG-2</td>
<td>0.294</td>
<td>0.388</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EVPC</td>
<td>MSSG-3</td>
<td>0.273</td>
<td>0.380</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EVPC</td>
<td></td>
<td>Distributional similarity</td>
<td>.165</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVPC</td>
<td></td>
<td>String similarity</td>
<td>.323</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVPC</td>
<td></td>
<td>State-of-the-art (C-S_all in Section 4.2)</td>
<td>.417</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNC</td>
<td>WORD2VEC-1</td>
<td>0.393</td>
<td><strong>0.460</strong></td>
<td>0.321</td>
<td>0.445</td>
</tr>
<tr>
<td>GNC</td>
<td>WORD2VEC-2</td>
<td>0.400</td>
<td>0.455</td>
<td>0.361</td>
<td>0.447</td>
</tr>
<tr>
<td>GNC</td>
<td>WORD2VEC-3</td>
<td>0.341</td>
<td>0.436</td>
<td>0.282</td>
<td>0.427</td>
</tr>
<tr>
<td>GNC</td>
<td>WORD2VEC-4</td>
<td>0.371</td>
<td>0.435</td>
<td>0.349</td>
<td>0.437</td>
</tr>
<tr>
<td>GNC</td>
<td>MSSG-1</td>
<td>0.181</td>
<td>0.367</td>
<td>0.122</td>
<td>0.346</td>
</tr>
<tr>
<td>GNC</td>
<td>MSSG-2</td>
<td>0.202</td>
<td>0.375</td>
<td>0.146</td>
<td>0.349</td>
</tr>
<tr>
<td>GNC</td>
<td>MSSG-3</td>
<td>0.155</td>
<td>0.358</td>
<td>0.101</td>
<td>0.338</td>
</tr>
<tr>
<td>GNC</td>
<td></td>
<td>Distributional Similarity</td>
<td>.140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNC</td>
<td></td>
<td>String Similarity</td>
<td>.343</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNC</td>
<td></td>
<td>State-of-the-art (Schulte im Walde et al. 2013)</td>
<td>.450</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.17 shows the results for the various methods, over a range of hyper-parameter settings, shown in Table 4.16, for each of WORD2VEC (vector dimensionality $d$; we also present results for CBOW vs. C-Skip) and MSSG (vector dimension-
ality \( d \) and window size \( w \)), informed by the experimental results in the respective publications. Note that for EVPC, we don’t use the vector for the particle, as discussed in Section 3.1.2; as such, there are no results for \( \text{comp}_2 \). For \( \text{comp}_1 \), \( \alpha \) is set to 1.0 for EVPC, and 0.7 for both ENC and GNC, also based on the findings of translation string similarity approach.

The results indicate that the approaches using both word2vec and MSSG outperform simple distributional similarity by a substantial margin. Further, over a variety of parameterizations, they surpass or perform as good as the state-of-the-art methods, when combined with string similarity, for ENC, EVPC and GNC.

Note that in each case, the state-of-the-art is achieved using varying levels of supervision over labelled data (ENC and EVPC) or language-specific pre-processing (GNC), whereas the word embedding methods use no labeled data and no pre-processing step is performed. As such, the answer to the first question of ‘whether the word embeddings are superior to conventional count-based models of distributional similarity, in the context of compositionality prediction’ would appear to be a resounding yes.

Looking to the second question, which was about the sensitivity of word embedding approaches to parameter optimisation, we conclude that the models are remarkably insensitive to hyper-parameter optimisation for EVPC, but there are slight deviations in the results for ENC and GNC. Having said that, they are largely between the different word embedding approaches, and the results for a given approach under different parameter settings are relatively stable. A large part of the cause of the drop in results and greater parameter sensitivity over GNC is the lower token frequencies, through a combination of the Wikipedia corpus being markedly smaller and our naive
Table 4.18: The frequency information for each dataset excluding the zero values.

tokenisation strategy having low recall over German due to the richer morphology. As such, the answer would appear to be a tentative “relatively insensitive, assuming high token frequencies”.

Finally, looking to the third question on the superiority of multi-prototype word embeddings to single-prototype word embeddings, the results show little separating \texttt{word2vec} and MSSG over \texttt{ENC} and \texttt{EVPC}, but over \texttt{GNC} dataset, \texttt{word2vec} had a clear advantage. Better performance of MSSG over \texttt{EVPC} suggests that multi-prototype representation is more appropriate for this dataset, where as we will show in Section 5.2.4, English verb particle constructions have a higher level of polysemy.

According to the results, the string similarity approach complements all word-embedding approaches. This is because the string similarity-based approach is not based on any corpus, and is thus not biased by the frequency of token instances in the corpus.

In the following experiments, the effect of some factors such as frequency of token instances in the corpus, context window size and vector size on \texttt{word2vec}, number of prototypes in MSSG and various MSSG operators to combine the components embeddings are studied.
Figure 4.16: Token frequency effect on corpus-based approaches. \textsc{word2vec}-i and \textsc{MSSG}-i represent the ith parameter settings of each method shown in Table 4.17.

### 4.3.4 Frequency Effect

In this section, we investigate the effect of MWE’s frequency in Wikipedia corpus on the three corpus-based approaches: distributional similarity (DS), \textsc{word2vec}, and \textsc{MSSG}. The data samples in each dataset are divided into two equal size groups, called low and high, based on the median of MWE frequencies (of each dataset) in the corpus (See Table 4.18). Figure 4.16 shows the Pearson’s correlation between
predicted scores and human judgments in each group.

According to the results, we almost always observe higher scores for the high group, which shows that corpus-based models perform better on high frequency MWEs. As a result, these models can predict the compositionality of MWEs with higher frequency better than less frequent MWEs.

This finding also explains why Pearson’s correlation for ENC is considerably higher than that for EVPC and GNC in Table 4.17. The reason is that we find fewer MWE instances due to the identification method (i.e., simple string match) for EVPC and GNC.

4.3.5 Window size

In this section, we study the effect of context window size as one of the parameters required to be set in WORD2VEC. This parameter chooses the number of neighbouring words to consider for training the model. Figure 4.17 shows how the context window size affects our prediction on the three datasets. Note that the results shown in Figure 4.17 are limited to WORD2VEC-1 results without string similarity (SS).

According to our results, window size does not have a significant effect when predicting the compositionality of MWEs. We also observe that WORD2VEC can efficiently weight irrelevant context words in a way that increasing the context window size will not decrease the correlation scores.
Figure 4.17: The effect of window size parameter on word2vec

4.3.6 Vector size

This section describes the effect of vector size on predicting the degree of MWE compositionality using word2vec. Here, vector size refers to the dimension of the projected latent space. Figure 4.18 shows how the correlation score changes when the latent space dimensionality is increased.\textsuperscript{17} The results suggest that smaller vector sizes (low dimensionality latent spaces) of 50 dimensions or lower cannot capture the information necessary to model word meanings for compositionality prediction. Therefore, we observe low correlation scores for all datasets with a vector size of

---

\textsuperscript{16}See Section 2.4.3 for more information about the context window

\textsuperscript{17}Note that word2vec allows up to 1000 dimensions.
50. According to Figure 4.18, the highest correlation for all datasets is when the vector size equals 100. Increasing the vector size beyond 100 shows no improvement in correlation. However, although the correlation slightly decreases by increasing the vector size, word2vec does not suffer significantly from the curse of dimensionality\textsuperscript{18} when we increase the dimensionality.

\textsuperscript{18}Curse of dimensionality occurs when the volume of the space arises and as a result the data instances become sparse. When the curse of dimensionality occurs, statistical analysis becomes less reliable (Keogh and Mueen 2010).
4.3.7 Multiple prototype effect

In Table 4.17, the number of embeddings for MSSG was set to 2 prototypes, based on the default recommendations of Neelakantan et al. (2014). To investigate the impact of this parameter on our results, we retrained MSSG over the range [1, 6] and reran our experiments for each set of embeddings over the three datasets (without string similarity, to isolate the effect of the number of embeddings), as shown in Figure 4.19. For both English datasets (ENC and EVPC), setting the number of prototypes to a value higher than 2 boosts the results slightly, with 5 prototypes appearing to be the optimal value. For the German dataset (GNC), on the other hand, the best results are actually achieved for a single prototype. This is inline with our observation in Table 4.17 that word2vec is superior to MSSG for this dataset.
Table 4.19: Pearson’s correlation \( (r) \) for the different operators to combine vector representations provided by MSSG, over the three datasets. SS stands for the string similarity approach.

### MSSG operator

MSSG provides multiple prototypes/vectors for each word/MWE; In this section, we explore various operators to use the provided vector representations of the MWEs and their components in predicting the degree of compositionality.

In our first experiment, the cosine similarity between all possible combinations of MWE prototypes and its component prototypes is measured. Then one of the following operators is applied to measure the degree of compositionality:

1. **MAX**, where the cosine similarity between all prototypes of the MWE and its components is measured and the maximum similarity is selected as the compositionality degree. Note that in all previous experiments, we used the MAX operator.

2. **MEAN**, where the average of the cosine similarity between all prototypes combi-
nations is measured.

(3) MEDIAN, where, we calculate the median of similarity scores between all prototypes combinations. MEDIAN is studied because of being more robust against outliers.

In the second experiment, in line with Mitchell and Lapata (2010), the prototypes for each token are combined using one of the two operators of ADD and MULT.

(1) ADD computes the vector addition of each word/MWE prototype.

(2) MULT measures the component-wise multiplication of prototypes for each word/MWE.

After using each of the two above operators, there will be one vector for each word/MWE. We will, then, apply cosine similarity to compute the compositionality score.

Table 4.19 provides the results of applying each operator over the three datasets. According to the results, MAX performs best for ENC and EVPC when using either \( \text{comp}_1 \) or \( \text{comp}_2 \), while for GNC ADD and MEDIAN perform best. Unlike the findings of Mitchell and Lapata (2010), we observe that MULT performs appreciably worse than the other operators, supporting the earlier claim that it is not suited to this task.

4.3.9 Summary of the Word Embeddings Approach

We presented the first approach to using word embeddings to predict the compositionality of MWEs. We showed that this approach, in combination with information from string similarity, surpassed, or was competitive with, the current state-of-the-art on three compositionality datasets.
We explore three questions relative to three compositionality prediction datasets spanning two MWE construction types (noun compounds and verb particle constructions) and two languages (English and German), and arrive at the following conclusions: (1) consistent with recent work over other NLP tasks, word embeddings are superior to count–based models of distributional similarity (and also translation-based string similarity); (2) the results are relatively stable under parameter optimisation for a given word embedding learning approach; and (3) based on simple approaches to composition, single word embeddings are empirically slightly superior to multi-prototype word embeddings overall.

The rest of our findings are:

- Corpus-based methods can predict more frequent MWEs better than MWEs with low frequency.

- Context window size, in the task of predicting the compositionality of MWEs, does not have a marked effect.

- Low dimensionality (i.e., 50) results in low correlation scores, while high dimensionality (i.e., 1000) does not suffer from the curse of dimensionality (i.e., the data sparsity problem, which often occurs when the representation size becomes overly large) when using WORD2VEC.

- Unlike the findings in Mitchell and Lapata (2010), we observe that MULT is noticeably performing worse than the other operators.

- String similarity approach complements all other corpus-based approaches and vice versa. The only exceptions are distributional approaches for GNC and
Chapter 4: Predicting the Compositionality Degree of MWEs

EVPC, in which clearly distributional similarity is performing very poorly and while string similarity complements these approaches, they cannot complement string similarity approach.

4.4 Summary

In this chapter, we introduced three approaches to predicting the degree of compositionality of multiword expressions: translation string similarity, translation distributional similarity, and word embedding-based approaches. The proposed approaches are applicable to a wide range of languages, as well as any type of multiword expression, and no language-specific assumptions are made. We studied our approaches on three datasets, covering two languages (English and German) and two different types of multiword expression (noun compounds and verb particle constructions). In the next chapter, we introduce a method to determine the compositionality of MWEs by classifying them as compositional or non-compositional based on the information provided by Wiktionary users. We will use the degree of compositionality of MWEs in machine translation evaluation in Chapter 6.
Chapter 5

Detecting Non-compositional MWE Components using Wiktionary

In Chapter 4, we proposed three approaches to predict the compositionality of MWEs. This chapter proposes a simple unsupervised approach to detecting non-compositional components in multiword expressions based on Wiktionary.\textsuperscript{1} The approach makes use of the definitions, synonyms and translations provided in Wiktionary. This method is applicable to any type of MWE in any language, assuming the MWE is contained in Wiktionary.

In Section 5.1, we explain why Wiktionary is a good resource for this task. The methodology consists of definition-based similarity, synonym-based definition expansion and a translation-based approach. We introduce each approach in Section 5.2,\textsuperscript{1}

\textsuperscript{1}http://www.wiktionary.org/
and discuss the experiments and results in Section 5.3. Finally, the findings of our error analysis are explained in Section 5.4.

5.1 Introduction

Wiktionary is the major resource we use in this chapter (see Section 3.2.3 on page 67 for more details on Wiktionary). Since our goal is to develop methods which are applicable to any type of MWE across a wide range of languages, we believe Wiktionary is an appropriate resource for the following reasons.

1. As a language evolves, new MWEs emerge. When an MWE becomes popular, users will add it to Wiktionary. Compared to Wiktionary, online dictionaries and resources such as Oxford dictionaries and WordNet are updated less frequently. For example, *selfie stick* and *beer o’clock* were added to Wiktionary in 2014 and 2011, respectively, while they were entered into the Oxford Dictionary in only 2015.²

2. Some MWEs are meaningful in only a specific domain, and meaningless in other areas such as *internet of things* in Technology,³ *bear market* in Economics,⁴ *Islets of Langerhans* in Medicine,⁵ *broken chord* in Music,⁶ and *finding of fact* in Law.⁷ While, there are certainly individual glossaries for specific domains, Wiktionary provides an environment for users from different backgrounds to

²http://www.oxforddictionaries.com/words/what-s-new
³“A proposed development of the Internet in which everyday objects have network connectivity, allowing them to send and receive data.”
⁴“A market in which share prices are falling, encouraging selling.”
⁵“Groups of cells secreting insulin and glucagon.”
⁶“A chord whose notes are played in sequential descending or ascending order”
⁷“verdict”
edit and add new words and MWEs for any domain. Based on Wiktionary editorial guideline, a term should be included if it’s likely that someone would run across it and want to know what it means.

3. Wiktionary entries are not only in English, but also in more than 100 language varieties. English has over 4 million lexical entries, French and Malagasy each have over 2 million, and other languages have less than 1 million at the time of writing of this thesis. We will only consider English in our experiments in this chapter, due to its high coverage.

4. Wiktionary provides translations of words and multiword expressions to other languages, although the set of languages for which translations are provided varies greatly across lexical entries.

For the above reasons, we believe Wiktionary has the potential to become a multilingual resource, which is rich enough to support a language-independent approach to determining the compositionality of MWEs. For the rest of this chapter, we only consider English, since English has the highest number of entries at the time of writing of this thesis. However, we do not make any language specific assumptions, to ensure the method is as generally applicable as possible.

5.2 Methodology

Our basic method relies on the analysis of lexical overlap between the component words and the definitions of the MWE in Wiktionary, in the manner of Lesk (1986).

---

8https://en.wiktionary.org/wiki/Wiktionary:Criteria_for_inclusion
That is, if a given component word can be found in the definition, then we infer that the MWE carries the meaning of that component. For example, the Wiktionary definition of *swimming pool* is “an artificially constructed *pool* of water used for *swimming*”, suggesting that the MWE is compositional relative to both *swimming* and *pool*. If the MWE is not found in Wiktionary, we use Wikipedia as a backoff, and use the first paragraph of the (top-ranked) Wikipedia article as a proxy for the definition.

As detailed below, we further extend the basic method to incorporate three types of information found in Wiktionary: (1) definitions of each word in the definitions, (2) synonyms of the words in the definitions using definition expansion, and (3) translations of the MWEs and components.

### 5.2.1 Definition-based Similarity

The basic method uses Boolean lexical overlap between the target component of the MWE and a definition. A given MWE will often have multiple definitions, however, begging the question of how to combine across them, for which we propose the following three methods.

**First Definition (FirstDef):** Use only the first-listed Wiktionary definition for the MWE, based on the assumption that this is the predominant sense.

**All Definitions (AllDefs):** In the case that there are multiple definitions for the MWE, calculate the lexical overlap for each independently and take a majority vote; in the case of a tie, label the component as non-compositional.
Idiom Tag (ITag): In Wiktionary, there is facility for users to tag definitions as idiomatic.\textsuperscript{9} If, for a given MWE, there are definitions tagged as idiomatic, use only those definitions; if there are no such definitions, use the full set of definitions.

5.2.2 Synonym-based Definition Expansion

In some cases, a component is not explicitly mentioned in a definition, but a synonym does occur, indicating that the definition is compositional in that component. In order to capture synonym-based matches, we optionally look for synonyms of the component word in the definition,\textsuperscript{10} and expand our notion of lexical overlap to include these synonyms.

For example, for the MWE \textit{china clay}, the definition is \textit{kaolin}, which includes neither of the components. However, we find the component word \textit{clay} in the definition for \textit{kaolin}, as shown below:

“A fine \underline{clay}, rich in kaolinite, used in ceramics, paper-making, etc.”

This method is compatible with the three definition-based similarity methods described above, and indicated by the +SYN suffix (e.g. \texttt{FirstDef+Syn} is \texttt{FirstDef} with synonym-based expansion).

5.2.3 Translations

A third information source in Wiktionary that can be used to predict compositionality is sense-level translation data. Due to the user-generated nature of Wiktionary,

\textsuperscript{9}Although the recall of these tags is low (Muzny and Zettlemoyer 2013).
\textsuperscript{10}After removing function words, based on a stopword list.
Table 5.1: Lexical coverage of WordNet, Wiktionary and Wiktionary+Wikipedia over our two datasets.

<table>
<thead>
<tr>
<th></th>
<th>ENC</th>
<th>EVPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td>91.1%</td>
<td>87.5%</td>
</tr>
<tr>
<td>Wiktionary</td>
<td>96.7%</td>
<td>96.2%</td>
</tr>
<tr>
<td>Wiktionary+Wikipedia</td>
<td>100.0%</td>
<td>96.2%</td>
</tr>
</tbody>
</table>

the set of languages for which translations are provided varies greatly across lexical entries. Our approach is to take whatever translations happen to exist in Wiktionary for a given MWE, and where there are translations in that language for the component of interest, use the LCS-based method proposed in Section 4.1 to measure the string similarity between the translation of the MWE and the translation of the components. Unlike Section 4.1, however, we do not use development data to select the optimal set of languages in a supervised manner, and instead simply take the average of the string similarity scores across the available languages. In the case of more than one translation in a given language, we use the maximum string similarity for each pairing of an MWE and component translation.

Unlike the definition and synonym-based approach, the translation-based approach will produce real rather than binary values (see Section 5.2.4). To combine the two approaches, we discretise the scores given by the translation approach. In the case of disagreement between the two approaches, we label the given MWE as non-compositional. This results in higher recall and lower precision for the task of detecting non-compositional components.
5.2.4 An Analysis of Wiktionary Coverage

A dictionary-based method is only as good as the dictionary it is based on. In the case of MWE compositionality analysis, our primary concern is lexical coverage in Wiktionary, i.e., what proportion of a representative set of MWEs is contained in Wiktionary. We measure lexical coverage relative to the two English datasets used in this research (described in detail in Section 3.1): ENC and EVPC. In each case, we calculated the proportion of the dataset that is found in Wiktionary, Wiktionary+Wikipedia (where we back off to a Wikipedia document in the case that a MWE is not found in Wiktionary) and WordNet (Fellbaum 1998). The results are found in Table 5.1, and indicate perfect coverage in Wiktionary+Wikipedia for the ENCs, and very high coverage for the EVPCs. In both cases, the coverage of WordNet is substantially lower, although still respectable, at around 90%.

Our results are not directly comparable with those of Chapter 4, where we evaluated in terms of a regression task, modelling the overall compositionality of the MWE. In this case, the task setup is a binary classification\footnote{Because our proposed methods use Boolean lexical overlap between the target component of the MWE and a definition} task relative to each of the two components of the MWE.

For the ENC dataset (see Section 3.1), 51% of noun compounds are compositional in the first component and 60% are compositional in the second component. We generate discrete labels by discretising the component-wise compositionality scores based on the partitions \([0, 2.5]\) and \((2.5, 5]\). On average, each NC in this dataset has 1.4 senses (definitions) in Wiktionary.

The EVPC dataset (see Section 3.1) was manually annotated for compositionality
on a binary scale for each of the head verb and particle. For the 160 EVPCs, 76% are verb-compositional and 48% are particle-compositional. On average, each EVPC in this dataset has 3.0 senses (definitions) in Wiktionary.

5.3 Experiments

The baseline for each dataset takes the form of looking for a user-annotated idiom tag in the Wiktionary lexical entry for the MWE: if there is an idiomatic tag, both components are considered to be non-compositional; otherwise, both components are considered to be compositional. We expect this method to suffer from low precision for two reasons: first, the guidelines given to the annotators of our datasets might be different from what Wiktionary contributors assume to be an idiom. Second, the baseline method assumes that for any non-compositional MWE, all components must be equally non-compositional, despite the wealth of MWEs where one or more components are compositional (e.g. from the Wiktionary guidelines for idiom inclusion,12 computer chess, basketball player, telephone box).

We also compare our method with: (1) “LCS”, the string similarity-based method of Section 4.1, in which 54 languages are used; (2) “DS”, the monolingual distributional similarity method of Section 4.2; (3) “DS+DSL2”, the multilingual distributional similarity method of Section 4.2, including supervised language selection for a given dataset, based on cross-validation; and (4) “LCS+DS+DSL2”, whereby the first three methods are combined using a supervised support vector regression model. In each case, the continuous output of the model is equal-width discretised to generate

12http://en.wiktionary.org/wiki/Wiktionary:Idioms_that_survived_RFD
### Table 5.2: Compositionality prediction results over the ENC dataset, relative to the first component (the modifier noun) and the second component (the head noun). The precision and recall calculated for the first and second component differ because the compositionality judgements differ between components.

<table>
<thead>
<tr>
<th>Method</th>
<th>First Component</th>
<th></th>
<th></th>
<th>Second Component</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-score</td>
<td>Precision</td>
<td>Recall</td>
<td>F-score</td>
</tr>
<tr>
<td>Baseline</td>
<td>66.7</td>
<td>68.2</td>
<td>67.4</td>
<td>66.7</td>
<td>83.3</td>
<td>74.1</td>
</tr>
<tr>
<td>LCS</td>
<td>60.0</td>
<td>77.7</td>
<td>67.7</td>
<td>81.6</td>
<td>68.1</td>
<td>64.6</td>
</tr>
<tr>
<td>DS</td>
<td>62.1</td>
<td>88.6</td>
<td>73.0</td>
<td>80.5</td>
<td>86.4</td>
<td>71.2</td>
</tr>
<tr>
<td>DS+DSL2</td>
<td>62.5</td>
<td>92.3</td>
<td>74.5</td>
<td>78.4</td>
<td>89.4</td>
<td>70.6</td>
</tr>
<tr>
<td>LCS+DS+DSL2</td>
<td>66.3</td>
<td>87.5</td>
<td>75.4</td>
<td>82.1</td>
<td>80.6</td>
<td>70.1</td>
</tr>
<tr>
<td>FirstDef</td>
<td>59.4</td>
<td>93.2</td>
<td>72.6</td>
<td>54.2</td>
<td>88.9</td>
<td>67.4</td>
</tr>
<tr>
<td>AllDefs</td>
<td>59.5</td>
<td>100.0</td>
<td>74.6</td>
<td>52.9</td>
<td>100.0</td>
<td>69.2</td>
</tr>
<tr>
<td>ITAG</td>
<td>60.3</td>
<td>100.0</td>
<td>75.2</td>
<td>54.5</td>
<td>100.0</td>
<td>70.0</td>
</tr>
<tr>
<td>FirstDef+Syn</td>
<td>64.9</td>
<td>84.1</td>
<td>73.3</td>
<td>63.8</td>
<td>83.3</td>
<td>72.3</td>
</tr>
<tr>
<td>AllDefs+Syn</td>
<td>64.5</td>
<td>90.9</td>
<td>75.5</td>
<td>60.4</td>
<td>88.9</td>
<td>71.9</td>
</tr>
<tr>
<td>ITAG+Syn</td>
<td>64.5</td>
<td>90.9</td>
<td>75.5</td>
<td>61.8</td>
<td>94.4</td>
<td>74.7</td>
</tr>
<tr>
<td><strong>FirstDef+Syn</strong></td>
<td><strong>82.9</strong></td>
<td>85.3</td>
<td>84.1</td>
<td><strong>81.9</strong></td>
<td>80.0</td>
<td>69.8</td>
</tr>
<tr>
<td><strong>AllDefs+Syn</strong></td>
<td><strong>81.2</strong></td>
<td>88.1</td>
<td><strong>84.5</strong></td>
<td><strong>87.3</strong></td>
<td>80.6</td>
<td>73.3</td>
</tr>
<tr>
<td><strong>ITAG+Syn</strong></td>
<td><strong>81.0</strong></td>
<td>88.1</td>
<td>84.1</td>
<td><strong>88.0</strong></td>
<td>81.1</td>
<td>73.9</td>
</tr>
</tbody>
</table>

Overall, the simple unsupervised methods proposed in this paper are comparable with the unsupervised and supervised methods of Section 4.1 and Section 4.2, with ITAG (among the three methods described in Section 5.2.1) achieving the highest F-score for the ENC dataset and for the verb components of the EVPC dataset. The
Table 5.3: Compositionality prediction results over the EVPC dataset, relative to the first component (the head verb) and the second component (the particle) inclusion of synonyms boosts results in most cases.

When we combine each of our proposed methods with the string and distributional similarity methods of Section 4.1 and Section 4.2, we see substantial improvements over the comparable combined method of “LCS+DS+DSL2” in most cases, demonstrating both the robustness of the proposed methods and their complementarity with the earlier methods. It is important to reinforce that the proposed methods make no language-specific assumptions and are therefore applicable to any type of MWE and any language, with the only requirement being that the MWE of interest be listed in the Wiktionary for that language.

Overall, we conclude that if we are after unsupervised classification of the components into compositional and non-compositional, and if Wiktionary coverage in that language is high enough, this approach can perform better than the supervised approaches in Chapter 4.
5.4 Error Analysis

We analysed all items in each dataset where the system score differed from that of the human annotators. For both datasets, the majority of incorrectly-labelled items were compositional but predicted to be non-compositional by our system, as can be seen in the relatively low precision scores in Tables 5.2 and 5.3. In many of these cases, the prediction based on definitions and synonyms was compositional but the prediction based on translations was non-compositional. In such cases, we arbitrarily break the tie by labelling the instance as non-compositional, and in doing so favour recall over precision.

Some of the incorrectly-labelled ENCs have a gold-standard annotation of around 2.5, or in other words are semi-compositional. For example, the compositionality score for game in game plan is 2.82/5, but our system labels it as non-compositional; a similar thing happens with figure and the EVPC figure out. Such cases demonstrate the limitation of approaches to MWE compositionality that treat the problem as a binary classification task.

On average, the EVPCs have three senses, which is roughly twice the number for ENCs. This makes the prediction of compositionality harder, as there is more information to combine across (an effect that is compounded with the addition of synonyms and translations).
5.5 Summary

In this chapter, we proposed an unsupervised approach for predicting the compositionality of an MWE relative to each of its components, based on lexical overlap using Wiktionary, optionally incorporating synonym and translation data. The proposed method makes no language-specific assumption and is applicable to any type of MWE. The only requirement of our proposed approach is that the given MWE be listed in Wiktionary. Our experiments showed that the various instantiations of our approach are superior to the supervised methods introduced in Sections 4.1 and 4.2. The results also demonstrate the complementary of this approach with the earlier proposed approaches.

The next chapter presents the first attempt to use compositionality scores in machine translation evaluation.
Chapter 6

Incorporating Compositionality into MT Evaluation

While the explicit identification of multiword expressions has been shown to be useful in various NLP applications, recent work has shown that automatic prediction of the degree of compositionality of MWEs also has utility, in applications including information retrieval (Acosta et al. 2011) and machine translation (MT) (Weller et al. (2014), Carpuat and Diab (2010) and Venkatapathy and Joshi (2006)). For instance, Acosta et al. (2011) showed that by considering non-compositional MWEs as a single unit, the effectiveness of document ranking in an IR system improves, and Carpuat and Diab (2010) showed that by adding compositionality scores to the Moses SMT system (Koehn et al. 2007), they could improve translation quality (see Section 2.5 for more details).

A large portion of recent studies on considering MWEs in NLP applications investigate how to integrate MWEs into machine translation systems. But fewer stud-
ies investigate using compositionality scores in other NLP applications, particularly those that require semantic understanding. Therefore, this chapter presents the first attempt to use MWE compositionality scores for the evaluation of MT system outputs. The basic intuition underlying this work is that we should sensitise the relative reward associated with partial mismatches between MT outputs and the reference translations, based on compositionality. For example, an MT output of *white tower* should not be rewarded for partial overlap with *ivory tower* in the reference translation, as *tower* is most naturally interpreted compositionally in the MT output, but non-compositionally in the reference translation. On the other hand, a partial mismatch between *traffic signal* and *traffic light* should be rewarded, as the usage of *traffic* is highly compositional in both cases. That is, we ask the question: can we better judge the quality of translations if we have some means of automatically estimating the relative compositionality of MWEs, focusing on compound nouns, due to being easily identifiable, and the TESLA machine translation metric, due to making it feasible to incorporate compositionality scores (Liu et al. 2010).

In this chapter, first we introduce MT evaluation and TESLA as the MT evaluation system we are using in this project. Section 6.2 provides our proposed methodology of integrating compositionality scores of multiword expressions into TESLA. We introduce the examined datasets in Section 6.3 followed by results and discussion sections.

### 6.1 Machine Translation Evaluation

Automatic MT evaluation methods score MT system outputs based on similarity with reference translations provided by human translators. Table 6.1 shows a ref-
Australia has banned web advertising by local companies and sites but cannot restrict overseas sites.

Australia ban online advertising by local companies and websites, but it cannot be limited foreign websites.

Australia banned the Internet advertising by local companies and Internet sites, but it cannot restrict foreign sites.

Table 6.1: An example of a reference translation, and the output of two machine translation systems.

The scoring can be based on: (1) simple string similarity (Papineni et al. 2002; Snover et al. 2006); (2) shallow linguistic information such as lemmatisation, POS tagging and synonyms (Banerjee and Lavie 2005; Liu et al. 2010); or (3) deeper linguistic information such as semantic roles (Giménez and Márquez 2008; Padó et al. 2009).

In this research, we focus on the TESLA MT evaluation metric (Liu et al. 2010), which falls into the second group and uses a linear programming framework to automatically learn weights for matching n-grams of different types, making it easy to incorporate continuous-valued compositionality scores of MWEs.

TESLA measures the maximum similarity ($S_{ms}$) between the unigrams of the two given sentences (MT output and reference translation) based on the following three
Chapter 6: Incorporating Compositionality into MT Evaluation

156

terms for each pairing of unigrams \( x \) and \( y \):

\[
S_{ms} = \begin{cases} 
1 & \text{if } \text{lemma}(x) = \text{lemma}(y) \\
\frac{a+b}{2} & \text{otherwise}
\end{cases}
\]

\[
S_{lem}(x, y) = I(\text{lemma}(x) = \text{lemma}(y))
\]

\[
S_{pos}(x, y) = I(\text{POS}(x) = \text{POS}(y))
\]

where:

\[
a = I(\text{synset}(x) \cap \text{synset}(y))
\]

\[
b = I(\text{POS}(x) = \text{POS}(y))
\]

\( \text{lemma} \) returns the lemmatised unigram, \( \text{POS} \) returns the POS tag of the unigram, \( \text{synset} \) returns the WordNet synsets associated with the unigram, and \( I(.) \) is an indicator function.

The similarity between two \( n \)-grams \( x = x_1, x_2, \ldots, x_n \) and \( y = y_1, y_2, \ldots, y_n \) is measured as follows:

\[
s(x, y) = \begin{cases} 
0 & \text{if } \exists i, s(x_i, y_i) = 0 \\
\frac{1}{n} \sum_{i=1}^{n} s(x_i, y_i) & \text{otherwise}
\end{cases}
\]

TESLA uses an integer linear program to find the phrase alignment that maximizes the similarity scores over the three terms \( (S_{ms}, S_{lem} \text{ and } S_{pos}) \) for all \( n \)-grams.

6.2 Using compositionality scores in Tesla

In order to add the compositionality score into TESLA, we first identify MWEs in the MT output and reference translation. If an MWE in the reference translation
Compute similarity score between Reference and MT System output n-grams, by rewarding matches between lemmas, and WordNet synonyms

For noun compounds in reference

\[ \text{Similarity}^* = \text{Similarity} \times \text{Compositionality score} \]

Calculate alignment of n-grams that maximises similarity using integer linear programming

Figure 6.1: Integrating compositionality scores of noun compounds into TESLA (shown in BOLD)

aligns exactly with an MWE in the MT output, the weight remains as 1. Otherwise, we replace the weight computed for the noun compound with the product of computed weight and the compositionality degree of the MWE (Figure 6.1). This forces the system to be less flexible when encountering less compositional noun compounds. For instance, in TESLA, if the reference sentence contains *ivory tower* and the MT output contains *white building*, TESLA will align them with a score of 1. However, by multiplying this weight with the compositionality score (which should be very low for *ivory tower*), the alignment will have a much lower weight.

### 6.2.1 Predicting the compositionality of MWEs

In order to predict the compositionality of MWEs, we calculate the similarity between the MWE and each of its component words, using the three approaches detailed below. We calculate the overall compositionality of the MWE via linear
interpolation over the component word scores, as:

\[
\text{comp}(\text{mwe}) = \alpha \text{comp}_c(\text{mwe}, w_1) + (1 - \alpha) \text{comp}_c(\text{mwe}, w_2)
\]

where \(\text{mwe}\) is, without loss of generality, made up of component words \(w_1\) and \(w_2\), and \(\text{comp}_c\) is the compositionality score between \(\text{mwe}\) and the indicated component word. We only consider noun compounds and therefore, based on the findings of Reddy \textit{et al.} (2011) and our findings in Chapter 4, we set \(\alpha = 0.7\).

**Distributional Similarity (DS):** the distributional similarity between the MWE and each of its components, calculated based on cosine similarity over co-occurrence vectors, in the manner of Schütze (1997), using the 51st–1050th most frequent words in the corpus as dimensions. Context vectors were constructed from English Wikipedia.

**SS+DS:** the arithmetic mean of DS and string similarity (“SS”), based on the findings of Section 4.1.5. SS is calculated for each component using the LCS-based string similarity between the MWE and each of its components in a number of translations, under the hypothesis that compositional MWEs are more likely to be word-for-word translations in a given language than non-compositional MWEs (see Section 4.1 for more details).

Since we focus on NCs in this chapter, we calculate the compositionality based on the languages found to work best for English noun compounds (see Table 4.3 on page 92), namely: Czech, Norwegian, Portuguese, Thai, French, Chinese, Dutch, Romanian, Hindi and Arabic.
Table 6.2: Kendall’s ($\tau$) correlation over WMT 2013 (all-en), for the full dataset and also the subset of the data containing a noun compound in both the reference and the MT output

<table>
<thead>
<tr>
<th></th>
<th>All sentences</th>
<th>Contains NC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>METEOR</strong></td>
<td>0.277</td>
<td>0.273</td>
</tr>
<tr>
<td><strong>BLEU</strong></td>
<td>0.216</td>
<td>0.206</td>
</tr>
<tr>
<td><strong>TESLA</strong></td>
<td>0.238</td>
<td>0.224</td>
</tr>
<tr>
<td><strong>TESLA-DS</strong></td>
<td>0.238</td>
<td>0.225</td>
</tr>
<tr>
<td><strong>TESLA-SS+DS</strong></td>
<td>0.238</td>
<td>0.225</td>
</tr>
<tr>
<td><strong>TESLA-0/1</strong></td>
<td>0.238</td>
<td>0.225</td>
</tr>
</tbody>
</table>

6.3 Dataset

We evaluate our method over the data from WMT 2013, which is made up of a total of 3000 translations for five to-English language pairs, namely French, German, Spanish, Czech and Russian (Bojar et al. 2013). As our judgements, we used: (1) the original pairwise preference judgements from WMT 2013 (i.e. which of translation A and B is better?); and (2) continuous-valued adequacy judgements for each MT output, as collected by Graham et al. (2014).

We used the Stanford CoreNLP parser (Klein and Manning 2003) to identify English noun compounds in the translations. Among the 3000 sentences, 579 sentences contain at least one noun compound.

6.4 Results

We performed two evaluations, based on the two sets of judgements (pairwise preference or continuous-valued judgement for each MT output). In each case, we use three baselines (each applied at the sentence level): (1) METEOR (Banerjee and Lavie
Chapter 6: Incorporating Compositionality into MT Evaluation

<table>
<thead>
<tr>
<th></th>
<th>All sentences</th>
<th>Contains NC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>METEOR</strong></td>
<td>0.436</td>
<td>0.500</td>
</tr>
<tr>
<td><strong>BLEU</strong></td>
<td>0.272</td>
<td>0.494</td>
</tr>
<tr>
<td><strong>TESLA</strong></td>
<td>0.303</td>
<td>0.467</td>
</tr>
<tr>
<td><strong>TESLA-DS</strong></td>
<td>0.305</td>
<td>0.464</td>
</tr>
<tr>
<td><strong>TESLA-SS+DS</strong></td>
<td>0.305</td>
<td>0.464</td>
</tr>
<tr>
<td><strong>TESLA-0/1</strong></td>
<td>0.308</td>
<td>0.464</td>
</tr>
</tbody>
</table>

Table 6.3: Pearson’s \( (r) \) correlation results over the WMT all-en dataset, and the subset of the dataset that contains noun compounds

2005), (2) **BLEU** (Papineni et al. 2002), and (3) **TESLA** (without compositionality scores). We compare these with **TESLA** incorporating compositionality scores, based on DS (“**TESLA-DS**”) and SS+DS (“**TESLA-SS+DS**”). We also include results for an exact match method which treats the MWEs as a single token, such that unless the MWE is translated exactly the same as in the reference translation, a score of zero results (“**TESLA-0/1**”). We did not experiment with the string similarity approach alone, because of the high number of missing translations in PanLex.

In the first experiment, we calculate the sentence level Kendall’s \( \tau \) following the method used in the WMT 2013 shared task, as shown in Table 6.2, including the results over the subset of the data which contains a compound noun in both the reference and the MT output (“contains NC”). When comparing **TESLA** with and without MWE compositionality, we observe a tiny improvement with the inclusion of the compositionality scores\(^1\) (magnified slightly over the NC subset of the data), but not great enough to boost the score to that of **METEOR**.

In the second experiment, we calculate Pearson’s \( r \) correlation over the continuous-

\(^1\)Over the full dataset, more accurate correlation scores for **TESLA**, **TESLA-DS**, **TESLA-SS+DS** and **TESLA-0/1** are 0.2375, 0.2378, 0.2377 and 0.2378, respectively
Table 6.4: The number of judgements that were ranked correctly by TESLA originally, but incorrectly with the incorporation of compositionality scores ("P→N") and vice versa ("N→P"), and the absolute improvement with compositionality scores ("Δ") valued adequacy judgments, as shown in Table 6.3, again over the full dataset and also the subset of data containing compound nouns. The improvement here is slightly greater than for our first experiment, but not at a level of statistical significance (Graham and Baldwin 2014). Perhaps surprisingly, the exact compositionality predictions produce a higher correlation than the continuous-valued compositionality predictions, but again, even with the inclusion of the compositionality features, TESLA is outperformed by METEOR. The correlation over the subset of the data containing compound nouns is markedly higher than that over the full dataset, but the \( r \) values with the inclusion of compositionality values are actually all slightly below those for the basic TESLA.

As a final analysis, we examine the relative impact on TESLA of the three compo-
sionality methods, in terms of pairings of MT outputs where the ordering is reversed based on the revised TESLA scores. Table 6.4 details, for each language pairing, the number of pairwise judgements that were ranked correctly originally, but incorrectly when the compositionality score was incorporated ("P→N"); and also the number of pairwise judgements that were ranked incorrectly originally, and corrected with the incorporation of the compositionality judgements ("N→P").

Overall, the two compositionality methods perform better than the exact match method, with the overall improvement sum of 25, 24 and 14 using DS, SS+DS and 0/1, respectively. Moreover, utilising compositionality has a more positive effect than negative. However, the difference between the numbers is, once again, very small, except for the ru-en language pair. The exact match method ("0/1") has a bigger impact, both positively and negatively, as a result of the polarisation of n-gram overlap scores for MWEs. We also noticed that the N→P sentences for SS+DS are a subset of the N→P sentences for DS. Moreover, the N→P sentences for DS are a subset of the N→P sentences for 0/1; the same is true for the P→N sentences.

6.5 Discussion

As shown in the previous section, the incorporation of compositionality scores can improve the quality of MT evaluation based on TESLA. However, the improvements are very small and not statistically significant. Part of the reason could be that we focus exclusively on noun compounds, which are contiguous and relatively easy to translate for MT systems (Koehn and Knight 2003b). Having said that, preliminary error analysis would suggest that most MT systems have difficulty translating non-
Table 6.5: Two examples from the all-en dataset. Each example shows a reference translation, and the outputs of two machine translation systems. In each case, the output of MT system 1 is annotated as the better translation.

<table>
<thead>
<tr>
<th>Reference</th>
<th>This means they are much better for our cash flow.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT system 1</td>
<td>That is why they are for our money flow of a much better.</td>
</tr>
<tr>
<td>MT system 2</td>
<td>Therefore, for our cash flow much better.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>'I felt like I was in a luxury store,’ he recalls.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT system 1</td>
<td>‘I feel as though I am in a luxury trade,’ recalls soldier.</td>
</tr>
<tr>
<td>MT system 2</td>
<td>‘I felt like a luxury in the store,’ he recalled the soldier.</td>
</tr>
</tbody>
</table>

compositional noun compounds, although then again, most noun compounds in the WMT 2013 shared task are highly compositional given that it is gathered from News, limiting the impact of compositionality scores. In fact, among 40 randomly examined noun compounds, we found only three to be non-compositional. We speculate that, for the method to have greater impact, we would need to target a larger set of MWEs, including non-contiguous MWEs such as split verb particle constructions (Kim and Baldwin 2010).

Further error analysis suggests that incorrect identification of noun compounds in a reference sentence can have a negative impact on MT evaluation. For example, *year student* is mistakenly identified as an MWE in *... a 21-year-old final year student at Temple ....*

Furthermore, when an MWE occurs in a reference translation, but not an MT system’s output, incorporating the compositionality score can sometimes result in an error. For instance, in the first example in Table 6.5, the reference translation contains the compound noun *cash flow*. According to the dataset, the output of MT system 1 is better than that of MT system 2. However, since the former translation does not contain an exact match for *cash flow*, our method decreases the alignment
score by multiplying it by the compositionality score for \textit{cash flow}. As a result, the overall score for the first translation becomes less than that of the second, and our method incorrectly chooses the latter as a better translation.

Incorrect estimation of compositionality scores can also have a negative effect on MT evaluation. In the second example in Table 6.5, the similarity score between \textit{luxury store} and \textit{luxury trade} given by TESLA is 0.75. The compositionality score, however, is estimated as 0.22, which is an incorrect estimate considering that the semantics of both \textit{luxury} and \textit{trade} are reflected in \textit{luxury trade}. The updated similarity between \textit{luxury trade} and \textit{luxury store} is therefore 0.16, which in this case results in our method incorrectly selecting the second sentence as the better translation.

\section{6.6 Summary}

This chapter described the first attempt at integrating MWE compositionality scores into an automatic MT evaluation metric. The intuition behind this work was to weigh mismatches between the MT output and reference translation based on compositionality scores. We chose the TESLA MT evaluation metric due to the possibility of incorporating compositionality scores easily. Our results show a marginal improvement with the incorporation of compositionality scores of noun compounds. The error analysis suggests that incorrect identification of noun compounds and incorrect estimation of compositionality scores are the two main shortcomings of our proposed method.
Chapter 7

Determining the Multiword Expression Inventory of a Language

Much previous research on multiword expressions has focused on the token- and type-level tasks of MWE identification and extraction, respectively. Such studies typically target known prevalent MWE types in a given language. This chapter describes the first attempt to learn the MWE inventory of a “surprise” language for which we have no explicit prior knowledge of MWE patterns, certainly no annotated MWE data, and not even a parallel corpus. Our proposed model is trained on a treebank with MWE relations of a source language, and can be applied to the monolingual corpus of the surprise language to identify its MWE construction types.
7.1 Introduction

Existing approaches to MWE identification/extraction typically target specific MWE types that are known to be prevalent in a given language, such as: (a) compound nouns in languages such as English (Copestake 2003; Ó Séaghdha 2008), German (Schulte im Walde et al. 2013) and Japanese (Tanaka and Baldwin 2003); (b) light verb constructions (LVCs) in languages such as English (Butt 2003), Persian (Karimi-Doostan 1997) and Italian (Alba-Salas 2002); and (c) compound verbs in languages such as Japanese (Uchiyama et al. 2005). Note here that the combination of highly-productive MWE types can vary greatly across languages: English is rich with compound nouns and LVCs are also common, but lacks compound verbs; Persian is rich with LVCs and adjective–noun compounds, but has very few compound nouns and compound verbs; and Japanese is rich with LVCs and compound nouns and verbs, but adjective–noun MWEs are rarer. Even for collocation extraction, this knowledge is generally assumed for a given language, in targeting only highly productive constructions such as adjective–noun or verb–noun collocations (Krenn and Evert 2001; Pecina 2008).

But what if the language of interest is one where no such prior knowledge exists, e.g. because it is a “surprise” language where rapid deployment of language technologies is required and there is no access to an informant with sufficient linguistic training to be able to inventorise the MWE types in the language (Oard 2003; Maynard et al. 2003)? Here, there is little expectation of success without an automatic method for determining the inventory and relative frequency of MWEs in a given language. This provides the motivation for our study: can we develop a method
for automatically profiling the MWE inventory of a novel language based simply on a monolingual corpus of that language, and a treebank in a second language such as English?

We carry out this research in the Universal Dependency ("UD") framework (Nivre et al. (2016)), using the method of Duong et al. (2015) to induce a delexicalised dependency parser for the surprise language, based on a supervised parsing model for a language such as English where we have a well-developed treebank in the UD. Given the parser output over a monolingual corpus in the surprise language, we then apply one of two methods to extract our MWE profile: (1) a baseline method, where we simply extract out delexicalised dependency tuples of relation type mwe or compound (including the POS tags), aggregate the counts of the pos–relation–pos triples, and extract the most frequent triples; and (2) a supervised reranker over the delexicalised dependency tuples, to better deal with noise in the output of the delexicalised dependency parser.

One additional contribution of this study is analysis of MWE annotation across different languages in the UD. We find that there are a number of competing styles of annotation, and very different levels of thoroughness in the annotation of MWEs. As part of this, we perform an “oracle” analysis of MWE extraction based on the gold-standard treebank annotations for a given language, and find that the results vary greatly between languages, due to annotation divergences. Using the supervised reranking method, however, and incorporating more and more languages for training (but holding out the surprise language), we find that we are able to “smooth” annotation differences between languages.
7.2 Background

As discussed in Section 2.3.1, there is a wealth of research on MWE identification (i.e. distinguishing MWEs from non-idiosyncratic combinations at the token level) and extraction (i.e. determining at the type level which word combinations in a corpus are MWEs). Many of these methods are customised to particular MWE constructions which are known to exist in a given corpus, e.g. noun compounds (Lapata and Lascarides 2003; Tanaka and Baldwin 2003), verb particle constructions ("VPCs": Baldwin and Villavicencio (2002); Baldwin (2005)), determinerless prepositional phrases (Baldwin et al. 2004a; van der Beek 2005), or compound verbs (Breen and Baldwin 2009). There is also a significant body of work on general-purpose MWE extraction, often based on statistical association measures applied to either a monolingual corpus (Evert and Krenn 2005; Pecina 2008; Ramisch 2012) or a parallel corpus (Melamed 1997; Moirón and Tiedemann 2006). Even here, however, POS-based constraints are generally applied on the types of MWE that are extracted (e.g. noun–noun or verb–noun bigrams). There are also methods for identifying MWEs in context using supervised models (Diab and Bhutada 2009; Li and Sporleder 2010; Schneider et al. 2014a), which require exhaustive annotation of MWE token occurrences in a corpus. All of this research differs from our work in that it either assumes knowledge of the type(s) of MWE to extract for a given language, or requires explicitly annotated MWE data in that language.

Closer to home, there has recently been work on general-purpose, unsupervised approaches to MWE extraction, making no assumptions about the types of MWE that exist in a given language (Newman et al. 2012; Brooke et al. 2014). Here, however,
the definition of MWE tends to be blurred somewhat to focus on index terms or “formulaic language”, i.e. idiomatic expressions with statistically-marked properties in a given corpus — blurred in the sense that many MWEs are not statistically marked, and also that they include formulaic expressions such as in this paper that are not formally MWEs.

Also related is recent work on resource development for low-resource languages, such as dependency parsing based on transfer learning from a higher-density language (Naseem et al. 2012; Täckström et al. 2013; Duong et al. 2015). For example, Duong et al. (2015) proposed a neural network-based parser that transfers dependency relations across languages without requiring a parallel corpus. They learn syntactic cross-lingual word embeddings by training the skip-gram model (Mikolov et al. 2013a) on a representation of the original text in which the context of each token is represented by its universal POS tags (Petrov et al. 2012). They then incorporate these word embeddings in a transition-based neural network dependency parser (Chen and Manning 2014).

Our proposed method is the first attempt to learn the MWE profile of a language with no knowledge of the target language except for POS tags (which themselves can be induced automatically, with little or no annotated data: Garrette and Baldridge (2013), Duong et al. (2014)), and no parallel corpus. We train a delexicalised dependency parser based on transfer learning (involving no syntactic annotations for the target language), and train a reranking model based on observed MWEs in only the source language(s).
Table 7.1: List of examined languages in this paper with their respective percentage of MWE tokens. Languages with a higher proportion of MWE tokens and patterns are shown in **bold**.

### 7.3 Resources

In this study, we use v1.2 of UD (see Section 3.2.4), focusing specifically on MWEs labelled as **mwe** and **compound**. As mentioned in Section 3.2.4, although the documentation for UD provides definitions of how to distinguish **mwe** and **compound** in labelling MWEs, there seem to be major inconsistencies in how they have been interpreted for particular languages: some languages do not use these relations at all, while others only annotate a subset of MWE types with these relations.

The languages examined in this study are shown in Table 7.1. These were selected based on the fact that they have at least 100 individual occurrences of the **mwe** or **compound** relation. The 5 languages in bold were selected as our test languages, based on the high prevalence of MWE annotations and diversity of MWE patterns. We discarded Hindi despite the high proportion of MWEs because: (1) it only covered **compound** relations, and has no **mwe** relations, and (2) it has a low number of distinct MWE patterns, and as such appeared skewed in its annotation. Here and for
the remainder of this chapter, we define “MWE pattern” to be an ordered tuple of the form \( \langle \text{pos}_h, \text{rel}, \text{pos}_d \rangle \), where \( \text{pos}_h \) is the POS of the head, and \( \text{pos}_d \) is the POS of the dependent in the triple. Based on this definition, English has 56 distinct MWE patterns, Croatian 49, Persian 48, Swedish 45, and Indonesian 26.

### 7.4 MWE Patterns

This study investigates the profile of MWE patterns in a given language, in the form of delexicalised dependency tuples. Figure 7.1 shows the normalised frequency of the MWE patterns found in the 5 selected languages, based on the gold-standard treebank annotations and the universal POS tags. The most frequent patterns in our 5 target languages with the \texttt{compound} relation are: \texttt{NOUN–NOUN}; \texttt{PROP–PROP} (i.e. proper noun dependencies, which should be annotated with the \texttt{name} relation rather than \texttt{compound}, according to the annotation guidelines); and \texttt{VERB–NOUN},
Figure 7.2: Cross-lingual similarity of MWE pattern distributions using JSD
Chapter 7: Determining the Multiword Expression Inventory of a Language

which includes LVCs. There are also other noticeable patterns such as VERB–ADV and VERB–ADP, corresponding to VPCs (Schulte im Walde 2004; Baldwin 2005).

**mwe** patterns are more diverse than **compound** patterns: **compound** patterns mostly involve nouns and verbs, while **mwe** patterns involve a diverse range of POS types, such as ADP–ADP or ADV–ADV, and pairings including CONJ or SCONJ (subordinating conjunctions).

We additionally measured the similarity between the MWE pattern probability distribution of the different languages using Jensen–Shannon divergence (JSD) and mean squared error (MSE) (Section 3.4.4), as shown in Figures 7.2 and 7.3, respectively for all languages in UD. To make comparison between related languages easier, we clustered the languages by language family. According to Figures 7.2 and 7.3, there is no clear indication that languages of the same family have similar MWE patterns, which is something that we might have expected.

These results suggest that although the ultimate goal of the UD project is to have compatible annotations, the MWE annotations are not, at present, consistent. In fact, annotation divergences would appear to be more noticeable than linguistic differences. For example, the Norwegian treebank annotates only VPCs (and not multiword compound nouns, e.g.), and the **mwe** relation is not used at all. That is, the observed differences in MWE patterns certainly reflect differences between languages, but greater than this, they capture differences in the annotation process between different languages.

We also examined the annotation consistency of MWEs between the Train+Dev sets and Test set of each language (based on the provided splits), and observed high
Figure 7.3: Cross-lingual similarity of MWE pattern distributions using MSE.
consistency (low JSD) between the existing patterns in these sets for the same language. The JSD on compound patterns are all below 0.10, except for Spanish (0.23) and French (0.18). Due to the diversity of mwe patterns, the JSD is less consistent within each language, with Croatian (0.64) and German (0.44) being notably high, and the rest of the languages below 0.25. This shows that annotation is quite consistent within each language.

Therefore, despite the cross-lingual annotation inconsistency, our corpora appear to be internally consistent enough to train a model over, based on the observed MWEs in a language.

\section*{7.5 Methodology}

In this work, we measure the likelihood of the triple $\langle \text{pos}_h, \text{rel}, \text{pos}_d \rangle$ being an MWE pattern in the target language. The scores are measured according to the respective lexical instances of each triple in the source language, aggregated to compute scores for each triple, and used to train a support vector regression (SVR) model (Section 3.3.2).

The gold-standard labels to train the model are based on the dependency relations: the value is set to 1 if the dependency is compound or mwe, and 0, otherwise.

We use 8 features in our proposed method: pointwise mutual information (PMI), $\phi$-square, the Dice coefficient, student’s $t$ test, log-likelihood ratio, pattern fixedness,
token/type ratio for a given triple, and token/token ratio across all relations.

\[
\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}
\]

\[
\phi^2 = \frac{(n(x, y)n(\bar{x}, \bar{y}) - n(x, \bar{y})n(\bar{x}, y))^2}{n(x)n(\bar{x})n(\bar{y})n(y)}
\]

\[
\text{Dice coefficient} = \frac{2n(x, y)}{n(x) + n(y)}
\]

\[
t(x, y) = \frac{n(x, y) - (n(x)n(y))}{\text{total words} \sqrt{n(x, y)}}
\]

where \(n(.)\) is the number of occurrences, and \(\bar{x}\) is the number of all instances except \(x\). Pattern fixedness is measured via entropy as \(H(\text{Pr}(D(x, y)))\), where \(D(x, y)\) is the difference between the linear position of the head and dependent, binned as follows: \(\text{posdiff} \in \{(-\infty, -2), -2, -1, 1, 2, (2, \infty)\}\).

These features are first computed for each lexical instance of a given pattern, and then aggregated to calculate the overall feature values for each triples, using either: (a) the mean (in the case of pattern fixedness); or (b) the median (in the case of the other measures).\(^1\)

After computing the features, we train an SVR model based on the dependency triples in the source language, and then apply the model to rank the triples in the target language. To avoid noisy annotations, we consider only those triples that occur at least twice in each corpus.

We further experiment with a simple ensemble method to combine source languages, in order to smooth over annotation and linguistic differences between languages: we combine the trained rerankers from multiple source languages by calcu-

\(^1\)MWEs components are usually seen in a fixed order and with fixed gap size. We use mean to aggregate the pattern fixedness scores in order to capture any lexical instances which are not used in a fixed order. However, we use the median for the other measures to suppress the impact of outliers. Our preliminary results also confirm that this is the best way to aggregate the scores.
lating the average of the predicted scores from each language.

7.6 Results

We report on two experiments. First, we train a model using features extracted from the gold-standard treebank in a given source language, and apply it to features extracted from the gold-standard treebank in a target language. We investigate how well our model is able to find the annotated MWE triples when gold-standard dependency relations are provided. This experiment also shows how our model can be used to find new MWE patterns in existing annotated treebanks (missing certain MWE types). Second, we investigate how our model performs in the more realistic scenario of no annotated treebank being available in the target language.

7.6.1 Experiment I: Learning given the gold standard treebank

In our first experiment, we assume access to gold standard annotations of POS tags and relation edges in both source and target languages, to determine the tractability of the task, assuming perfect parses.

Since the output of our model is a score in the range [0, 1], we evaluate based on the area under the curve (AUC) from a ROC curve. Figure 7.4 shows the ROC curve for predicting MWEs when English and Norwegian are the source languages. English is among our 5 selected languages — i.e., one of the languages with the highest number of multiword expression patterns — while for Norwegian, the mwe
Figure 7.4: Selecting a language as a source language
relation is not used at all and only \textbf{compound [:prt]} is annotated. According to these results, the average AUC for predicting MWE patterns is 0.63 when English is the source language (averaged across all target languages, excluding English), while it is 0.50 when Norwegian is the source language. This shows that a source language with less annotated patterns makes for a weaker model. The average scores when our 5 selected languages are used as the source language are remarkably similar as shown in Table 7.2.

To investigate further, Figure 7.5 shows how adding more source languages affects the results for MWE pattern extraction over our 5 selected languages. According to these results, using more than one language can increase the AUC, however, using more than 3 languages does not improve the average AUC greatly.

The best source language for the prediction of MWEs in each of the 5 selected languages is shown in Table 7.3. As expected, Norwegian is always among the least successful languages because of the small number of annotated patterns.

Finally, we show the top predicted MWE patterns in English in Table 7.4. We observe errors such as \langle \textsc{Noun}, \textsc{punct}, \textsc{Num} \rangle and \langle \textsc{Sym}, \textsc{punct}, \textsc{Sym} \rangle, because of their idiosyncratic properties across token instances. However, our model also predicts that \langle \textsc{Noun}, \textsc{ccomp}, \textsc{Adj} \rangle is an MWE.

<table>
<thead>
<tr>
<th>Source language</th>
<th>Average target AUC scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.63</td>
</tr>
<tr>
<td>Persian</td>
<td>0.64</td>
</tr>
<tr>
<td>Croatian</td>
<td>0.64</td>
</tr>
<tr>
<td>Indonesian</td>
<td>0.61</td>
</tr>
<tr>
<td>Swedish</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 7.2: Average AUC scores for each source language.
7.6.2 Experiment II: Learning without gold standard dependency relations

In our second experiment, we evaluate under the more realistic task setting of there being no gold standard treebank in the target language. Instead, we use the cross-lingual parser proposed by Duong et al. (2015) to parse the corpus in the target language (see Section 7.2). Note that we still use gold-standard POS tags, but this isn’t entirely unrealistic given the relative maturity of methods for inducing universal POS taggers (Das and Petrov 2011; Täckström et al. 2013; Duong et al. 2014).

Obviously, due to the fact that the parser has no access to dependency annotations in the target language, the parser output will be noisy. However, this emulates a true
### Chapter 7: Determining the Multiword Expression Inventory of a Language

#### Table 7.3: The best source languages for predicting the MWE patterns in each of the 5 selected languages.

<table>
<thead>
<tr>
<th>Target language</th>
<th>Best source languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>da &gt; sv &gt; hr &gt; de &gt; fa &gt; id &gt; ... &gt; es &gt; no &gt; hu</td>
</tr>
<tr>
<td>Croatian</td>
<td>ga &gt; bg &gt; fr &gt; de &gt; da &gt; eu &gt;... &gt; id &gt; hu &gt; no</td>
</tr>
<tr>
<td>Persian</td>
<td>da &gt; fr &gt; de &gt; sv &gt; bg &gt; hr &gt; ... &gt;no &gt; it &gt; nl</td>
</tr>
<tr>
<td>Swedish</td>
<td>eu &gt; de &gt; fr &gt; fi &gt; bg &gt; da &gt; ... &gt;hi &gt; nl &gt;no</td>
</tr>
<tr>
<td>Indonesian</td>
<td>ga &gt; bg &gt; en &gt; hi &gt; nl &gt; de &gt; ... &gt; et &gt; he &gt;no</td>
</tr>
</tbody>
</table>

#### Table 7.4: The top predicted MWE patterns in English, by combining all other languages.

<table>
<thead>
<tr>
<th>Score</th>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.327</td>
<td>\langle NOUN, ccomp, ADJ \rangle</td>
<td>sure place</td>
</tr>
<tr>
<td>0.306</td>
<td>\langle X, compound, PROPN \rangle</td>
<td>Indo Lanka</td>
</tr>
<tr>
<td>0.301</td>
<td>\langle NOUN, appos, SYM \rangle</td>
<td>$ value</td>
</tr>
<tr>
<td>0.298</td>
<td>\langle ADJ, nmod:npmod, ADV \rangle</td>
<td>little more</td>
</tr>
<tr>
<td>0.295</td>
<td>\langle NOUN, punct, NUM \rangle</td>
<td>5 &quot;</td>
</tr>
<tr>
<td>0.285</td>
<td>\langle SYM, punct, SYM \rangle</td>
<td>— —</td>
</tr>
<tr>
<td>0.283</td>
<td>\langle AUX, advcl, ADV \rangle</td>
<td>as can</td>
</tr>
<tr>
<td>0.277</td>
<td>\langle NOUN, mwe, SCONJ \rangle</td>
<td>in case</td>
</tr>
<tr>
<td>0.270</td>
<td>\langle SCONJ, mwe, ADP \rangle</td>
<td>due to</td>
</tr>
<tr>
<td>0.268</td>
<td>\langle NOUN, mwe, ADP \rangle</td>
<td>in order</td>
</tr>
</tbody>
</table>

In order to evaluate our proposed method and compare it with the gold-standard treebank, we change the evaluation method slightly in order to better reflect the expected inconsistencies in the parser output. In terms of gold-standard labeling, we exhaustively consider every edge between all pairs of tokens in each sentence, and consider an edge to be a positive instance if there is an MWE dependency between its token pairs in the gold-standard treebank, and a negative instance otherwise. To evaluate the parser output, which is the baseline in this experiment, we use the generated dependency edges and labels, and evaluate this against the exhaustively-generated surprise language setup, where we have no prior knowledge of MWEs or dependency structure in the target language.
Chapter 7: Determining the Multiword Expression Inventory of a Language

Table 7.5: AUC scores when English is used as the source language to transfer dependency links and to train our reranker model. In “Baseline + gold”, the trained model is applied to the gold-standard annotation of the target language rather than the parsed corpus.

gold-standard. To evaluate our system’s performance, we use the dependency edges given by the parser and aggregate the reranker’s predicted scores at the level of the delexicalised dependency triples, as per the first experiment. Unlike the first experiment, we evaluate using ROC AUC over the token pairs instead of $\langle pos_h, rel, pos_d \rangle$ triples.

Table 7.5 shows the AUC scores when English is used as the source language to parse the target language, and English is also used to train our reranking model. Our proposed model produces above-baseline results for all target languages except German, Danish, Italian and Croatian. We observe a very high percentage (63%) of compounds being predicted as noun–noun compounds in German, which is a large part of the strong results for that language.

In order to compare with a collocation extraction methods, we contrast this with a ranking based on the average PMI score for each dependency relation (“PMI”). The results show that for half of the languages simple PMI scores can lead to higher AUC scores, while for the other half, the reranker model (which incorporates PMI scores but is trained on another language), performs better.

The final row in Table 7.5 is the result of providing the baseline method with gold standard dependency relations (with unknown label, to avoid trivialising the
Figure 7.6: Combining source languages given the noisy dependency relations. In (a)–(c), the English treebank is used as the source language for the cross-lingual parser, and in (d)–(f) Swedish is used.

Similar to the previous experiment, we also experimented with an ensemble of rerankers.

We use English and Swedish as the source language to parse Persian, Croatian
Table 7.6: Top-ranking Persian and Croatian MWE patterns extracted using English and Swedish as the source language. Those patterns which match the top-ranking gold standard patterns are shown with “†”.

<table>
<thead>
<tr>
<th>Source = English</th>
<th>Source = Swedish</th>
<th>Gold standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(NOUN, compound, NOUN) †</td>
<td>(ADP, case, NOUN)</td>
<td>(NOUN, +, VERB)</td>
</tr>
<tr>
<td>(ADJ, amod, NOUN) †</td>
<td>(NOUN, nsubj, VERB) †</td>
<td>(ADJ, +, ADP)</td>
</tr>
<tr>
<td>(NOUN, nmod, NOUN) †</td>
<td>(NOUN, nmod : poss, NOUN) †</td>
<td>(ADJ, +, VERB)</td>
</tr>
<tr>
<td>(NOUN, nsubj, VERB) †</td>
<td>(NOUN, nmod, VERB) †</td>
<td>(NUM, +, NOUN)</td>
</tr>
<tr>
<td>(NOUN, nmod, ADJ)</td>
<td>(VERB, acl : relcl, NOUN)</td>
<td>(NUM, +, NOUN)</td>
</tr>
<tr>
<td>(DET, det, NOUN)</td>
<td>(ADJ, amod, NOUN) †</td>
<td>(CONJ, +, PRON)</td>
</tr>
<tr>
<td>(ADV, advmod, NOUN)</td>
<td>(CONJ, cc, NOUN)</td>
<td>(CONJ, +, NOUN)</td>
</tr>
<tr>
<td>(NOUN, conj, VERB) †</td>
<td>(ADJ, nsubj, VERB) †</td>
<td>(ADJ, +, NOUN)</td>
</tr>
<tr>
<td>(NOUN, nmod, VERB) †</td>
<td>(DET, det, NOUN)</td>
<td>(NUM, +, ADP)</td>
</tr>
<tr>
<td>(NOUN, conj, SCONJ)</td>
<td>(NUM, nummod, NOUN) †</td>
<td>(CONJ, +, NOUN)</td>
</tr>
<tr>
<td>Croatian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ADJ, amod, NOUN) †</td>
<td>(ADJ, amod, NOUN) †</td>
<td>(PRON, +, VERB)</td>
</tr>
<tr>
<td>(NOUN, compound, NOUN) †</td>
<td>(NOUN, nmod, NOUN) †</td>
<td>(ADJ, +, NOUN)</td>
</tr>
<tr>
<td>(ADP, case, NOUN) †</td>
<td>(ADP, case, NOUN) †</td>
<td>(NUM, +, NOUN)</td>
</tr>
<tr>
<td>(NOUN, nmod, NOUN) †</td>
<td>(NOUN, dobj, VERB)</td>
<td>(PROP., +, PROPN)</td>
</tr>
<tr>
<td>(NOUN, dobj, VERB)</td>
<td>(NOUN, nmod, PROP)</td>
<td>(PROP., +, NOUN)</td>
</tr>
<tr>
<td>(NOUN, compound, PROP)</td>
<td>(NOUN, nmod : poss, NOUN) †</td>
<td>(NUM, +, NOUN)</td>
</tr>
<tr>
<td>(NOUN, nmod, VERB)</td>
<td>(NOUN, nmod, VERB)</td>
<td>(ADP, +, NOUN)</td>
</tr>
<tr>
<td>(PROP., compound, PROP) †</td>
<td>(NOUN, nsubj, VERB)</td>
<td>(X, +, X)</td>
</tr>
<tr>
<td>(NOUN, nsubj, VERB)</td>
<td>(AUX, aux, VERB)</td>
<td>(PROP., +, ADP)</td>
</tr>
<tr>
<td>(NOUN, conj, NOUN) †</td>
<td>(PRON, nsubj, VERB) †</td>
<td>(PRON, +, SCONJ)</td>
</tr>
</tbody>
</table>

and Indonesian. Figure 7.6 shows how increasing the number of source languages and combining the trained models affects the AUC scores. According to Figure 7.6, our proposed method on average beats the baseline, when using only one language to train the reranker. However, unlike the previous experiment, combining multiple source languages does not improve the reranker. Additionally, comparing English with Swedish, we observe that the source language used to induce the dependency parser plays an important role.

### 7.7 Error Analysis

Finally, we perform error analysis to better understand the performance of the proposed method, focusing on two languages: Persian and Croatian, with 322 and
225 patterns to rank, respectively. We selected these two languages primarily because of the diversity of the MWE annotations in the treebanks (Section 7.3), and we had access to expert native-speaker annotators. The most frequent annotated patterns in the original treebanks are shown as “Gold standard” in Table 7.6 (the relation between pos\(_h\) and pos\(_d\) are either mwe or compound or both). The dependency parser and reranker are trained on English and Swedish as the source languages, individually. The top-10 most frequent patterns in the first quartile of the output of the reranker are shown in Table 7.6. The patterns which match with the gold standard pattern are marked with “†”.

When English is the source language, noun compounds are correctly selected as a very frequent pattern in Persian. The pattern of \(\langle \text{ADJ}, \text{amod}, \text{NOUN} \rangle\) is selected as the second most common pattern in Persian, for which almost all token instances are also true MWEs, such as Islamic republic, Islamic revolution, right wing and fundamental law.\(^2\). Instances of \(\langle \text{NOUN}, \text{nmod}, \text{NOUN} \rangle\) are more institutionalized such as seminar presentation and Iftar time. Persian is rich with LVCs, which shows up in the first column as \(\langle \text{NOUN}, \text{nsubj}, \text{VERB} \rangle\), i.e. misanalysed as verb–subject rather than verb–object pairs, but containing predominantly LVCs. In fact, among the top-20 most frequent token instances of this pattern, 17 are LVCs. As we work our way down the list of dependency triples in Table 7.6, there are fewer and fewer actual MWE token instances associated with the pattern. For example, the number of MWE instances associated with \(\langle \text{ADV}, \text{advmod}, \text{NOUN} \rangle\) is less than non-MWEs (MWE examples are before Christ, and before revolution). The primary sources of error were parser errors.

\(^2\)Note that our model predicts the MWE patterns rather than MWE instances. Therefore, whether an individual MWE is also an MWE in the target language or not, does not affect the final results.
or the triple being a fragment of a larger MWE. Using Swedish as the source language, we observed a similar trend.

For Croatian, almost all of the tokens associated with the top-2 patterns for both English and Swedish are MWE instances, with the tokens associated with \(\langle \text{NOUN, compound, NOUN} \rangle\) based on English corresponding very closely with \(\langle \text{NOUN, nmod, NOUN} \rangle\) based on Swedish. As with Persian, as we go down the list, the patterns become more noisy and the MWE tokens sparser, with the exception of \(\langle \text{NOUN, compound/nmod, PROPN} \rangle\), for which almost all instances are MWEs (e.g. president Erdogan) or part of a larger MWE. Also, the instances of \(\langle \text{PROPN, compound, PROPN} \rangle\) are all named entities. None of the token instances associated with \(\langle \text{NOUN, nsubj, VERB} \rangle\) and \(\langle \text{AUX, aux, VERB} \rangle\) were MWEs.

7.8 Conclusion

In this chapter, we proposed a method for automatically determining the MWE composition of a novel language, based on delexicalised universal dependency patterns of the form \(\langle \text{pos}_h, \text{rel}, \text{pos}_d \rangle\). The method is based on determination of MWEs in a source language from a dependency treebank, and training of a model over delexicalised dependency patterns for that language. This is then applied to a target language to rerank patterns, in terms of MWEhood. In our initial experiments, we used gold-standard dependency information for the target language, and found the method to be highly successful at ranking dependency patterns. This both validates the method, as well as suggests the potential for the use of the method in cross-checking the consistency of the UD treebanks. We then applied our method under
the more realistic setting of having no gold-standard dependency data for the target language, but instead the output of a dependency parser induced for the target language based only on a POS-tagged monolingual corpus in the target language (and gold-standard data in the source language). We found the method to produce above-baseline results for the majority of languages tested, and that for the false positives associated with higher token frequencies, many of the associated tokens were actually true instances of MWEs (with the wrong dependency relation).
Chapter 8

Conclusion

The goal of this thesis has been to investigate approaches to analyzing multiword expressions, which, unlike most previous studies in this domain: (1) do not make any language specific assumptions; and (2) are applicable to any type of MWE. We have focused on three studies, considering the above mentioned goals: (1) modeling the compositionality of multiword expressions and their components; (2) integrating compositionality scores of MWEs into machine translation evaluation; and (3) determining the MWE inventory of a language.

This chapter first summarizes the research questions, proposed methods and findings of each chapter of this thesis. Then, Section 8.2 discusses the limitations of our work and outlines future research directions.
8.1 Summary of Chapters and Contributions

In Chapter 2, we provided a detailed definition of the concept of multiword expression, and discussed the linguistic features of MWEs including lexical, semantic, syntactic, pragmatic and statistical idiosyncrasy. Then, we introduced the most famous types of MWEs in addition to their individual characteristics. We reviewed the computational literature on MWEs, and explained the challenges and approaches to the identification and extraction of MWEs. In addition, we discussed various studies related to the semantic of MWEs.

A large part of this thesis focuses on compositionality of MWEs, and therefore, we surveyed research on measuring the compositionality of MWEs. Our review showed that a large number of studies make language-specific or type-specific assumptions, or require language-specific pre-processing steps.

At the end of Chapter 2, we explained the importance of explicitly capturing MWEs and modeling their compositionality degree, in reviewing previous studies which investigate the integration of MWEs into NLP applications. This review shows that most previous research has chosen machine translation as the target NLP application, and reported improvement in performance and accuracy after this integration.

Chapter 3 introduced the datasets, resources, machine learning approaches and evaluation metrics used in this thesis. The datasets include two types of MWEs (noun compounds and verb particle constructions) and two languages (English and German). This chapter detailed the resources used in our experiments for predicting compositionality scores: (1) PanLex, which is a multilingual dictionary used to translate MWEs and their constituents (used in Chapter 4); (2) Wikipedia, which provides
monolingual corpora in multiple languages (used in Chapter 4); and (3) Wiktionary, which provides definitions of words and multiword expressions, as well as translations to other languages (used in Chapter 5). We further provided information about the Universal Dependency (UD) treebank, which provides annotations based on a universal part-of-speech and dependency relation set. This resource helps us to build language-independent models to determine MWE patterns of a language in Chapter 7. At the end of this chapter, we provided details of our classification and regression frameworks, which are support vector machine and support vector regression. We concluded the chapter by reviewing the evaluation metrics used throughout this thesis.

In Chapter 4, we proposed three approaches to predict the compositionality of multiword expressions. The approaches proposed in this chapter are general, in the sense that they are applicable to any type of MWE in a wide range of languages. The three proposed approaches are: (1) translation-based string similarity, (2) translation-based distributional similarity and (3) word embeddings.

The translation-based string similarity approach is based on the assumption that compositional MWEs are more likely to be word-for-word translations in a given language compared to non-compositional MWEs. In this approach, the given MWE and its components are translated to multiple languages, and the translation of each of the components is compared with the translation of the MWE using string similarity measures. The results show that using multiple languages, rather than a single language, improves the prediction accuracy. Comparing to a state-of-the-art method which requires identification of MWE instances in text, we observed that this ap-
proach performs better when MWE instances cannot be easily identified in the text. Our error analysis showed the shortcoming of our proposed method when it comes to multiword expressions which have both compositional and non-compositional senses. Loan translations are another source of error in our proposed approach.

Translation-based distributional similarity is the second approach proposed in Chapter 4 to predict the degree of compositionality. This approach is based on previous state-of-the-art studies of Reddy et al. (2011) (English noun compound), Schulte im Walde et al. (2013) (German noun compound) and Bannard (2006) (English verb particle constructions), which use distributional similarity to compare the semantics of an MWE with that of its components. In our proposed approach, instead of using distributional similarity in only the source language, we measure the distributional similarity between the translations of an MWE and the translations of its components in other languages. Our experiments showed that using languages other than the source language improves the prediction performance for English verb particle constructions, which are hard to identify in English and German noun compounds, which have low recall in identification because of the rich morphology of German. We also found that the string similarity approach complements the distributional similarity approach. Our error analysis showed that not having entries in PanLex is the major source of errors in this method.

Lastly, in Chapter 4, we investigated the strengths and weaknesses of word embeddings in the task of predicting the degree of compositionality. We noted that, consistent with recent work over other NLP tasks, word embeddings are superior to count–based models of distributional similarity. We also compared single word embeddings
with multi-prototype embeddings and observed slight superiority of single-prototype embeddings to multi-prototype word embeddings in this task. Our experiments also showed that the translation-based string similarity approach complements the word embedding approach. Furthermore, we discussed several experiments to examine the sensitivity of word-embedding approaches to the initial hyperparameter settings, and observed low sensitivity overall. Our error analysis showed that the performance of word embedding approaches drops more markedly than other methods for MWEs with low frequency.

**Chapter 5** proposed an unsupervised approach to identify non-compositional components of an MWE using definitions and translations provided in Wiktionary. The only requirement of this approach is that the given MWE be listed in Wiktionary. The experiments showed the superiority of this method to the supervised methods introduced in Chapter 4. However, we observed that polysemy is one of the main challenges in this approach; for instance, we noticed that English verb particle constructions have three senses on average.

In **Chapter 6**, we introduced the first attempt to integrate the compositionality of MWEs into machine translation evaluation. The main hypothesis was to reward partial mismatches between noun compounds according to the compositionality degree. In this study, we limited our focus to noun compounds, as they are common in English and can be easily identified. We chose TELSA as the machine translation evaluation system in this study, because it is easy to incorporate compositionality scores.

The results show a marginal improvement of **TELSA** with the incorporation of
compositionality scores. Based on our error analysis, we identified that incorrect identification of noun compounds and incorrect estimation of compositionality scores are the two main limitations of our method.

**Chapter 7** discussed a different aspect of MWE analysis. In this chapter, we proposed a method to determine the MWE inventory of a language, without making any assumptions about the existing patterns of MWEs in that language. The proposed model is trained over delexicalised dependency patterns of other languages. Then, the trained model is used to rank the patterns of the target language in terms of the likelihood of it being an MWE pattern.

We provided some insights into how MWEs are annotated across Universal Dependency treebanks for different languages, and showed that MWE patterns are annotated differently from one language to another. Using the gold-standard dependency relations, we performed some experiments to validate the proposed method and suggested the potential for the use of it in cross-checking the provided annotation in UD treebank.

Finally, for a more realistic scenario, instead of using the gold-standard dependency relations, we used the output of a cross-lingual dependency parser trained on a POS-tagged monolingual corpus. We showed above-baseline results for the majority of the languages tested, with the dependency parser being the primary source of errors.
8.2 Future work

This thesis detailed language independent approaches to: (1) predicting the compositionality of MWEs; (2) identifying non-compositional components of MWEs; (3) integrating compositionality of MWEs into machine translation evaluation; and (4) identifying multiword expression patterns of a language. This section first outlines potential short-term extensions to our proposed approaches in each chapter, and then returns to discuss untouched directions for future research.

8.2.1 Short-term Extensions

In Chapter 4, we proposed three approaches to predict the degree of compositionality. In Section 4.1 a translation-based string similarity approach was proposed, whereby the translations of MWEs were compared with the translations of components using string similarity metrics. Since there might be more than one translation for a word/MWE, we measured the maximum similarity score (\( \text{MAX} \)) between all translation pairs. In this approach, noisy and wrong translations can result in inaccurate similarity scores. To improve this approach, we suggest to find the most accurate or probable translation for each MWE and components instead of using the \( \text{max} \) operator.

One way to find the best translation, without requiring language specific tools, is to use Wikipedia. Wikipedia documents have comparable documents in multiple languages, meaning that while the documents are not exact translations of one another, they more or less convey the same semantic information. Our intuition is that the best translation of a word/MWE can be found by using documents which contain a
given word/MWE and looking at comparable documents in other languages. Then, among the provided translations in PanLex, we can choose the one which occurs more in the comparable documents.

As discussed in Section 4.1, we had some experiments to find out why the top languages were selected as the best target languages; We found that target languages which had more translation coverage were selected as top languages. While PanLex and other resources such as Babelnet¹ (Navigli and Ponzetto 2010) are growing in terms of translation coverage, we suggest to perform a more detailed research when all languages have relatively similar coverage. Such experiments can provide more linguistic insight into the similarities and differences between multiword expressions across languages.

In Section 4.2, we proposed the translation-based distributional similarity approach, in which the distributional similarity between the translation of MWEs and translation of components was measured. Similarly to the translation-based string similarity approach, we suggest looking for the most accurate translation instead of using the MAX operator. To make the study generally applicable to a wide range of languages, we did not perform any lemmatization or stemming. Although we observed promising results, we anticipate there is still space for more improvement by lemmatizing the words especially in morphologically rich languages. In line with our language-independent assumption, we suggest to use unsupervised lemmatizing methods to lemmatize the words without requiring any language specific training data, tools or assumptions (Goldsmith 2001; Creutz and Lagus 2002; Poon et al. 2009).

¹http://lcl.uniroma1.it/babelnet/
In Section 4.3, we proposed using word embeddings to compute the compositionality degree of MWEs. We assumed that we knew the MWE instance before training the models and producing the word embeddings. In fact, we only considered the MWEs in our datasets and trained the model by concatenating the components of instances of that limited dataset. However, to enhance this approach, it might be better to identify all the MWEs in text using general approaches such as the method proposed by Schneider et al. (2014a).

In another direction, based on our findings of Section 4.2 we found that using distributional similarity in multiple languages enhances the compositionality prediction. As a future extension to this work, we like to study whether using more than one language can enhance the word embeddings approach as well.

In Chapter 5, we proposed an unsupervised method to identify non-compositional components of an MWE using Wiktionary. In this work, we only experimented with English noun compounds and English verb particle constructions. However, although we did not make any language specific assumptions, we did not experiment with MWEs in other languages due to limited entries for other languages at the time of writing this thesis. As Wiktionary has the potential to grow in other languages as well as English, it would be valuable to examine if we can get the same results as for English MWEs. Another limitation of this work was when we encountered MWEs with more than one sense. There are studies on word sense disambiguation for MWEs (Hashimoto and Kawahara 2008; Fothergill and Baldwin 2012) and word sense distribution learning (McCarthy et al. 2004; Bennett et al. 2016), which can be used to better predict which sense to consider.
In Chapter 6, we proposed to integrate our predictions of the degree of compositionality into machine translation evaluation. We limited this work to identifying noun compounds and integrating their compositionality. Although we noticed a marginal improvement, we realized that machine translation systems are performing well on translating noun compounds. For future work, we suggest to consider MWEs with more complex structures. To identify MWEs, we suggest to use the approach proposed by Schneider et al. (2014a). In their supervised method, they propose to use a feature-rich sequence tagger to identify all kinds of MWEs including those containing gaps (e.g. VPCs in split forms).

In Chapter 7, we presented the first attempt to determine the MWE inventory of a language. The proposed approach does not make any assumptions about the language. The ultimate goal of this project is to help linguists and computational linguists to get a better understanding of MWE patterns in a language. We suggest to use this method to extract dominant patterns in low-resource languages with no explicit information on MWEs. For future work, we also suggest to extend this approach to extract MWE tokens of a language in an unsupervised manner. In this approach, the potential MWEs are extracted using MWE patterns. Then, the degree of MWEhood for each instance is measured using statistical association measures.

8.2.2 Future Directions

Along with the proposed improvements above, there are some related areas that we did not touch on in this thesis. This section details these areas.

Semantic interpretation of compositional MWEs: In Chapters 4 and 5
we proposed methods to measure the compositionality degree of MWEs and predict to what extent the meaning of the components is reflected in MWE. However, we did not study the meaning of the MWEs. Several methods have been proposed methods to interpret the meaning of English noun compounds, such as automatically interpreting that *chocolate cake* is cake made of chocolate, while *honey bee* is a bee that makes honey (Lapata 2002; Kim and Baldwin 2005; Girju 2007). There are, in fact, lists of possible semantic relations between the components of English noun compounds (Lapata 2002; Kim and Baldwin 2005; Girju et al. 2007; Hendrickx et al. 2009; Hendrickx et al. 2013). However, there are less studies regarding the semantic of other types of MWEs. For example, Cook and Stevenson (2006) study verb particle constructions, but, they only focus on English VPCs with the particle *up*. More studies are required for other types of MWEs and languages other than English.

There are limited studies on interpreting noun compounds of other languages such as Chinese (Zhao et al. 2007) and Korean (Yoon et al. 2001).

We showed how cross-lingual information can help in better predicting the compositionality of MWEs in Chapter 4. Girju (2007) observed a similar effect by considering cross-lingual information to interpret semantic relations. Our intuition is that the use of cross-lingual information can be extended to interpret semantic relations of other types of MWEs in multiple languages.

Recent work on relation modeling using vector subtraction over word embeddings has shown to be successful (Mikolov et al. 2013a; Mikolov et al. 2013b). A famous example in this task is the automatic prediction of the word *queen* for the combination of *king − man + woman* using word embeddings. We suggest to examine this
approach to interpret the relations between the MWE components. For example, we hypothesize that for \( MWE_1 \) (with \( component_{11} \) and \( component_{12} \) and \( MWE_2 \) (with \( component_{21} \) and \( component_{22} \)), the relations (interpretation) should be the same if the following equation is true:

\[
MWE_1 - (component_{11} + component_{12}) \approx MWE_2 - (component_{21} + component_{22})
\]

Using this intuition, we can interpret MWEs according to their similarity to annotated data. Furthermore, we can use this approach to cluster the relation types in a language where we have no previous knowledge.

In a similar vein, we suggest studying the potential interaction between compositionality and semantic relations within noun compounds (Barker and Szpakowicz 1998; Ó Séaghdha and Copestake 2007; Tratz and Hovy 2010). First, we hypothesise that different semantic relations have different tendencies to occur within compositional or non-compositional NCs. Second, we hypothesize that compositionality can be determined based on how much we can support the existence of a given semantic relation. For example, consider *ivory tower*, which has a MATERIAL (i.e., made of) relation. Analysis of the context of the NC will tend not to support the MATERIAL relation between *ivory* and *tower*, but analysis of the component words will, based on which, we can potentially predict that *ivory tower* is non-compositional. On the other hand, a compositional NC such as *bus driver* will provide contextual evidence to support the semantic relation between the components (here, AGENT). This suggests a joint approach to the analysis of compositionality and semantic relations.

**Meaning of non-compositional MWEs:** Modeling the meaning of non-compositional MWEs is another important task which is neglected and requires further research.
There exist a number of studies on metaphors and similes. For example, *sound like prophet* is automatically associated with characteristics such as *wise, insightful, prescient* and *enlightened* (Veale and Hao 2007; Li *et al.* 2012; Qadir *et al.* 2016).

This study can be extended to non-compositional but figurative MWEs such as *kill two birds with one stone*. Such MWEs are understandable to non-native speakers of the language despite being non-compositional, through visualizing the phrase and reasoning by comparing it with the context. A more long-term goal would be figuring out the meaning of idioms automatically from context. In recent studies, word embeddings have been shown to be quite strong in capturing the semantic of words and finding semantically similar words. Therefore, we suggest to examine word embeddings in this task. However, we anticipate the main challenge in this study would be dealing with the problem of low frequency.

**Identifying Equivalent Idioms across Languages** Our third suggestion for future study is to look for approaches to find equivalents of an idiom in other languages, for example, a system which can learn that *evil eye* in English is equivalent to *salty eye* in Persian. Some equivalent idiom pairs translate word-for-word across languages; others vary slightly across languages, such as *in the black* and its equivalent in Portuguese *no azul* (“*in the blue*”) (Villavicencio *et al.* 2004).

Idioms are semantically non-compositional multiword expressions, in which the overall meaning diverges from the meanings of some or all of the individual components. Idioms are not only interesting theoretically and computationally, but are also relevant in fields such as sociolinguistics and anthropological linguistics as idioms can reflect cultural and national characteristics of the speakers of a particular language.
(Pike 1954). For example, *cut your cloth according to your coat* in English is equivalent to *lay your feet according to your gelim* in Persian, and *all that glitters is not gold* in English is equivalent to *all the things that are white are not milk* in Tamil.

Despite their importance, idioms have traditionally not been handled well in natural language processing due to their syntactic variability, semantic non-compositionality and sparsity in corpora. All these characteristics make them hard to identify and semantically understand.

Pershina *et al.* (2015) compared English idioms and extracted semantically similar pairs (e.g., *cloud nine* and *seventh heaven*). They use lexical similarity and word embeddings to capture semantic similarity between definitions rather than the idioms themselves. This study can be extended to identify equivalent idioms in a bilingual setting rather than monolingual.

---

2 A gelim is a small rug, which Iranians used to sit on.
Bibliography


BIBLIOGRAPHY


Calzolari, Nicoletta, Charles J Fillmore, Ralph Grishman, Nancy Ide,


—-, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014.


Graham, Yvette, and Timothy Baldwin. 2014. Testing for significance of in-


Hendrickx, Iris, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way clas-


Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. 2003. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.


LESK, MICHAEL. 1986. Automatic sense disambiguation using machine readable
dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings
of the 5th Annual International Conference on Systems Documentation*, 24–26,
Ontario, Canada.

LI, BIN, HAIBO KUANG, YINGJIE ZHANG, JIAJUN CHEN, and XURI TANG. 2012.
Using similes to extract basic sentiments across languages. In *Proceedings of the
International Conference on Web Information Systems and Mining*, 536–542,
Chengdu, China.

LI, LINLIN, and CAROLINE SPORLEDER. 2010. Linguistic cues for distinguishing
literal and non-literal usages. In *Proceedings of the 23rd International Confer-
ence on Computational Linguistics (COLING 2010), Posters Volume*, 683–691,
Beijing, China.

LI, WEI, XIUHONG ZHANG, CHENG NIU, YUANKAI JIANG, and ROHINI SRIHARI.
2003. An expert lexicon approach to identifying English phrasal verbs. In *Pro-
cedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acqui-
sition and Treatment*, 513–520, Sapporo, Japan.

of the 31st Annual Meeting on Association for Computational Linguistics (ACL
1993)*, 112–120, Columbus, USA.

———. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the
17th International Conference on Computational Linguistics (COLING 1998)*,
768–774, Montreal, Canada.
BIBLIOGRAPHY


—, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In
Proceedings of Advances in Neural Information Processing Systems, 3111–3119, Lake Tahoe, USA.


Passos, Alexandre, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 78–86, Ann Arbor, USA.


Pershina, Maria, Yifan He, and Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In Proceedings of the EMNLP 2015 Workshop
on Linking Models of Lexical, Sentential and Discourse-level Semantics, 76–82, Lisbon, Portugal.


PICHOTTA, KARL, and JOHN DeNERO. 2013. Identifying phrasal verbs using many


— —, — —, and Timothy Baldwin. 2014a. Detecting non-compositional MWE com-


— —, SPENCER ONUFFER, NORA KAZOUR, EMILY DANCHIK, MICHAEL T MOR-


BIBLIOGRAPHY


Villavicencio, Aline, Timothy Baldwin, and Benjamin Waldron. 2004. A


Wierzbicka, Anna. 1982. Why can you have a drink when you can’t *have an eat? Language 58.753–799.


## Appendix A

### List of Languages

<table>
<thead>
<tr>
<th>ISO 639-1 code</th>
<th>Language</th>
<th>Language Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>sq</td>
<td>Albanian</td>
<td>Albanian</td>
</tr>
<tr>
<td>ar</td>
<td>Arabic</td>
<td>Semitic</td>
</tr>
<tr>
<td>eu</td>
<td>Basque</td>
<td>Basque</td>
</tr>
<tr>
<td>bn</td>
<td>Bengali</td>
<td>Indic</td>
</tr>
<tr>
<td>br</td>
<td>Breton</td>
<td>Celtic</td>
</tr>
<tr>
<td>bg</td>
<td>Bulgarian</td>
<td>Slavic</td>
</tr>
<tr>
<td>ca</td>
<td>Catalan</td>
<td>Romance</td>
</tr>
<tr>
<td>ceb</td>
<td>Cebuano</td>
<td>Austronesian</td>
</tr>
<tr>
<td>cs</td>
<td>Czech</td>
<td>Slavic</td>
</tr>
<tr>
<td>da</td>
<td>Danish</td>
<td>Germanic</td>
</tr>
<tr>
<td>nl</td>
<td>Dutch</td>
<td>Germanic</td>
</tr>
<tr>
<td>en</td>
<td>English</td>
<td>Germanic</td>
</tr>
<tr>
<td>fi</td>
<td>Finnish</td>
<td>Uralic</td>
</tr>
<tr>
<td>fr</td>
<td>French</td>
<td>Romance</td>
</tr>
<tr>
<td>gl</td>
<td>Galician</td>
<td>Romance</td>
</tr>
<tr>
<td>ka</td>
<td>Georgian</td>
<td>Caucasian</td>
</tr>
<tr>
<td>de</td>
<td>German</td>
<td>Germanic</td>
</tr>
<tr>
<td>el</td>
<td>Greek</td>
<td>Greek</td>
</tr>
<tr>
<td>he</td>
<td>Hebrew</td>
<td>Semitic</td>
</tr>
<tr>
<td>hi</td>
<td>Hindi</td>
<td>Indic</td>
</tr>
<tr>
<td>hu</td>
<td>Hungarian</td>
<td>Uralic</td>
</tr>
<tr>
<td>is</td>
<td>Icelandic</td>
<td>Germanic</td>
</tr>
<tr>
<td>io</td>
<td>Ido</td>
<td>Celebes</td>
</tr>
<tr>
<td>ISO 639-1 code</td>
<td>Language</td>
<td>Language Family</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>id</td>
<td>Indonesian</td>
<td>Indonesian</td>
</tr>
<tr>
<td>it</td>
<td>Italian</td>
<td>Romance</td>
</tr>
<tr>
<td>ja</td>
<td>Japanese</td>
<td>Japanese</td>
</tr>
<tr>
<td>jv</td>
<td>Javanese</td>
<td>Indonesian</td>
</tr>
<tr>
<td>ko</td>
<td>Korean</td>
<td>Altaic</td>
</tr>
<tr>
<td>la</td>
<td>Latin</td>
<td>Italic</td>
</tr>
<tr>
<td>lt</td>
<td>Lithuanian</td>
<td>Baltic</td>
</tr>
<tr>
<td>lb</td>
<td>Luxembourgish</td>
<td>Germanic</td>
</tr>
<tr>
<td>mk</td>
<td>Macedonian</td>
<td>Slavic</td>
</tr>
<tr>
<td>cmn</td>
<td>mandarin</td>
<td>Chinese</td>
</tr>
<tr>
<td>mr</td>
<td>Marathi</td>
<td>Indic</td>
</tr>
<tr>
<td>no</td>
<td>Norwegian(Bokml)</td>
<td>Germanic</td>
</tr>
<tr>
<td>nn</td>
<td>Norwegian(Nynorsk)</td>
<td>Germanic</td>
</tr>
<tr>
<td>oc</td>
<td>Occitan</td>
<td>Romance</td>
</tr>
<tr>
<td>fa</td>
<td>Persian</td>
<td>Iranian</td>
</tr>
<tr>
<td>pl</td>
<td>Polish</td>
<td>Slavic</td>
</tr>
<tr>
<td>pt</td>
<td>Portuguese</td>
<td>Romance</td>
</tr>
<tr>
<td>ro</td>
<td>Romanian</td>
<td>Romance</td>
</tr>
<tr>
<td>ru</td>
<td>Russian</td>
<td>Slavic</td>
</tr>
<tr>
<td>sk</td>
<td>Slovak</td>
<td>Slavic</td>
</tr>
<tr>
<td>sl</td>
<td>Slovenian</td>
<td>Slavic</td>
</tr>
<tr>
<td>es</td>
<td>Spanish</td>
<td>Romance</td>
</tr>
<tr>
<td>su</td>
<td>Sundanese</td>
<td>Indonesian</td>
</tr>
<tr>
<td>sv</td>
<td>Swedish</td>
<td>Germanic</td>
</tr>
<tr>
<td>tl</td>
<td>Tagalog</td>
<td>Austronesian</td>
</tr>
<tr>
<td>ta</td>
<td>Tamil</td>
<td>Dravidian</td>
</tr>
<tr>
<td>te</td>
<td>Telugu</td>
<td>Dravidian</td>
</tr>
<tr>
<td>th</td>
<td>Thai</td>
<td>Kam-thai</td>
</tr>
<tr>
<td>tr</td>
<td>Turkish</td>
<td>Turkic</td>
</tr>
<tr>
<td>uk</td>
<td>Ukrainian</td>
<td>Slavic</td>
</tr>
<tr>
<td>vi</td>
<td>Vietnamese</td>
<td>Viet-Muong</td>
</tr>
<tr>
<td>cy</td>
<td>Welsh</td>
<td>Celtic</td>
</tr>
</tbody>
</table>

Table A.1: List of languages used in Sections 4.1 and 4.2