UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

SILVIO RICARDO CORDEIRO

Distributional models of multiword expression compositionality prediction

Thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Computer Science

Advisor: Dr. Aline Villavicencio
Coadvisor: Dr. Alexis Nasr
Coadvisor: Dr. Carlos Ramisch

Porto Alegre
January 2018

CIP — CATALOGING-IN-PUBLICATION

Cordeiro, Silvio Ricardo

Distributional models of multiword expression compositionality prediction / Silvio Ricardo Cordeiro. – Porto Alegre: PPGC da UFRGS, 2018.

167 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2018. Advisor: Aline Villavicencio; Coadvisor: Alexis Nasr.

I. Villavicencio, Aline. II. Nasr, Alexis. III. Título.

ABSTRACT

Natural language processing systems often rely on the idea that language is compositional, that is, the meaning of a linguistic entity can be inferred from the meaning of its parts. This expectation fails in the case of multiword expressions (MWEs). For example, a person who is a *sitting duck* is neither a duck nor necessarily sitting. Modern computational techniques for inferring word meaning based on the distribution of words in the text have been quite successful at multiple tasks, especially since the rise of word embedding approaches. However, the representation of MWEs still remains an open problem in the field. In particular, it is unclear how one could predict from corpora whether a given MWE should be treated as an indivisible unit (e.g. *nut case*) or as some combination of the meaning of its parts (e.g. *engine room*). This thesis proposes a framework of MWE compositionality prediction based on representations of distributional semantics, which we instantiate under a variety of parameters. We present a thorough evaluation of the impact of these parameters on three new datasets of MWE compositionality, encompassing English, French and Portuguese MWEs. Finally, we present an extrinsic evaluation of the predicted levels of MWE compositionality on the task of MWE identification. Our results suggest that the proper choice of distributional model and corpus parameters can produce compositionality predictions that are comparable to the state of the art.

Keywords: Distributional semantics. Multiword expressions. Compositionality. Idiomaticity.

Modelos distribucionais para a predição de composicionalidade de expressões multipalavras

RESUMO

Sistemas de processamento de linguagem natural baseiam-se com frequência na hipótese de que a linguagem humana é composicional, ou seja, que o significado de uma entidade linguística pode ser inferido a partir do significado de suas partes. Essa expectativa falha no caso de expressões multipalavras (EMPs). Por exemplo, uma pessoa caracterizada como *pão-duro* não é literalmente um pão, e também não tem uma consistência molecular mais dura que a de outras pessoas. Técnicas computacionais modernas para inferir o significado das palavras com base na sua distribuição no texto vêm obtendo um considerável sucesso em múltiplas tarefas, especialmente após o surgimento de abordagens de *word embeddings*. No entanto, a representação de EMPs continua a ser um problema em aberto na área. Em particular, não existe um método consolidado que prediga, com base em *corpora*, se uma determinada EMP deveria ser tratada como unidade indivisível (por exemplo *olho gordo*) ou como alguma combinação do significado de suas partes (por exemplo *tartaruga marinha*). Esta tese propõe um modelo de predição de composicionalidade de EMPs com base em representações de semântica distribucional, que são instanciadas no contexto de uma variedade de parâmetros. Também é apresentada uma avaliação minuciosa do impacto desses parâmetros em três novos conjuntos de dados que modelam a composicionalidade de EMP, abrangendo EMPs em inglês, francês e português. Por fim, é apresentada uma avaliação extrínseca dos níveis previstos de composicionalidade de EMPs, através da tarefa de identificação de EMPs. Os resultados obtidos sugerem que a escolha adequada do modelo distribucional e de parâmetros de *corpus* pode produzir predições de composicionalidade que são comparáveis às observadas no estado da arte.

Palavras-chave: Semântica distribucional. Expressões multipalavras. Composicionalidade. Idiomaticidade..

ACKNOWLEDGMENTS

My deep gratitude goes first to my advisers, Aline Villavicencio, Alexis Nasr and Carlos Ramisch, who expertly guided me through my graduate education and who shared the excitement of my three years of discovery and in-depth research in the field of NLP. Their unwavering enthusiasm for the field has kept me constantly engaged with my research, and I cannot thank them enough for these years of scientific mentorship.

I would also like to thank the reviewers of my papers, who have immensely contributed to this thesis through their feedback and advice from an external point of view. In particular, I thank the reviewers of this thesis, Timothy Baldwin and Stefan Evert, as well as the two other members of the thesis committee, Antoine Doucet and Helena Caseli.

I would especially like to thank all the coauthors of my papers. In particular, I thank my advisers, Leonardo Zilio, Rodrigo Wilkens, Marco Idiart, Manon Scholivet, Marie Candito, and Agata Savary. I have immensely enjoyed the multiple discussions we have had on how to best pursue the research at hand. I also feel the need to thank Agata Savary one more time. After all, the PARSEME network of researchers might not have existed without her. I am also indebted to the members of PARSEME, with whom I have shared research ideas in multiple occasions.

This thesis would not have been possible without the financial support from the Brazilian PhD scholarship from CAPES, along with the Brazilian project "Simplificação Textual de Expressões Complexas" sponsored by Samsung Eletrônica da Amazônia Ltda, the European project PARSEME (COST IC1207), the French project PARSEME-FR (ANR-14-CERA-0001), as well as the French-Brazilian projects CAMELEON (CAPES-COFECUB #707-11) and AIM-WEST (FAPERGS-INRIA 1706-2551 /13-7).

I would also like to thank my family and friends, who have supported me through these years of PhD studies. A special thanks goes to Robert and Tanya, for their companionship in the hours when the lab lights were off.

Finally, I would like to express my gratitude to everyone who, directly or indirectly, contributed to the completion of this thesis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The ever-growing presence of computers in our society has put forward an increasing demand for new ways of dealing with human-generated content. The field of Natural Language Processing (NLP) has the goal of mediating this interaction between computers and human language, ranging from the interpretation of written texts on the web to the interaction with spoken commands on hand-held devices. A common theme to many of the NLP tasks is the requirement of *semantic* interpretation (i.e. determining the meaning of the text).

One of the most fundamental assumptions in the field of semantics is that the meaning of a phrase, expression or sentence can be determined from the meanings of its parts. Part of the appeal of this *principle of compositionality*[1] is that it implies that a meaning can be assigned by humans even to sentences that have never been seen before, through the combination of the meaning of familiar words (GOLDBERG, 2015). In the case of NLP, semantic composition can also be an attractive way of deriving the meaning of larger units from their smaller parts. By employing the principle of compositionality, one could design generic NLP systems, able to perform the semantic interpretation of any text.

The representation of the meaning of individual words and their combinations in computational systems has often been addressed by *distributional semantic models* (DSMs). DSMs are based on Harris' distributional hypothesis that the meaning of a word can be inferred from the context in which it occurs (HARRIS, 1954; FIRTH, 1957). In these models, words are usually represented as vectors that, to some extent, capture cooccurrence patterns in corpora. These vectors are assumed to be good proxies for meaning representations. Traditionally, a vector can be built for a target word by explicitly counting all its cooccurrences with context words (LIN, 1998; LANDAUER; FOLTZ; LAHAM, 1998). These models, also known as *count-based models* (BARONI; DINU; KRUSZEWSKI, 2014), result in sparse vectors that are often projected into a low-dimensionality space using a statistical technique such as singular value decomposition. The more recent neural-network models, often referred to as *word embeddings*, also represent words as real-valued vectors projected onto some low-dimensional space, but these are obtained as a by-product of training a neural network to learn a function between words and their contexts (MIKOLOV et al., 2013).

[1] Attributed to Frege (1892/1960).

Evaluation of DSMs has focused on obtaining accurate semantic representations for single words, and it is on this basis that many optimizations have been proposed (LIN, 1999; ERK; PADÓ, 2010; BARONI; LENCI, 2010). For instance, state-of-the-art models are already capable of obtaining a high level of agreement with human judgments for predicting synonymy or similarity between single words (FREITAG et al., 2005; CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2015; LAPESA; EVERT, 2017). Although there seems to be a reasonable understanding of the strengths and weaknesses of vector representations for single words, the same is not true for larger units such as sentences. There exist some proposals for modeling the *composition* of individual words to create representations for larger units such as phrases, sentences and even whole documents (MITCHELL; LAPATA, 2010; MCCARTHY; KELLER; CARROLL, 2003; REDDY; MCCARTHY; MANANDHAR, 2011; MIKOLOV et al., 2013; FERRET, 2014). They include the use of simple additive and multiplicative vector operations (MITCHELL; LAPATA, 2010), syntax-based lexical functions (SOCHER et al., 2012), and the application of matrices and tensors as word-vector modifiers (BARONI; LENCI, 2010; BRIDE; CRUYS; ASHER, 2015). These operations usually assume the principle of compositionality when building representations for larger units.

However, this assumption is challenged in the case of idiomatic expressions, whose meanings may not be straightforwardly related to their parts (SAG et al., 2002). In fact *multiword expressions* (MWEs) display a wide spectrum of idiomaticity, from more compositional to more idiomatic cases (BALDWIN; KIM, 2010). For instance, although the meaning of *olive oil* can be derived from its parts (as *oil* extracted from *olives*), this is not the case for *snake oil*, which is used to refer to any product of questionable benefit (not necessarily *oil* and certainly not extracted from *snakes*). Such constructions are notoriously challenging for semantically-focused systems, as they are very numerous in a speaker's lexicon (JACKENDOFF, 1997), but they are often not *compositional.* For example, a non-compositional expression such as *dead end* should not be literally translated into French as **fin morte*, as it would lose its intended meaning. It is therefore crucial to determine to what degree the principle of compositionality applies to a specific expression to ensure its correct semantic interpretation.

The task of *compositionality prediction* consists in assigning a numerical score to a word combination indicating to what extent the meaning of the whole combination can be directly computed from the meanings of its component words. This score can then be used to decide how the combination should be represented in downstream tasks and applications. Given that idiomatic expressions are quite frequent in human languages, compositionality prediction is relevant to any NLP task and application that performs some form of semantic processing. For instance, in machine translation, idiomatic expressions must be translated as an indivisible whole (CAP et al., 2015; SALEHI et al., 2015; CARPUAT; DIAB, 2010; REN et al., 2009). In semantic parsing, one needs to identify complex predicates and their arguments to avoid erroneous analyses (JAGFELD; PLAS, 2015; HWANG et al., 2010). For word-sense disambiguation, no sense should be ascribed to the individual words pertaining to an idiomatic expression (SCHNEIDER et al., 2016a; KULKARNI; FINLAYSON, 2011). In all of these cases, there is a need for the preprocessing task of *MWE token identification* in running text, and this operation may benefit from the availability of compositionality scores.

In this thesis, we discuss approaches for automatically predicting the compositionality of MWEs on the basis of their semantic representation and those of their component words represented using DSMs. To determine to what extent these models are adequate cross-lingually, we evaluate them in three languages, English, French and Portuguese. Since MWEs encompass a large amount of related but distinct phenomena, we focus exclusively on a sub-category of MWEs: *nominal compounds* (NCs).[2] Nominal compounds (such as *nut case* or *milk tooth*) represent an ideal case study for the work in this thesis, thanks to their relatively homogeneous syntax (as opposed to e.g. verbal idioms such as *take into account*, which may take internal arguments and modifiers), as well as their pervasiveness in the languages under consideration. We assume that, in the future, models able to predict the compositionality of nominal compounds could be generalized to include other categories of MWEs by addressing their morphological and syntactical variability.

By using DSM instances to predict the compositionality of nominal compounds, we are also indirectly evaluating those instances themselves. While evaluations of DSMs based on single words abound, their evaluation on tasks involving MWEs are currently lacking. Some notable exceptions include the works of Reddy, McCarthy and Manandhar (2011), who compare additive and multiplicative combinations of traditional DSMs, and of Salehi, Cook and Baldwin (2015) who look at addition-based models for compositionality prediction using both traditional and neural-network DSMs (see Section 4.1 for more details). However, these works do not explore the vast landscape of existing DSM configurations, and may be unable to draw conclusions that are generalizable across languages, DSMs, their parameters and the corpora they are learned on.

[2] A generalization over compound nouns, see Section 2.3.1.

The main goal of this thesis is to bridge this gap by presenting a framework for MWE compositionality prediction along with a broad cross-lingual evaluation. We evaluate both intrinsically and extrinsically to what extent DSMs can accurately model the semantics of NCs with various levels of compositionality compared to human judgments. The following section details the contributions that allow us to reach this goal.

1.1 Contributions

The main contributions of this thesis are the following:

Compositionality dataset We construct and evaluate three datasets containing NCs ranging from fully compositional to fully idiomatic. These NCs were manually annotated based on their degree of compositionality. The datasets span multiple languages (English, French and Portuguese), and can be useful in the evaluation of MWE compositionality prediction techniques. Section 3.1 presents a detailed account of the data collection process. The dataset has also been described in a publication (RAMISCH et al., 2016).

Dataset analysis We report the results of a thorough analysis of the three constructed datasets, studying the correlation between compositionality and related linguistic variables. Part of these results has been published in Ramisch et al. (2016) (focusing on the distribution of annotations) and in Cordeiro, Ramisch and Villavicencio (2016a) (focusing on inter-annotator agreement). Section 3.2 expands on these results by analyzing the correlation between human-rated scores and distributional characteristics of the NCs.

Compositionality prediction framework We propose a language-independent framework for the prediction of the degree of compositionality in MWE expressions. As part of this framework, we also systematize a set of parameters that can be evaluated across different DSMs, allowing a sound comparison of multiple DSMs under a variety of settings. In Chapter 4, we extend the underlying model with the possibility of six compositionality prediction strategies — two of which are original strategies (*maxsim* and *geom*), proposed in the scope of this thesis. The implementation of this framework is freely available as part of the mwetoolkit. The predictive framework has been described in a publication (CORDEIRO; RAMISCH; VILLAVICENCIO, 2016b).

Intrinsic evaluation We evaluate the proposed compositionality prediction model under a variety of settings: different DSMs, different DSM parameters, different corpora parameters, different prediction strategies. Chapter 5 of the thesis extends these results with predictions for Portuguese datasets, including previously unpublished results for one DSM (lexvec) and multiple prediction strategies. We additionally evaluate corpus-specific parameters such as corpus size and a new technique of parallel predictions. Furthermore, we consolidate the interpretation of these results through a large set of previously unpublished sanity checks and detailed error analyses. The results have been published in Cordeiro et al. (2016).

Extrinsic evaluation We perform an extrinsic evaluation of the proposed compositionality prediction model by using predicted scores as features in the task of MWE identification. Chapter 6 presents the implementation of an MWE identifier based on syntactic patterns (CORDEIRO; RAMISCH; VILLAVICENCIO, 2016c), and compare its accuracy with the one achieved by a technique of sequence modeling, with and without the help of the predicted scores (SCHOLIVET; RAMISCH; CORDEIRO, 2017).

1.2 Investigated hypotheses

This thesis investigates a series of hypotheses concerning MWEs and their compositionality, both in relation to the human perception of compositionality and in relation with DSM-based representations. The hypotheses described in this section guide our work — whose main goal is to propose, implement and evaluate distributional methods for the compositionality prediction of MWEs.

Central to this thesis is a framework of compositionality prediction based on DSM representations of semantics. The main assumption behind this framework is that, when the semantics of a compositional MWE can be derived from a combination of its parts, this should be reflected in DSMs. In particular, the vector for the compositional MWE should be similar to the combination of the vectors of its parts. Conversely we can use the lack of similarity between the MWE vector representation and a combination of its parts to detect non-compositionality. We formulate this assumption in the form of the general hypothesis $\mathrm{h}_{\text{pred-comp} \approx \text{comp}}$: **MWE compositionality as assessed by human annotators is correlated with compositionality predictions**, where the predictions are based on the distributional representation of MWE elements and MWEs themselves.

We begin our work with the construction of three datasets of nominal compounds with human-annotated compositionality scores. This dataset is analyzed so as to evaluate the hypothesis $\mathrm{h}_{\mathrm{idiom}\,\approx\,\mathrm{distr}}$: **idiomaticity is correlated with distributional characteristics of MWEs**. In particular, we consider the following sub-hypotheses:

- $\mathrm{h}_{\mathrm{idiom}\,\approx\,\mathrm{distr.freq}}$ The level of idiomaticity of an MWE is positively correlated with its frequency. The intuition is that exceptional constructions (such as idiomatic MWEs) need to be frequent to ensure their survival in the language (PINKER, 1995).

- $\mathrm{h}_{\mathrm{idiom}\,\approx\,\mathrm{distr.convent}}$ The level of idiomaticity of an MWE is positively correlated with its level of conventionalization. This follows from the literature on MWE type extraction, which uses estimators of conventionalization to identify the idiomatic expressions among a list of MWE candidates (FAZLY; STEVENSON, 2006; BU; ZHU; LI, 2010; GURRUTXAGA; ALEGRIA, 2013; MAAROUF; OAKES, 2015).

For each dataset, we instantiate DSMs under a variety of configurations, generating a total of more than 8 thousand sets of compositionality predictions. We then evaluate what kinds of variables may influence the accuracy of the highest-ranking configurations. We consider the general hypothesis $\mathrm{h}_{\mathrm{accur}\,\leftarrow\,\mathrm{MWE}}$: the **accuracy of the model depends on MWE-specific properties**, and we formulate four non–mutually-exclusive sub-hypotheses:

- $\mathrm{h}_{\mathrm{accur}\,\leftarrow\,\mathrm{MWE.idiom}}$ The accuracy of predicted scores is higher for MWEs that were classified by humans as more compositional (i.e. less idiomatic). The intuition is that DSM representations should be more faithful to the reality for compositional MWEs, which follow the regularities that are normally exploited in other works in the literature (MITCHELL; LAPATA, 2010; MIKOLOV et al., 2013).

- $\mathrm{h}_{\mathrm{accur}\,\leftarrow\,\mathrm{MWE.diffic}}$ The accuracy of predicted scores is lower for MWEs that are more difficult to annotate for humans, as measured through the level of agreement among annotators. The intuition is that one would expect the predictive model to have difficulty in the same MWEs that posed a problem for humans, either due to some inherent difficulty in the MWE or due to less reliability of the data.

- $\mathrm{h}_{\mathrm{accur}\,\leftarrow\,\mathrm{MWE.freq}}$ The accuracy of predicted scores is positively correlated with the frequency of the MWE in the corpus. The intuition is that low-frequency expressions should have a less trustworthy representation inside DSMs.

- $\text{h}_{\text{accur} \leftarrow \text{MWE.convent}}$ The accuracy of predicted scores is positively correlated with the conventionalization of the MWE. The intuition is that the elements of highly conventionalized MWEs are more likely to share contexts, and thus have a more compatible vector representation.

In addition to these MWE-centric hypotheses, we consider the specific choice of DSMs and internal parameters in the task of compositionality prediction. The hypothesis $\text{h}_{\text{accur} \leftarrow \text{DSM}}$ is that the **accuracy of the model depends on DSM-specific parameters**. In particular, we consider two sub-hypotheses:

- $\text{h}_{\text{accur} \leftarrow \text{DSM.window}}$ The accuracy of compositionality prediction depends on the amount of content that is taken into account at each occurrence of a word (i.e. the size of the context window). More context should lead to a more precise representation and thus result in better predictions.

- $\text{h}_{\text{accur} \leftarrow \text{DSM.dims}}$ The accuracy of compositionality prediction depends on the number of dimensions in each vector generated by the DSM. A higher number of dimensions should allow for a more fine-grained representation of the data.

Along with an influence from DSM-specific parameters, we also consider the impact of different corpus-specific configurations. The hypothesis we evaluate is $\text{h}_{\text{accur} \leftarrow \text{corpus}}$: **accuracy of the model depends on corpus-specific parameters**. We evaluate the following sub-hypotheses:

- $\text{h}_{\text{accur} \leftarrow \text{corpus.wordform}}$ Higher-quality compositionality predictions are obtained when the corpus is preprocessed so as to reduce the sparseness of the word occurrences (e.g. through lemmatization). The intuition is that the reduction in sparseness should allow DSMs to generate vectors from a more varied number of contexts, and that these vectors would thus be more robust than the ones generated without preprocessing.

- $\text{h}_{\text{accur} \leftarrow \text{corpus.size}}$ DSM representations built from larger corpora outperform representations built from smaller corpora. The intuition is that more varied occurrences of each word allow the construction of DSM representations that more faithfully correspond to the actual semantics.

- $\text{h}_{\text{accur} \leftarrow \text{corpus.parallel}}$ Multiple parallel DSM representations built from different parts of the corpora can be combined to achieve equivalent compositionality predictions. The intuition is that these representations could still provide a high-quality description of the underlying semantics, while allowing for the final predictions to be calculated through a combination of multiple computational resources (e.g. computer clusters).

Regarding the compositionality prediction model, we consider six different predictive strategies (defined in Section 5.4). The hypothesis $\mathrm{h}_{\text{strat}}$ is that the **accuracy of the model depends on the predictive strategy**. In particular, we consider these three sub-hypotheses:

- $\mathrm{h}_{\text{strat.partial-info}}$ Predictions derived only from parts of an MWE (i.e. its syntactic head) will be less accurate than predictions that consider all words in the MWE. The intuition is that, by using only part of the available distributional information, these strategies are limited in their ability of predicting the compositionality of the MWE as a whole.

- $\mathrm{h}_{\text{strat.maxsim}}$ We can improve the score prediction of compositional MWEs through a strategy that favors compositional interpretations. More specifically, we can improve predictions if we assign weights to the vector representation of each member word of an MWE so as to maximize its compositionality score. The *maxsim* strategy proposed in this thesis models this assignment of weights favoring composition.

- $\mathrm{h}_{\text{strat.geom}}$ We can improve the score prediction of idiomatic MWEs through a strategy that favors idiomatic interpretations. More specifically, we propose the *geom* strategy that multiplies individual predictions of compositionality for all words in an MWE. If any word has been identified as idiomatic (i.e. having a small level of compositionality), the compositionality score of the MWE as a whole will be reduced.

One final hypothesis we consider is related to the use of predicted compositionality scores in an extrinsic evaluation. Here, we consider the task of MWE token identification, which can be seen as an important preprocessing step to semantic applications (such as machine translation or text simplification). The relevant hypothesis is $\mathrm{h}_{\text{pred-comp} \rightarrow \text{ident-accur}}$: **predicted compositionality scores are useful in the task of MWE identification**. This hypothesis is evaluated by a comparison of the accuracy of the MWE identification with and without the use of predicted compositionality scores.

1.3 Context of the thesis

The work in this thesis has been produced as part of a joint supervision (cotutelle) between two universities: Universidade Federal do Rio Grande do Sul (UFRGS, Brazil) and the Aix-Marseille Université (AMU, France). I have spent part of my research time in each university, and most of the publications derive from this cooperation.

This thesis is centered on the topic of multiword expressions (MWEs) and distributional semantic models (DSMs), which are two areas of research that have been growing in the latest decade. The increasing interest in MWE research has been the main motivation behind the formation of the PARSEME network of researchers (ICT COST Action IC1207). Some of the work done in conjunction with PARSEME researchers, notably the shared task on verbal MWE identification (SAVARY et al., 2017b), falls out of the scope of this thesis, but has been paramount to a broadening of my view of the field. Moreover, part of the work described in this thesis was carried out in the context of PARSEME-FR (ANR-14-CERA-0001), a French-language spin-off of PARSEME.

This thesis is also closely related to a software called mwetoolkit (RAMISCH, 2015), which was both used and extended during this thesis. The mwetoolkit is one of the major resources published by the research group in the side of UFRGS, and it is currently being maintained in joint work with Carlos Ramisch in AMU. The contributions from this thesis are integrated and freely available as part of the mwetoolkit.[3]

1.4 Publications

I present below a list of papers that have been published or accepted for publication during the period of my PhD studies. The papers that are directly relevant to the topic of this thesis are the following ones:

- Cordeiro, Ramisch, Idiart, Villavicencio. *Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time.* In: ACL 2016.

- Ramisch, Cordeiro, Zilio, Idiart, Villavicencio, Wilkens. *How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality.* In: ACL 2016.

[3] <http://mwetoolkit.sf.net>

- Cordeiro, Ramisch, Villavicencio. *mwetoolkit+sem: Integrating Word Embeddings in the mwetoolkit for Semantic MWE Processing.* In: LREC 2016.

- Cordeiro, Ramisch, Villavicencio. *Filtering and Measuring the Intrinsic Quality of Human Compositionality Judgments.* In: MWE 2016.

- Cordeiro, Ramisch, Villavicencio. *UFRGS&LIF: Rule-Based MWE Identification and Predominant-Supersense Tagging.* In: SemEval 2016.

- Scholivet, Ramisch, Cordeiro. *Sequence Models and Lexical Resources for MWE Identification in French.* In: PMWE 2017. Submitted for review.

- Wilkens, Zilio, Cordeiro, Paula, Ramisch, Idiart, Villavicencio. *LexSubNC: A Dataset of Lexical Substitution for Nominal Compounds.* In: IWCS 2017.

- Cordeiro, Ramisch, Villavicencio. *Token-based MWE Identification Strategies in the mwetoolkit.* In: PARSEME 2015.

- Cordeiro, Ramisch, Villavicencio. *Nominal Compound Compositionality: A Multilingual Lexicon and Predictive Model.* In: PARSEME 2016.

- Cordeiro, Ramisch, Villavicencio. *MWE-aware corpus processing with the mwetoolkit and word embeddings.* In: W-PROPOR 2016.

Other papers published (or submitted for review) during this period are not directly implicated in the present thesis. Nevertheless, they all have in common the topic of MWE processing and semantic representation:

- Savary, Ramisch, Cordeiro, Sangati, Vincze, QasemiZadeh, Candito, Cap, Giouli, Stoyanova, Stoyanova, Doucet. *The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.* In: MWE 2017.

- Zilio, Wilkens, Möllmann, Wehrli, Cordeiro, Villavicencio. *Joining forces for multiword expression identification.* In: PROPOR 2016.

- Savary, Ramisch, Cordeiro, Candito, Vincze, Sangati, QuasemiZadeh, Giouli, Cap, Stoyanova, Stoyanova, Doucet. *The PARSEME shared task on automatic identification of verbal multiword expressions.* In: PMWE 2017. Submitted for review.

- Savary, Candito, Mititelu, Bejček, Cap, Čéplö, Cordeiro, Eryiğit, Giouli, Gompel, Hacohen-Kerner, Kovalevskaite, Krek, Liebeskind, Monti, Escartín, Plas, Qasemizadeh, Ramisch, Sangati, Stoyanova, Vincze. *PARSEME multilingual corpus of verbal multiword expressions.* In: PMWE 2017. Submitted for review.

1.5 Thesis structure

The remainder of this thesis is structured as follows:

- Chapter 2 presents the terminology and necessary background in the statistical representation of languages. It also presents a literature review on MWE research, as well as related work on the semantic representation of words and MWEs.
- Chapter 3 describes the methodology used in the construction of three new datasets of nominal compounds annotated with compositionality scores. It also presents an analysis of the score distribution and difficult of annotation, as well as the correlation between the annotated score and distributional characteristics of the MWEs.
- Chapter 4 presents our framework of MWE compositionality prediction, with a detailed description of the experimental setup that will be used in the following Chapter.
- Chapter 5 presents a large-scale intrinsic evaluation of our model of compositionality prediction against datasets of human-rated compositionality scores, focusing on nominal compounds as a specific category of MWE.
- Chapter 6 presents an extrinsic evaluation of our model of compositionality prediction, in the form of its application in the task of MWE identification.
- Chapter 7 presents some conclusions and perspectives of future work.

2 BACKGROUND

This chapter describes the background information that is essential for the remainder of the thesis. Section 2.1 defines the basic terminology that will be used in the rest of the thesis. Section 2.2 presents the use of statistics in NLP, going from co-occurrence measures to evaluation methods. Section 2.3 then presents the motivation and challenges associated with the research on MWEs. Finally, Sections 2.4 and 2.5 present an overview of lexical semantic representations in NLP, on the level of single words and MWEs. Since it constitutes the core of this thesis, the state of the art on compositionality prediction will be presented later, in Section 4.1, along with the definition of the model that we propose and evaluate.

2.1 Basic terminology

The work in this thesis focuses on the interpretation of written texts. Linguists often classify the interpretation of texts into different layers of abstraction. For example, consider the sentence *"this student published a paper"*. These are some of the levels in which this sentence can be analyzed:

- Morphology: A word can change its base form to express a change of grammatical category (derivational morphology; e.g. *"publish"* → *"publisher"*) or to represent a variation such as gender, number or tense (inflectional morphology; e.g. *"publish"* → *"published"*).

- Syntax: Just like morphology works on the level of words and their internal structure, syntax can be used to analyze sentences based on their internal structure (word order). Such analysis, when applied to the above sentence, could inform us that *"published"* is the main verb, and that *"this student"* and *"a paper"* are respectively the subject and direct object of this verb.

- Semantics: The field of semantics looks at the meaning of words and phrases. For example, although the word *"paper"* is often used as reference to the material that is created from cellulose, in this particular context it should be interpreted as a scientific document (e.g. it can be in electronic format).

- Pragmatics: While one may be able to build an accurate abstract mental model for a word such as *“student”* given only the surrounding words, the specific reference in *“this student”* can only be resolved by taking the underlying context into account. If this sentence is uttered by a person who is pointing at someone else, for example, one may conclude through pragmatics who is the referent of *“this student”*.

In the area of NLP, the operations that can be performed on texts are usually conceptually organized in tasks. Many of these tasks may take as an input a collection of sentences and have as an output the same sentences with an added layer of information (JURAFSKY; MARTIN, 2008). Tasks are often performed in succession as a pipeline, where the output of a task is the input for the next one. These are some of the tasks that are often performed in any end-to-end application in NLP involving text analysis:

- Tokenization: The goal of tokenization is to break a written text into *tokens*, which correspond somewhat to the linguistic notion of *words*. While the precise definition of word depends on language-specific semantics, a *token* is a pragmatic concept used in NLP, and its definition will often vary depending on the tokenizer at hand. For example, the sentence

  > `Mr. Smith doesn’t eat bananas with a fork.`

  could be tokenized as such:

  > `Mr. Smith does n’t eat bananas with a fork .`

  The punctuation associated with the abbreviations (e.g. *“Mr.”*) is often tokenized along with the preceding characters, while the period at the end of the sentence is considered a token apart. Also note that the contraction *“doesn’t”* may be separated in two tokens, indicating the two underlying words *“do+not”*. The use of tokens allows a trade-off between linguistic accuracy and practicality of implementation.

- POS tagging: In a given context, every word can be associated with a grammatical class, known as its Part-of-Speech (POS). The task of POS tagging consists in identifying the POS of each token in the text according to a *tagset*. A list of POS tags for the sentence above could be the following one:

  > `PROPN PROPN AUX PART VERB NOUN ADP DET NOUN PUNCT`

These tags follow the Universal POS tagset, which standardizes a common set of tags across a variety of languages (PETROV; DAS; MCDONALD, 2011).

- Lemmatization: The goal of lemmatization is to identify a canonical form for the tokens in the text. For each token in its *surface form*, the lemmatizer will produce the corresponding *lemma*. For the example above, these could be the resulting lemmas:

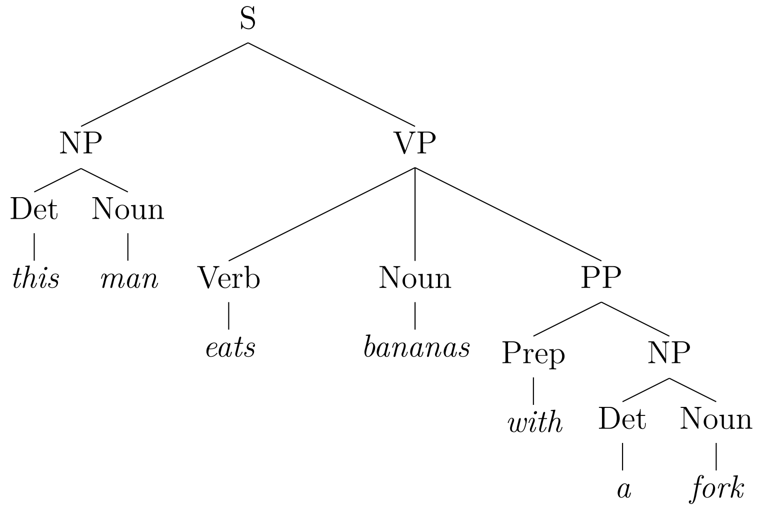
  > `Mr. Smith do not eat banana with a fork .`

  Lemmatization is usually applied to neutralize distinctions caused by morphological inflection (e.g. *"publish/publishes/published/publishing"* → *"publish"*).

- Parsing: In the goal of understanding a sentence, it is often useful to have a representation of how words are grouped to form larger structures. This syntactic information can be encoded in different ways, and it usually corresponds to one of two classes of grammar theories: phrase structure and dependency grammar.

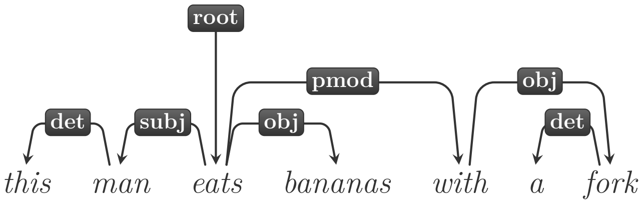
  Figure 2.1 presents a syntax tree for the sentence *"this man eats bananas with a fork"* according to the phrase structure grammar. In this representation, words that behave syntactically as a single unit are grouped as *phrases* (also known as constituents). For example, the phrase *"this man"* behaves as a single noun, and is thus grouped under a single node of the tree, known as a noun phrase (NP).

Figure 2.1: Representing syntax: constituency tree.



In a dependency tree representation, words are connected by labeled edges, in what is known as a dependency relation (see Figure 2.2). This relation connects *heads* (e.g. a verb) to their *dependents* (e.g. a verb's subject). The root of the tree is usually the main verb (in the example, the verb *eats*), and everything else is connected to this verb through a chain of dependency relations.

Figure 2.2: Representing syntax: dependency tree.



Many works in NLP are centered around the idea of a *corpus*. A corpus is a large body of (written or spoken) text that can be considered representative of a human language (MITKOV, 2005). In NLP, this is often taken to mean that corpora should be as large as possible. The construction of a corpus is often performed automatically (e.g. by crawling the web), digitally storing the data according to some file format for further processing. In the case of written corpora, the preprocessing pipeline may include tasks such as the aforementioned tokenization, POS tagging and lemmatization.

One important distinction that must be made in NLP is the one between *types* and *tokens*. While a given corpus may have many concrete instances of a given word (tokens), it is often useful to talk about the unique concept that instantiated those tokens (the type). In the same manner that tokens can be assembled in sentences to form corpora, types can be assembled in dictionary entries. Such dictionaries are also known as lexicons, or more generally as lexical resources. The set of all types instantiated in a corpus is known as its *vocabulary*, which we will denote in this thesis as V.

Finally, most applications in NLP involve some kind of prediction, where the computational system attempts to replicate the outcome that a human would have produced. This prediction is then compared against a blind *test set*. Test sets are often handcrafted based on direct human knowledge, which forms an authoritative sample of test cases and their solutions (in which case it is also known as a *gold standard*). Some systems are designed so as to perform their predictions solely based on a programmed algorithm. Others, known as *supervised* approaches, require a *training set* with examples from which the system may learn to perform the predictions (MANNING; SCHÜTZE, 1999; JURAFSKY; MARTIN, 2008)

2.2 Language and statistics

Now that we have defined the basic terminology, we turn to a statistical view of the properties of human language. We start with with some background on the occurrence counts of isolated words (Section 2.2.1) and of word pairs in the same context (Section 2.2.2). We then present measures of association between word pairs (Section 2.2.3). We also consider the creation of human-rated datasets, focusing on measures of inter-rater agreement (Section 2.2.4). Finally, we consider different ways of measuring the accuracy of computational predictions of human annotation, both in the case of continuous data (Section 2.2.5) and in the case of categorical data (Section 2.2.6).

2.2.1 Occurrence counts

Many properties of words can be inferred from their statistical behavior with regards to human language. One of the simplest statistical measures that can be considered is the number of *occurrences* of each word in a corpus. For example, consider the following three toy corpora: English-language (`EN`) Wikipedia entry "U.S.A", French-language (`FR`) Wikipedia entry "France" and Portuguese-language (`PT`) Wikipedia entry "Brasil" (*Brazil*).[1] Table 2.1 presents the tokens with highest number of occurrences (#occur) in each of these toy corpora.

Two notable categories of words can be seen in the table:

- Words that have been biased by the input corpus (e.g. *States*). These words are an artifact of the chosen corpora, and would have a notably lower ranking in a more general corpus.

[1] Wikipedia entries collected on 2017-08-10.

Table 2.1: Most frequent words in toy corpus.

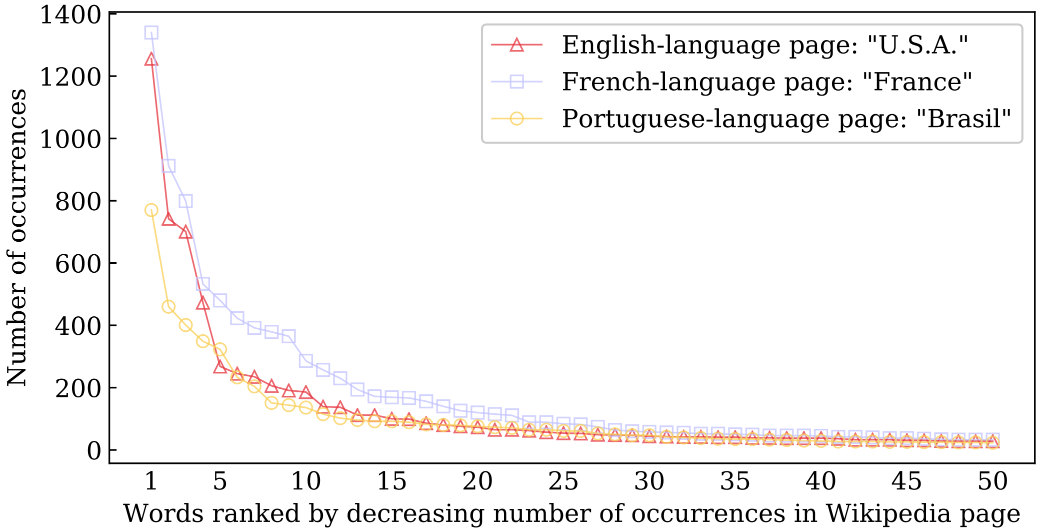
Rank	#occur	EN word	#occur	FR word	#occur	PT word
1	1256	the	1340	de *(of)*	770	de *(of)*
2	742	of	912	la *(the)*	460	e *(and)*
3	701	and	799	et *(and)*	401	do *(of+the)*
4	473	in	533	le *(the)*	349	a *(the/to)*
5	268	The	480	des *(the/of+the)*	323	o *(the)*
6	245	United	423	en *(in)*	233	em *(in)*
7	235	to	392	les *(the)*	204	da *(of+the)*
8	206	a	379	du *(of the)*	151	Brasil *(Brazil)*
9	191	States	365	à *(to)*	144	no *(in+the)*
10	186	is	286	est *(is)*	136	que *(that)*

- Words that only convey a general meaning (e.g. prepositions, articles), serving mainly to connect other words according to the underlying grammar. These words are known as function words, and contrast with the more semantically distinguished content words (e.g. nouns, verbs). In NLP, a similar concept to function words is the one of *stopwords*. When dealing with semantic tasks, one very common step of corpus preprocessing involves the removal of stopwords from the text. The definition of stopwords may be based on their POS tag (e.g. removing all prepositions from the text) or based on a list of e.g. the 50 or 100 words with the highest number of occurrences in the corpus.

When looking at the number of occurrences per word, we can also see a clear pattern: the number of occurrences of each word decreases proportionally to its ranking. This pattern, known as Zipf's Law, can more easily be seen in a graph, as in Figure 2.3, which shows the number of occurrences of the top 50 most frequent words in the three Wikipedia pages. If we consider the number of occurrences for the English word at rank 5, we see that it is about 5 times smaller than the number of occurrences for the word at rank 1. Similarly, the word at rank 10 appears about 10 times less often in the corpus.

The above effects can be observed in all human languages, even in these severely small corpora (e.g. the English toy corpus has less than 20 thousand tokens). When comparing across different languages, one might be tempted to consider the number of occurrences itself, but this number presents a pitfall: it is directly dependent on the size

Figure 2.3: Most frequent words in toy corpus.



of the corpus. An alternative solution is the use of *frequency*, defined as the ratio between the number of occurrences and the size of the underlying corpus:

$$\text{freq}(w) = \frac{\#\text{occur}(w)}{\text{corpus size}}.$$

Frequency values range between 0.0 and 1.0 regardless of corpus size, and can be thus regarded as a normalized version of the number of occurrences.

2.2.2 Co-occurrence counts

Similarly to how one may derive statistical information from the occurrences of isolated words, one may consider the statistical properties of pairs of words that *co-occur* (i.e. that occur in the same context). Co-occurrence may be calculated based on syntactic information (i.e. two words co-occur if they are connected in a dependency tree), or on a sliding window of adjacent words (i.e. two words are considered to co-occur if there are at most k other words in between). For example, Table 2.2 presents the top 10 most frequent co-occurrences of the pattern *"federal <noun>"* in the toy corpus, along with their co-occurrence counts (#co-occur).[2] In this thesis, we will refer to every word of interest (e.g. *"federal"*) as a *target word*. The words that co-occur with a target word will be called *context words* (e.g. *"government"*, *"courts"*, etc).

This type of word-pair information is often structured as a *co-occurrence matrix*.

[2] Extracted using the sliding window method with $k = 0$ (i.e. only adjacent word pairs were considered)

For a vocabulary V of size $|V|$, a co-occurrence matrix of dimension $|V| \times |V|$ contains the number of occurrences of each (target, context) pair as seen in the corpus. One can think of these matrices as big square tables of numbers, with $M_{i,j}$ representing the value at row i and column j. Words are mapped to row/column indexes (e.g. *"government"* could be word$_{53}$), so that both row$_i$ and column$_i$ refer to word$_i$. The co-occurrence between word$_j$ and word$_k$ can then be obtained at the matrix position $M_{j,k}$. Due to the symmetry of this statistical relation, the matrix position $M_{k,j}$ yields the same result.

Table 2.2: Most frequent word pairs involving *"federal"*.

#co-occur	word pair
9	federal government
4	federal courts
4	federal district
4	federal level
4	federal taxes
2	federal debt
2	federal law
2	federal outlays
2	federal republic

According to the distributional hypothesis, the meaning of words is associated with the context that they share. For the example in Table 2.2, all of these nouns have the word *"federal"* as a context in common, from which we can infer that they have related meaning. Moreover, the co-occurrence of two words can be seen as an evidence that they are semantically "close" in a sense — e.g. these nouns are closer in meaning to the word *"federal"* than other nouns not shown here. The underlying idea is that semantically related words tend to be used together, and hence words that often appear together can be assumed to be semantically related. For example, when talking about a *"government"*, one will often use the word *"federal"*. Similarly, among all things *"federal"*, the *"government"* is a notable example that will come to a speaker's mind. Co-occurrence counts can also be used to build vector models of word semantics (as presented in Section 2.4.2).

2.2.3 Association measures

Lexical co-occurrence can be considered a rudimentary way of estimating the level of *association* between a pair of words. Word association is a statistical property that can indicate the predictability of a specific combination of words. For example, the word *“population”* appears 41 times in the text, but since none of these is paired up with the word *“federal”*, it is not taken into account when representing its semantics through co-occurrence, suggested that these two words are not associated.

There is a downside to the use of co-occurrence counts as measures of association: some word pairs may co-occur with high frequency solely due to the fact that one of the words in the pair is frequent itself. For example, even in a syntactically restricted word-pair pattern such as *“federal <word>”*, the toy corpus contains 2 occurrences of *“federal and”*, in the coordinations *“federal and state”* and *“federal and military”*. The word *“and”* is not particularly meaningful to the definition of *“federal”*, and its co-occurrence with *“federal”* is due to its high overall frequency (788 occurrences in the toy corpus).

Other measures can better capture the association between word-pairs by reducing the effect of words that have high isolated corpus occurrence counts. One such measure is the pointwise mutual information (PMI), which can be defined as

$$\mathrm{PMI}(w_1, w_2) = \log\left(\frac{\mathrm{freq}(w_1 w_2)}{\mathrm{freq}(w_1)\ \cdot\ \mathrm{freq}(w_2)}\right).$$

PMI considers the co-occurrence frequency as well as the individual frequencies of the words. Its value can range from negative infinity to positive infinity. In an extreme case, if the word w_1 always appears alongside the word w_2 in the corpus, $\mathrm{freq}(w_1) = \mathrm{freq}(w_1 w_2)$, and thus $\mathrm{PMI} = \log\left(\frac{1}{\mathrm{freq}(w_2)}\right) = \log\left(\frac{\text{corpus size}}{\#\mathrm{occur}(w_2)}\right)$, which can reach high values for moderate numbers of occurrence of word_2. On the other hand, if word_1 appears too often by itself and not as often in conjunction with word_2, the denominator will be considerably high, and thus the PMI between the two words will be the logarithm of a low number — i.e. a low number itself (CHURCH; HANKS, 1990).

Table 2.3 presents the number of individual occurrences, word-pair occurrence and PMI for the word pairs mentioned in Table 2.2. Note how the PMI is close to zero for the non-associated word pair *“federal and”*, but is much higher when calculated between the word *“federal”* and highly associated nouns (e.g. *“government”*). Moreover, even among these associated pairs, we can see a distinction in the strength of PMI: the highest score belongs to *"federal outlays"*, reflecting the fact that *"outlays"* only appears in the corpus in conjunction with the word *"federal"*. A significantly lower score is obtained by *"federal law"*, as the word *"law"* often appears in other contexts inside the corpus (with only 2 out of the 19 occurrences being the expression *"federal law"*).

Table 2.3: PMI between *"federal"* and associated nouns.

word$_1$	word$_2$	#occur(w_1)	#occur(w_2)	#co-occur	PMI
federal	government	54	38	9	4.4
federal	courts	54	7	4	5.3
federal	district	54	10	4	4.9
federal	level	54	16	4	4.4
federal	taxes	54	15	4	4.5
federal	debt	54	9	2	4.3
federal	law	54	19	2	3.6
federal	outlays	54	2	2	5.8
federal	republic	54	5	2	4.9
federal	and	54	788	2	-0.1

The frequency value freq(w_k) can be interpreted as the probability of a random word being precisely w_k. As a consequence of this, negative PMI values imply that the word-pair is occurring less often than would be predicted by chance (through the multiplication freq(w_1) $\cdot$ freq(w_2), which estimates the expected probability of two independent words co-occurring). However, such negative scores are often deemed unreliable (JURAFSKY; MARTIN, 2008). A common solution is to use Positive PMI (PPMI), which eliminates the negative scores:

$$\text{PPMI}(w_1, w_2) = \max\{0,\ \text{PMI}(w_1, w_2)\}.$$

Association measures are often used for the estimation of the level of *conventionalization* of expressions. Conventionalization refers to the degree to which the specific choice of words and word order can be seen as fixed in the language, when referring to a particular concept. For example, while one may talk about the *"forecast of the weather"*, there is a distinctive markedness to this choice of words, and native speakers are more likely to refer to this concept as *"weather forecast"* instead. Constructions such as the latter are thus said to be more conventionalized. They have also be described as *collocations*, i.e. statistically idiosyncratic word combinations (FARAHMAND; SMITH; NIVRE, 2015; BALDWIN; KIM, 2010).

Other association scores include the Dice coefficient, Student's t-test, χ^2 tests, the log-likelihood ratio test (DUNNING, 1993) and NPMI (BOUMA, 2009). A thorough description of commonly used association measures, along with their evaluation in the context of collocation extraction, can be found in Pecina (2010).

2.2.4 Inter-rater agreement measures

One of the goals of NLP is to be able to replicate human-like processing of natural language. To this end, datasets containing human annotation of some linguistic phenomenon are often useful in the tuning and evaluation of NLP systems. The construction of these datasets may be performed by experts (with appropriate linguistic background), or may be framed in such a way that the task may be performed by laypeople. In both cases, there must be a set of guidelines, which instruct the annotator on what must be annotated and provide a solution to common corner cases.

Human language can be often ambiguous (especially when it comes to semantics), and this ambiguity may cause disagreement of annotation even among highly well-trained experts. One indicator of the quality and objectivity of both the annotation guidelines and the resulting dataset could be the fraction of annotations in which all annotators agree:

$$p_{\text{agree}} = \frac{\text{cases of agreement}}{\text{all annotations}}$$

In the case of two annotators, another solution would be to calculate the linear correlation and categorical evaluation measures (see Section 2.2.5).

However, these measures do not take into account the probability of *chance agreement* among annotators. For example, consider a POS-tag annotation task with an inventory of 17 possible POS tags. Even if two annotators decide to assign POS tags randomly to every word in the corpus, they will still be in agreement in around $\mathrm{p}_{\text{chance}} = \frac{1}{17} = 0.058 = 5.8\%$ of the cases. An alternative measure that does take this probability of chance agreement into account is Cohen's *kappa* coefficient (ARTSTEIN; POESIO, 2008):

$$\kappa = \frac{p_{\text{agree}} - p_{\text{chance}}}{1 - p_{\text{chance}}}$$

Kappa scores are usually lower-bounded by 0.0 (which indicates pure chance agreement), and are always upper-bounded by 1.0 (which indicates perfect agreement). The kappa score can be generalized for more than two annotators by using an appropriate estimate of $\mathrm{p}_{\mathrm{chance}}$.

One downside of the kappa coefficient is the fact that it is restricted to categorical data, with the same weight applied to all disagreements. An alternative measure that calculates multi-annotator agreement while taking into account the distance between ordinal ratings is Krippendorff's *alpha* score (ARTSTEIN; POESIO, 2008):

$$\alpha = 1 - \frac{D_{\text{disagreement}}}{D_{\text{chance disagreement}}},$$

in which the measure of disagreement D is an average of the variance in the distance between ordinal ratings for all paired-up annotators.

2.2.5 Evaluation measures for continuous data

When dealing with large amounts of data, it is useful to have a way of summarizing results. In particular, one often needs to measure how similar are two sets of numerical scores. For example, consider Table 2.4, in which each word pair has been manually annotated by a human rater based on the perceived level of association between the words. While the human and PMI scores of each word pair can be individually compared with ease, it is not immediately obvious whether the PMI values are a good estimation of the human ratings. The similarity of these two variables can be calculated through *correlation* measures.

Table 2.4: Human-rated association scores vs PMI.

word_1	word_2	Human score	PMI score
federal	government	6.0	4.4
federal	courts	8.8	5.3
federal	district	3.0	4.9
federal	level	0.7	4.4
federal	taxes	5.0	4.5
federal	debt	2.2	4.3
federal	law	1.5	3.6
federal	outlays	9.5	5.8
federal	republic	8.0	4.9
federal	and	0.0	-0.1

One way of calculating the correlation between two paired datasets X and Y (each with N elements) is through the covariance:

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

The covariance between human and PMI scores in Table 2.4 is 3.86. Covariance values close to zero serve as an evidence that the datasets are not correlated: a positive deviation from the mean in the first dataset $(X_i - \bar{X})$ is equally likely to be paired up with a positive and a negative deviation in the second dataset $(Y_i - \bar{Y})$. Covariance values distant from zero (whether negative or positive) indicate a tendency of both datasets towards higher deviations for the same data points. The covariance score is upper-bounded by the max of the variances $\sigma^2(X)$ and $\sigma^2(Y)$, and is similarly lower-bounded by the min of $-\sigma^2(X)$ and $-\sigma^2(Y)$. The variance itself is calculated as a special case of the covariance:

$$\sigma^2(X) = \mathrm{cov}(X,X).$$

For the values in Table 2.4, the variance in human scores was 12.16, while the variance in PMI was 2.64. Variance scores are either 0.0 or positive, with higher values representing data that differs more heavily from the average.

Covariance scores are generally deemed hard to interpret, as they depend on how the datasets themselves deviate from the mean. One way of dealing with this difficulty is to normalize the covariance based on the variance of both datasets. This is what is done in *Pearson*'s r coefficient:

$$\mathrm{r}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\sigma^2(X)} \cdot \sqrt{\sigma^2(Y)}},$$

This normalization permits a more straightforward interpretation of the correlation, as Pearson scores range from -1.0 (perfect negative correlation) to 1.0 (perfect positive correlation). A neutral score of 0.0 represents no correlation, and values in between represent partial levels of correlation. The Pearson coefficient between human and PMI scores in Table 2.4 is 0.68, indicating that the values are highly correlated.

Pearson scores measure the linear correlation between paired-up values from two datasets. Sometimes, it may be considered more appropriate to interpret these values as an indicator of the *ranking* of two data items instead. In this case, the correlation between the values themselves may not be as interesting as the ability of both sets of scores to rank the items in the same order. Thus, an alternative measure of correlation is *Spearman*'s ρ rank coefficient, which represents the linear correlation between the *ranks* from two datasets. Spearman's ρ can be defined in terms of Pearson's r:

$$\rho = \text{r}(\text{rank}(X), \text{rank}(Y)),$$

where the rank operation maps increasing values in the dataset to consecutive integer ranks. For example, for the human scores in Table 2.4, rank(X) would map $[0.0,\ 0.7,\ 1.5\ \dots\ 9.5] \mapsto [\#1,\ \#2,\ \#3\ \dots\ \#10]$. Similarly, rank$(Y)$ would map PMI scores as $[-0.1,\ 3.6,\ 4.6\ \dots\ 5.8] \mapsto [\#1,\ \#2,\ \#3\ \dots\ \#10]$. As in the case of Pearson, Spearman scores range from -1.0 (perfect negative rank correlation) to 1.0 (perfect positive rank correlation).

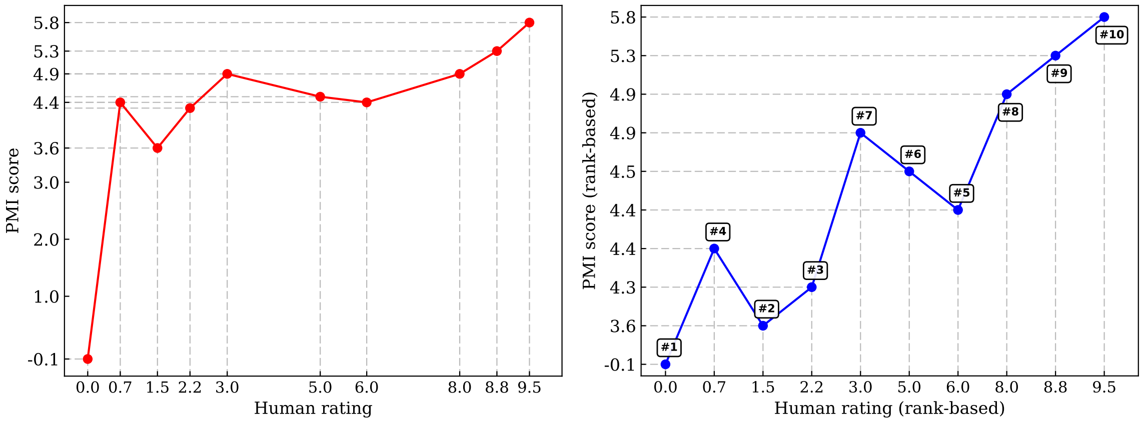
Figure 2.4 presents a graphical visualization of these two approaches:

- On the left, we see the PMI values as a function of human ratings themselves. We can visually confirm a tendency of PMI scores to be higher for higher values of human rating. Pearson's r calculates the level of this correlation between the values themselves (visually: how much the points in this graph resemble a straight line). The Pearson coefficient in this example is $r = 0.68$.

- On the right, we see the PMI score ranks as a function of the rank of the human ratings. Values in the x-axis were ordered based on successive ranks (e.g. $0.0 \mapsto \#1$, $0.7 \mapsto \#2$, and so on). Successive ranks in the y-axis are indicated in little squares (e.g. $5.3 \mapsto \#9$).[3] In both axes, we see that consecutive points are separated by an equal amount of space, regardless of the actual value, as each point is plotted based on the successive integer ranks. Spearman's ρ calculates the linear correlation between these ranks (visually: how much the points in this graph resemble a straight line). The Spearman rank coefficient in this example is $\rho = 0.87$.

[3] Tie-breaking is not performed in this example for 4.4 and 4.9. In the remainder of the thesis, tie-breaking is performed by assigning the average of all tied ranks (e.g. $4.9 \mapsto \#7.5$).

Figure 2.4: Visual representation of Pearson's r (left) and Spearman's ρ (right) between human-rated association scores and PMI.



2.2.6 Evaluation measures for categorical data

Correlation measures are able to adequately compare two sets of items with quantitative (e.g. real-valued) scores. However, they are unfit for comparisons involving qualitative categorical scores. The need for comparing categorical scores arises, for example, in the task of POS tagging (see Section 2.1). Consider this sample excerpt from the English toy corpus: *"The first inhabitants of North America migrated from Siberia by way of the Bering land bridge"*. Table 2.5 presents an example of automatic system prediction of POS tags along with the corresponding human-annotated POS tags.

Table 2.5: Human-annotated and system-predicted POS tags.

Word	Human POS	System POS
The	DET	DET
first	ADJ	ADJ
inhabitants	NOUN	NOUN
of	ADP	ADP
North	PROPN	NOUN
America	PROPN	PROPN
migrated	VERB	VERB
from	ADP	ADP
Siberia	PROPN	PROPN
by	ADP	ADP
way	NOUN	ADV
of	ADP	ADP
the	DET	DET
Bering	PROPN	PROPN
land	NOUN	VERB
bridge	NOUN	NOUN

Consider the human and system annotations of the NOUN tag. Humans annotated 4 occurrences (*"inhabitants"*, *"way"*, *"land"*, *"bridge"*), while the system predicted 3 occurrences (*"inhabitants"*, *"North"*, *"bridge"*). Out of these, 2 occurrences were *true positives*, i.e. system predictions that matched human annotations (*"inhabitants"* and *"bridge"*). One way of summarizing the predictive quality of this system for NOUN tags is by means of *precision* and *recall*:

$$\text{precision} = \frac{\text{true positives}}{\text{total system predictions}}, \qquad \text{recall} = \frac{\text{true positives}}{\text{total human annotations}}.$$

The precision measures how many of the predictions were correct, while the recall measures how many of the tags were correctly predicted. These two measures are often further combined into a single score using the harmonic mean, yielding the F_1 *score*:

$$\text{F}_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Table 2.6 present these statistics for all POS tags, comparing human and system annotations. In particular, it can be seen how the scores for NOUN were obtained: the precision of 2/3 represents the fraction of correct system predictions, while the recall of 2/4 represents the fraction of human-annotated NOUN tags that were correctly found by the system.[4]

Table 2.6: Binary evaluation measures for POS-tag prediction.

POS tag	Total human	Total system	True Positives	Precision (TP/system)	Recall (TP/human)	F_1
ADJ	1	1	1	1/1 = 1.00	1/1 = 1.00	1.00
ADP	4	4	4	4/4 = 1.00	4/4 = 1.00	1.00
ADV	0	1	0	0/1 = 0.00	0/0 = 1.00	0.00
DET	2	2	2	2/2 = 1.00	2/2 = 1.00	1.00
NOUN	**4**	**3**	**2**	**2/3 = 0.67**	**2/4 = 0.50**	**0.57**
PROPN	4	3	3	3/3 = 1.00	3/4 = 0.75	0.86
VERB	1	2	1	1/2 = 0.50	1/1 = 1.00	0.67

In this thesis, we will also consider a variant of the F_1 score known as the Best F_1 score (BF_1). This measure is useful when comparing continuous system predictions to categorical judgments in a dataset. It is obtained by calculating the F_1 score for the

[4] To handle edge cases, we define 0/0 as 1.0.

top k entries classified as positive (i.e. the k highest-scoring system predictions), for all possible values of k. In other words, for a dataset X ordered by predicted scores:

$$\mathrm{BF}_1(X) = \max_k \mathrm{F}_1(X_{1\ldots k}).$$

This measure will be used in Chapter 5 when comparing the continuous predictions from our system with the categorical judgments from the *Farahmand* dataset.

2.3 Multiword expressions

Sentences in human languages are more than an unordered collection of words. In order to convey a specific meaning, the words in a sentence must be structured, grouped so as to create phrases, which themselves can be recursively grouped until the whole sentence has been internally connected. The semantics of the whole sentence can then be derived from the semantics of its individual components and from the way in which they relate to each other.

In every known human language, there is a class of expressions that does not necessarily behave in this compositional manner, known as multiword expressions (MWEs). Examples of MWEs would be the English verb–particle construction *give up*, the French NC *carte bleue* (‘bank card’, lit. *blue card*), and the Portuguese idiom *bater as botas* (‘kick the bucket, die’, lit. *hit the boots*). The meaning of an MWE is not always formed by the application of regular rules from the grammar. Rather, each MWE constitutes a semantic unit that spans over multiple lexemes (SAG et al., 2002), and which often needs to be analyzed as an indivisible entity. MWEs may present lexical, morphological, syntactic, semantic, pragmatic and statistical idiosyncrasies (BALDWIN; KIM, 2010), as exemplified below:

- Lexical idiosyncrasy: MWEs may contain words that do not otherwise exist in the language, and thus cannot appear by themselves[5]. Examples include EN *of yore*; PT *no entanto* ‘however’ (lit. *in-the entanto*); and FR *au fur et à mesure* ‘in keeping with’ (lit. *to-the fur and to-the measure*).

- Morphological idiosyncrasy: MWEs may contain words that do not respect the normal rules of inflection, as in EN *spill the beans*, where the word *bean* must always

[5] Often called *cranberry words*.

be pluralized; `PT` *azul-marinho* 'navy blue' (lit. *marine blue*), where the *marine* adjective does not inflect; and `FR` *grand-mère* 'grandmother' (lit. *big mother*), where the adjective *big* does not inflect.

- Syntactic idiosyncrasy: MWEs may not conform to the regular syntactic rules of the language. Examples include `EN` *by and large*, which behaves as an adverb; `PT` *faz de conta* 'make-believe' (lit. *makes of account*), which behaves as a noun; and `FR` *bon marché* 'cheap' (lit. *good market*), which behaves as an adjective. In the case of *by and large*, the MWE comprises a sequence of elements (preposition+conjunction+adjective) that would otherwise not even be allowed by the grammar of the English language.

- Semantic idiosyncrasy: the meaning of the whole expression may not come from the combination of the meaning of its parts. This can be seen in the idiomatic meaning of an MWE such as `EN` *to kick the bucket*, with the equivalent `PT` *bater as botas* (lit. *to hit the boots*) and `FR` *casser sa pipe* (lit. *to break one's pipe*).

- Pragmatic idiosyncrasy: MWEs may only occur in a particular extra-linguistic context, as in the case of `EN` *all aboard*, `PT` *bom dia* 'good morning' (lit. *good day*), `FR` *au revoir* 'goodbye' (lit. *to see-again*).

- Statistical idiosyncrasy: the expression may have been conventionalized in a specific form, even though a substitution by a synonym could happen in principle. This is the case of `EN` *many thanks* (compare with **a lot of thanks*), `PT` *café preto* ('black coffee', compare with **café negro*), `FR` *noir et blanc* ('black and white', compare with **blanc et noir*).

Note that the definition of MWEs in the literature may exclude purely statistically idiosyncratic expressions (also known as collocations). In this thesis, an MWE is understood as a more general term that encompasses all types of idiosyncratic units that cross word boundaries (SAG et al., 2002; BALDWIN; KIM, 2010), as the main interest of our research is the variation of these expressions in a continuum of idiomaticity.

The idiosyncratic behavior of MWEs might lead one to think that such expressions constitute rare exceptions in the language. However, estimations on the number of MWEs in a given speaker's lexicon may reach at least the same order of magnitude as the number of single words (JACKENDOFF, 1997). In the context of specialized domains, this number is expected to be even higher, as they naturally favor the apparition of multiword technical terms (SAG et al., 2002).

The abundance of MWEs, coupled with the fact that their occurrences are not obvious and their meaning is not predictable, contributes to making this an essential topic of research for NLP (SAVARY et al., 2015). A correct analysis of MWEs is essential to any computer application that has the goal of somehow interpreting written texts, such as machine translation and text simplification.

2.3.1 Nominal compounds

A category of MWE that is of particular interest in this thesis is the *nominal compound.* A nominal compound (NC) is defined as a syntactically well-formed and conventionalized noun phrase containing two or more content words, the head of which is a noun.[6] Such compounds can express different levels of semantic idiosyncrasy: their interpretation may come directly from the meaning of its components (e.g. *climate change*), or be highly idiomatic (e.g. *cloud nine*), with partially idiomatic cases in-between (e.g. *spelling bee, middle school*) (NAKOV, 2013).

The syntactic realization of NCs varies across languages. In English, they are often expressed as a sequence of nouns, usually N_1 N_2 (with the head noun N_2 modified by N_1). This is the most frequent annotated POS-tag pattern in the MWE-annotated English corpus DiMSUM (SCHNEIDER et al., 2016a). In French and Portuguese, NCs often assume the form of ADJ N or N ADJ, where ADJ is an adjective that modifies the head noun N. Examples of these constructions include the French ADJ N compound *petite annonce* (‘classified ad’, lit. *small announcement*) and the Portuguese N ADJ compound *buraco negro* (‘black hole’, lit. *hole black*). Additionally, NCs in the three languages may also include prepositions that provide a hint on the role of the modifier noun with respect to the head (e.g. the N PREP N compound *rule of thumb*). Most prepositions are highly polysemous, and it is not clear how they should be represented in the context of distributional semantic models. In the remainder of this thesis, we will focus on 2-word NCs that follow the form N_1 N_2 (in English), N ADJ (in Portuguese and French) and ADJ N (in the three languages).

[6] The terms *noun compound* and *compound noun* are usually reserved to noun–noun compounds. These are typical of Germanic languages, but not as common in Romance languages.

2.3.2 Type discovery

The goal of MWE discovery is to automatically find new MWEs in corpora, collecting these MWEs in a lexicon for future use (CONSTANT et al., 2017). The earliest works toward MWE discovery and lexicon building involved attempts at extracting collocations. These are expressions that present some level of statistical idiosyncrasy, regardless of other levels of idiosyncrasy. By taking into account some statistical measures, pairs of collocated words could be extracted from corpora with high accuracy (SMADJA, 1993).

Subsequent works have applied linguistic knowledge in order to target specific categories of MWE, such as noun compounds (e.g. *milk tooth*), verb–particle constructions (e.g. *look up*) and light-verb constructions (e.g. *take a shower*) (JUSTESON; KATZ, 1995; FRANTZI; ANANIADOU; MIMA, 2000; STEVENSON; FAZLY; NORTH, 2004; EVERT; KRENN, 2005). More recent works also focus on more general type-independent approaches to MWE discovery (SERETAN, 2011; AGRAWAL; AGGARWAL et al., 2013; TSVETKOV; WINTNER, 2014).

MWE discovery techniques usually rely on word frequency and association measures, which are inexpensive language-independent methods of detecting the conventionalization of MWEs. The construction of such MWE lists can be greatly simplified by using an MWE extractor, such as the mwetoolkit (RAMISCH, 2015). This toolkit includes multiple tools for MWE discovery and manipulation, including an extraction algorithm that builds upon the notion of regular-expression patterns to match token properties. For example, given a noun compound pattern such as `(Noun Noun`$^{+}$`)` and a POS-tagged corpus, the extraction yields all occurrences of at least two subsequent nouns in the text. The mwetoolkit can then be used to calculate association measures and to filter out MWE candidates.

A major downside of such commonly used extraction techniques is that they do not take semantic information into account when filtering MWE candidates. One of the contributions of this thesis is an extension of the mwetoolkit that calculates whether an extracted expression can be treated as a combination of its parts or whether it should be treated as a standalone semantic unit (see Chapter 4 for more details).

2.3.3 Token identification

While MWE type discovery focuses on building lexicons of new MWE types, token identification has the goal of automatically annotating the tokens that correspond to an MWE occurrence in running text (CONSTANT et al., 2017). The accurate identification of MWE tokens is a fundamental task in the pipeline of many NLP applications. For example, MT systems need to know when a group of words must be translated as a unit, and parsers need to recognize the cases where a seemingly unrelated group of words should be joined as a single lexeme or constituent.

Identification of MWE tokens in corpora usually requires an MWE lexicon as input, and can be seen as a tagging process, akin to POS tagging. The goal is to look for occurrences of MWEs in a corpus and to output an augmented version of the corpus that explicitly indicates where each expression occurs. This indication can range from simply joining the MWE components as a single word (using a special "MWE separator" character) to more complex metadata representations, such as indicating each MWE by the index of its component tokens (SCHNEIDER et al., 2016a; SAVARY et al., 2017c).

MWE identification tools such as `jMWE` (KULKARNI; FINLAYSON, 2011) are often used to annotate sentences based on preexisting lexicons. Finite-state transducers can also be used to take into account the internal morphology of component words and perform efficient tokenization based on MWE dictionaries (SAVARY, 2009). The problem of MWE identification has also been modeled using supervised machine learning, where the data is encoded in a begin-inside-outside scheme, from which one can learn sequence taggers such as CRFs (CONSTANT; SIGOGNE, 2011; SCHNEIDER et al., 2014; SCHOLIVET; RAMISCH; CORDEIRO, 2017).

These solutions have some shortcomings. One of the problems is that MWEs do not always appear contiguously in the text. For example, they may contain internally inserted modifiers or arguments, as in the expressions *to give [something] up* and *to take [a very long] shower*. In such cases, contiguous identifiers will fail to detect these MWEs. One way of dealing with this problem would be the use of parsing-based approaches (CONSTANT; NIVRE, 2016), but these require the existence of annotated treebanks for training, which are not available for most languages.

Another shortcoming of using separate tools for type discovery and token identification is that one misses the opportunity of sharing information. This has negative results both in terms of CPU time and in the inability to guarantee that all MWE candidates extracted by one tool have been projected back onto the source corpus by the other tool. One of the contributions of this thesis is an extension of the mwetoolkit that identifies MWE occurrences in text. The implementation addresses the latter by allowing the integration of type and token identification in the same pipeline, and the former by enabling non-contiguous matches (e.g. *eat [food] up*, as in *eat that wonderful chocolate cake up*) and optional and variable elements in MWEs (e.g. *throw [person] to the lions/wolves*). See Section 6.1.1 for more details.

MWE token identification is also closely related to the problem of disambiguation in lexical semantics (SCHNEIDER et al., 2016a). For example, a machine translation system would need to decide whether an expression such as *a piece of cake* should be interpreted as a single unit (e.g. *the test was a piece of cake*) or as a composition of its parts (e.g. *he just ate a piece of cake*), so that it may translate it into an equivalent meaning in the target language. Semantically-aware MWE token identification is a current topic of research (CONSTANT et al., 2017).

In this thesis, we will focus on MWE type-level compositionality prediction. The work on token-level identification is seen as a means to an end: collecting candidate MWE tokens so as to be able to calculate type-level semantics.

2.4 Word semantics

In the past years, NLP work on lexical semantics has been shifting from a focus on symbolic representations (often in the form of handcrafted resources) to more numerical techniques that can automatically extract word representations from large bodies of text. We present below an overview of both approaches.

2.4.1 Symbolic representations

One of the earliest formal representation of semantics was the one promoted by Montague during the 1960s and 1970s. This approach relied on the use of formal logic to treat natural language semantics in the same rigorous way as Chomskian grammar would deal with syntax. The resulting Montague Grammar is still today an active area of research in the field of Linguistics (PARTEE, 2014). Words are defined as sets of elements: the noun *house* stands for the set of everything that could be considered a house, while the adjective *yellow* would be seen as the set of all things yellow. A yellow house would then be anything in this intersection (BARONI; ZAMPARELLI, 2010). This representation is somewhat limited, as can be seen in the expressions *red watermelon* (inside) versus *green watermelon* (outside).

The meaning of individual words may also be approximated through a set of semantic labels. This is the approach used in SemCor, one of the earliest sense-annotated corpus for the English language. SemCor was constructed by tagging the content words in the Brown corpus based on a set of semantic *sense* labels[7] (LANDES; LEACOCK; TENGI, 1998). More recently, the STREUSLE corpus of web reviews has been annotated in terms of on noun, verb and preposition *supersenses*. Supersense tags (such as *person*, *location*, and *event*) are a more coarse-grained way of representing the semantics that allows some level of comparability between different words (CIARAMITA; JOHNSON, 2003; SCHNEIDER et al., 2016b).

Another way of representing semantics is through a graph, where each node represents a different meaning[8], and edges represent some kind of relation between those meanings. For example, a graph of hypernymy relations (i.e. word generalizations) could contain nodes such as DOG, MAMMAL and ANIMAL, with edges connecting DOG→MAMMAL as well as MAMMAL→ANIMAL. Lexical databases have been constructed for some languages, with the most prominent example being the hand-crafted WordNet (FELLBAUM, 1998), which includes relations such as hypernymy, meronymy (i.e. part–whole), and antonymy (i.e. words with opposite polarity) for the English language. Such databases require dedicated human work for each word in the language, requiring periodic updates to keep in line with the lexical evolution of the language over time.

Automatically-created databases are a current topic of research, and may be a more feasible alternative to manual annotation (especially for languages that dispose of fewer financial resources), at the expense of accuracy (NAVIGLI; PONZETTO, 2012). Such resources may be created from parallel texts, determining the semantics of each word based on its possible translations (GANITKEVITCH; DURME; CALLISON-BURCH, 2013). However, this approach is inherently reliant on the alignment of large amounts of parallel data, which is a scarce resource for most languages.

[7] The sense labels correspond to WordNet synsets.

[8] Words that may refer to the same meaning are grouped in what is known as a synset, so e.g. the words *dog* and *hound* could be part of a synset called DOG.

2.4.2 Numerical representations

While symbolic approaches often require an impractical amount of high-quality data for a given language or domain, numerical solutions such as distributional semantic models (DSMs) tend to be more malleable in their requirements. DSMs use context information (such as the co-occurrence matrix, described in Section 2.2) to represent the meaning of lexical units in the form of numerical vectors. The central idea is that the meaning of a word is naturally learned by human beings based on the contexts where it appears — or, as popularized by Firth (1957), "you shall know a word by the company it keeps". Distributional models can be built from any large-enough corpus, and do not particularly require any level of data preprocessing.

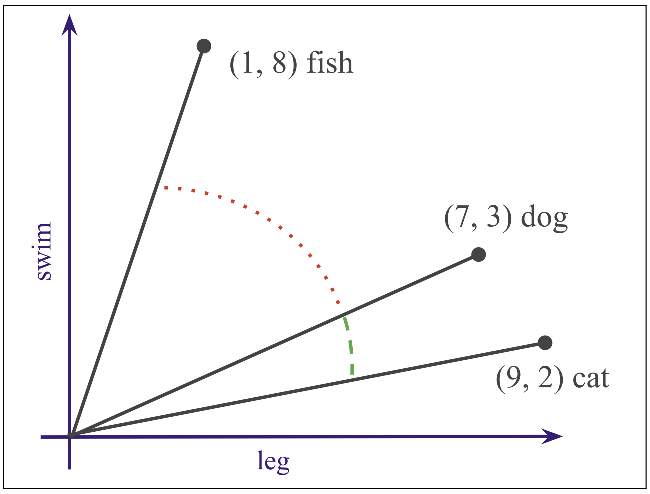
Consider the target–context matrix in Table 2.7. Each row corresponds to a target word (*fish, dog, cat*) followed by its context vector. Each dimension in the context vector represents a measure of co-occurrence between the target (e.g. *dog*) and a context that was seen close to this target in a corpus (e.g. *swim*). For example, the target *fish* has a co-occurrence of 1 in the dimension labeled *leg* and a co-occurrence of 8 in the dimension labeled *swim*.

Table 2.7: Matrix with targets (rows) × contexts (columns)

vocabulary		leg	swim
w_1	fish	1	8
w_2	dog	7	3
w_3	cat	9	2

The essence of DSMs consists in treating each of these rows as a numerical D-dimensional representation of its target word (with $D = 2$ in this toy example). The assumption is that the meaning of a word can be derived from this vector. For example, the word *cat* would be represented as the vector $\mathbf{v}(w_3) = [9, 2]$. Figure 2.5 presents a visual representation of these vectors. The results exemplify something that is often seen in DSMs: because both words *dog* and *cat* have a numerical representation that is strongly associated with *leg* and weakly associated with *swim*, their vectors are closer to each other, from which we can infer that *dog* is overall more similar to *cat* than it is to *fish*.

Figure 2.5: DSM vectors of the target words *"fish"*, *"dog"* and *"cat"*, considering the contexts *"lex"* and *"swim"*.



The similarity between the vector representation of two words can be mathematically measured through the use of a function such as the cosine, which can be defined for two n-dimensional vectors as their normalized dot product:

$$\cos(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| \cdot |\mathbf{w}|},$$

where $|\mathbf{v}|$ represents the norm of the vector $\mathbf{v}$. This definition can be further expanded as:

$$\cos(\mathbf{v}, \mathbf{w}) = \frac{\sum_{i=1}^{n} \mathbf{v}_i \cdot \mathbf{w}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{v}_i^2} \cdot \sqrt{\sum_{i=1}^{n} \mathbf{w}_i^2}},$$

The cosine has a value closer to 1 for vectors that are closer to each other and closer to 0 for vectors that are perpendicular in most dimensions.[9] In the case of DSMs, this perpendicularity indicates that those vectors do not share many distributional features, and might suggest that they are not semantically related. Other measures could be used for measuring word vector similarity, in particular the euclidean distance. One of the advantages of the cosine is the fact that it is invariant with regards to vector length (which is not usually considered meaningful in a DSM representation).

Formally, distributional semantic models attempt to encode the representation of each word in a vocabulary V as a vector of real numbers $\mathbb{R}^{|V|}$. Traditionally, this representation is built by weighting the level of co-occurrence of all pairs of words. This weighting can either be done by counting the number of co-occurrences (BARONI; DINU; KRUSZEWSKI, 2014) or by calculating some measure of the association between target

[9] The cosine can also be negative, e.g. the vectors point to opposite directions in most dimensions.

and context, such as the PPMI (LEVY; GOLDBERG; DAGAN, 2015). Words are usually deemed to co-occur if they appear in the text under a small fixed-size window of words. Alternatively, some works define co-occurrence based on whether the two words share a syntactic dependency relation (LIN, 1998; PADÓ; LAPATA, 2007). In all cases, the result is a target–contexts mapping $M = V \times \mathbb{R}^{|V|}$ where many of the context weights are zero (as most word pairs in $V \times V$ will almost never co-occur), and it is thus often implemented as a sparse matrix. A threshold on the number of word pair co-occurrences is often applied to discard low weights.

An alternative to working on a sparse representation is the use of *word embeddings*, in which the vectors are transformed so as to have a significantly smaller number of dimensions.[10] Two solutions are commonly employed in the literature: global contexts and dimensionality reduction. In the case of global contexts, only the top k most frequent words are considered as contexts, producing a $|V| \times k$ matrix (SALEHI; COOK; BALDWIN, 2014; PADRÓ et al., 2014). A second alternative is the use of dimensionality reduction techniques. Assuming that all vectors represent data points on a space whose mean is $\mu = 0$, a technique known as single value decomposition (SVD) may be used to transform the matrix rows $M_1..M_n$ in such a way that the largest variance now occurs on $M_{i,1}$ (i.e. maximizing the variance $\sigma_i^2(M_{i,1})$), the second largest variance on $M_{i,2}$, and so on. Only the first k components of each vector are then kept; the rest is discarded. The rationale is that higher variances represent actual structure, while lower variances represent noise in the data.

SVD achieves its results through a matrix factorization technique (SHLENS, 2014). Formally, it decomposes the matrix $M_{m \times n}$ into three other matrices:

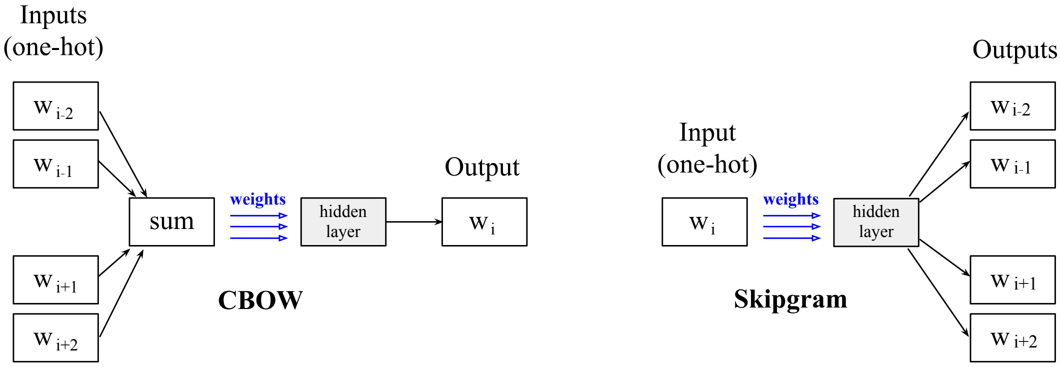
$$M = U_{m \times m} \cdot \Sigma_{m \times n} \cdot V_{n \times n},$$

where U and V specify rotations and Σ is a diagonal matrix which specifies a scaling operation. The product $U\Sigma$ has the aforementioned property in which lower indexes correspond to higher variances. It can be truncated into an $m \times k$ matrix for smaller values of k, effectively obtaining a version of M that has a reduced number of dimensions. Similar methods have also been published focusing on factorizing the logarithm of the co-occurrence matrix (PENNINGTON; SOCHER; MANNING, 2014) and factorizing a matrix of PPMI values (SALLE; VILLAVICENCIO; IDIART, 2016).

[10] Each of these smaller vectors is then called a *word embedding*.

Another word embedding technique is the one adopted by word2vec, which trains a neural network in a task that involves predicting target–context relationships. Two approaches have been proposed: training a network to predict a target word given a window of surrounding contexts, known as the Continuous Bag-of-Words (CBOW) model; and training a network to predict likely contexts for a given target, known as the skipgram model (MIKOLOV et al., 2013). Figure 2.6 presents the general architecture of both approaches, for a window of 2+2 words around each target. In both cases, input words are represented as a one-hot vector in $\{0, 1\}^{|V|}$. Each entry in the input vector is connected to a hidden layer of d neurons, which learns to generate an $\mathbb{R}^{|V|}$ output that predicts[11] the one-hot vector of w_i (for CBOW) or its contexts (for skipgram). After training, the input weights from the hidden layer are then taken as the set of d-dimensional word embeddings. In both word2vec approaches, the network automatically adapts itself to encode useful semantic information as a side effect of trying to solve its prediction task.

Figure 2.6: Architecture of word2vec (CBOW and skipgram).



When building a distributional semantic model, there are some common parameters that must be considered, but that are orthogonal to the choice of DSM technique. One of the main considerations is the type of information that will be provided for each word in the corpus: using the word's surface form may generate a sparser model (due to inflections), while using the lemma may merge unrelated occurrences, ignoring relevant morphological distinctions. The use of POS tags may also contribute to disambiguate polysemous words (e.g. compare *the circle* with *we circle*), but it risks the introduction of tagging errors. Another consideration is the removal of stopwords: common function words (such as *the*, *of*, and *for*) do not contribute much to the semantics of the text, and their removal may allow DSM techniques to better capture relevant co-occurrence

[11] During training, the output layer is followed by a softmax layer, which generates the probability for each element in V (and which is compared to the one-hot in the output).

patterns in the text. In Chapter 5, we present some contributions to the understanding of how these parameters affect different DSMs.

2.5 MWE semantics

Human language is often modeled as though it were compositional, i.e. assuming that the meaning of the whole can be built from the meaning of its parts. However, MWEs may display a wide range of idiomaticity, ranging from compositional cases (e.g. *tennis championship*) to idiomatic non-compositional cases (e.g. *gravy train*). For the latter, the meaning of the expression cannot be understood directly from the meaning of its parts (e.g. a *gravy train* refers to a low-effort lucrative endeavor). Even when there is a level of compositionality in the expression, the contribution of each word may vary considerably, independently from its status as a syntactic head or modifier, as *tears* (head) in *crocodile tears* versus *cash* (modifier) in *cash cow.* In this section, we present the state-of-the-art approaches towards the representation of compositional and idiomatic MWE semantics, using symbolic as well as numerical representations.

2.5.1 Symbolic representation

In the case of compositional MWEs, Lauer (1995) argues that prepositions (such as *from*, *for*, *in*) can be used to classify the role of each component (e.g. *olive oil* is *oil from olives*). These prepositions are explicitly part of some NCs in Romance languages (e.g. `FR` *huile d'olive* and `PT` *azeite de oliva*). More generally, Girju et al. (2005) present and compare several inventories of semantic relations between nouns inside NCs, ranging from fine-grained to coarse senses. These relations include syntactic and semantic classes such as *subject*, *instrument* and *location.*

Free paraphrases have also been used to model compositional MWE semantics based on the meaning of the components. Nakov (2008) suggests using unsupervised generation of paraphrases combined with web search engines to classify NCs. This was further extended in SemEval 2013, in a task where free paraphrases were ranked according to their relevance for explicitly describing the underlying semantic relations in the compounds (HENDRICKX et al., 2013). For instance, with respect to the expression *flu virus*, the paraphrases at the top of the rank contained verbs such as *cause, spread* and *create* (i.e. *virus that causes/spreads/creates the flu*).

Regarding the representation of idiomatic MWEs, lexical resources such as the aforementioned WordNet (FELLBAUM, 1998) will often include them alongside single words. However, the drawbacks from single-word lexical resources still hold true for MWEs. In particular, for automatically constructed lexicons, the relatively lower frequency and high syntactic variability of MWEs may exacerbate the discrepancy between single-word and MWE coverage even further.

Regarding the representation of idiomatic MWEs, some works extend the *supersense* approach used for single words to identify every MWE as an indivisible unit pertaining to a semantic class. Recent versions of the SemCor corpus (LANDES; LEACOCK; TENGI, 1998) annotate MWEs as well as single-word units based on a set of supersenses derived from top-level WordNet hypernym senses. A similar approach is adopted by the STREUSLE corpus, which contains supersense labels for nouns, verbs and prepositions, both when acting as single words and as part of an MWE (SCHNEIDER et al., 2014a; SCHNEIDER et al., 2016b). This latter corpus was also extended in the DiMSUM shared task, with the joint goal of token-based MWE identification and sense disambiguation (SCHNEIDER et al., 2016a).

2.5.2 Numerical representation

Numerical approaches to MWE semantic representation usually focus on a continuum of compositionality ranging from very compositional expressions to very idiomatic ones. The meaning of an MWE is then expressed through a numerical compositionality score. A low score indicates a completely idiomatic meaning, while a high compositionality score indicates that the meaning of the MWE comes directly from its parts. For example, using a range from 0 to 1, the idiomatic expression *sitting duck* could be associated with the compositionality score 0.2, while the compositional expression *swimming pool* could be assigned a high score such as 0.9.

Separate scores can also be used to represent the literality associated with each individual word (REDDY; MCCARTHY; MANANDHAR, 2011). For example, the expression *spelling bee* could be 80% literal with regards to *spelling* and 0% literal with regards to *bee*, while *milk tooth* could be 20% literally related to *milk* and 100% related to tooth.

The meaning of an MWE can also be specified in terms of its *conventionalization*

(FARAHMAND; SMITH; NIVRE, 2015). Just like compositionality scores can distinguish MWEs in a continuum of idiomaticity, conventionalization scores can be used to identify the level of perceived statistical idiosyncrasy in an expression. For example, the expression *tap water* could be considered 20% conventionalized, while *spelling bee* could be judged as 100% conventionalized. Note that conventionalization does not imply idiomaticity (e.g. *time machine* is highly conventionalized while being fairly compositional), and both measures could be used in a single dataset to more precisely specify the semantics of an expression.

Annotating the semantics of MWEs is a considerably hard task, and annotators may disagree on the exact compositionality score. Therefore, scores are often average among multiple annotators. One source of divergence that may be found among annotators is that some datasets do not take polysemy into account, as the authors ask annotators to think about the most common sense of an MWE without providing any context. Some of these datasets address this issue by providing example sentences to attenuate this problem.

Numerical scores can sometimes be considered a more flexible alternative to symbolic representations of semantics. For one, they allow the interpretation of compositionality in a continuum, which is in line with the perception that the meaning of some MWEs can be more easily guessed than the meaning of the others. Numerical representations can be more readily applied in other numerical contexts, such as DSM representations of semantics.

Numerical representations also allow for fine-grained distinctions that are not possible in a strictly categorical setting. For example, while one may stipulate the categories *idiomatic* and *compositional*, any further attempts at representing partial levels of compositionality would rely on an (implicit) ranking among the categories, tending towards the numerical representations. The possibility of fine-grained distinctions can also be seen as one of the major downsides of numerical scores, as it may introduce uncertainty into the dataset due the subtle differences with which different people see the MWEs. Symbolic representations may also be preferable in the case of highly polysemous MWEs, as the distinction between the multiple senses may not be feasible with bases on commonly studied dimensions such as compositionality or conventionalization.

2.5.3 Compositionality datasets in the literature

We present below a list of relevant datasets representing MWEs alongside human-rated compositionality scores:

- Baldwin and Villavicencio (2002) collected binary type-level judgments for 3 078 English phrasal verbs. Each entry is classified by two experts as either compositional (e.g. *give back*) or idiomatic (e.g. *pull over*), with 14% of the MWEs in the dataset being judged as idiomatic.

- McCarthy, Keller and Carroll (2003) present a dataset of 116 English verb–particle constructions, annotated with type-level compositionality scores by three native speakers, on a scale ranging from 0 (idiomatic) to 10 (compositional). Five of these were unknown to at least one judge, and were removed from the dataset for their experiments.

- Bannard (2006) collected binary judgments for 160 English verb–particle constructions. In this work, compositionality judgments for each expression were collected from multiple annotators, allowing more fine-grained distinctions in idiomaticity.

- Reddy, McCarthy and Manandhar (2011) collected judgments for a set of 90 English noun–noun and adjective–noun compounds, in terms of three numerical scores: the compositionality of the compound as a whole and the literal contribution of each of its parts individually, using a scale from 0 to 5. Compounds included in the dataset were selected to balance frequency range and degree of compositionality (low, middle and high). The dataset was built through crowdsourcing, and the final scores are the average of 30 judgments per compound. This dataset will be used in our intrinsic evaluation experiments in Chapter 5, where it will be referred to simply as *Reddy*.

- Gurrutxaga and Alegria (2013) had three experts classify 1200 Basque noun–verb expressions according to one of three possibilities: idiomatic, compositional collocation, or free combination.

- Roller, Walde and Scheible (2013) collected judgments for a set of 244 German noun–noun compounds, each compound with an average of around 30 judgments on a compositionality scale from 1 to 7, obtained through crowdsourcing. The resource was later enriched with feature norms (ROLLER; WALDE, 2014).

- Farahmand, Smith and Nivre (2015) had 4 experts annotate 1042 English noun–noun compounds. Each annotator provided binary judgments for every MWE regarding idiomaticity (non-compositionality) and conventionalization (when a particular choice of words has been crystallized as part of the language, even if synonyms would have been understandable). A hard threshold can be applied so that compounds are considered as non-compositional if at least two annotators say so (YAZDANI; FARAHMAND; HENDERSON, 2015), and the total compositionality score is given by the sum of the 4 binary judgments. This dataset will be used in our intrinsic evaluation experiments in Chapter 5, where it will be referred to as .

- Walde et al. (2016) collected judgments for a set of 868 German noun–noun compounds, with human judgments of compositionality ranging on a scale of 1 to 7. The dataset is also annotated for in-corpus frequency, productivity and ambiguity, and a subset of 180 compounds has been selected so as to be balanced with respect to these variables. The different annotations were performed by the paper authors, linguists, and through crowdsourcing. A similar dataset has been collected for verb–particle constructions (BOTT et al., 2016).

Some of these datasets were constructed with binary judgments, while others were constructed with a more malleable representation of idiomaticity, requesting human raters to specify their judgment over a range of possible values. Note, however, that even binary judgments could constitute a numerical dataset. As long as there are enough annotators, the average of the judgments can be taken as a numerical estimate of its perceived idiomaticity.

While the compositionality judgments from the datasets above could be used by themselves as features to semantic tasks such as MWE token identification (described in Section 2.3.3), the size of these datasets may be a limiting factor in the results obtained. On the other hand, these datasets are particularly useful as a way of evaluating the quality of automatic models of compositionality prediction (which themselves may then be used to predict the compositionality of a much larger set of MWEs). In Chapter 3, we present three new datasets with human annotation of compositionality scores. Chapter 5 then evaluates a framework of compositionality prediction on these new datasets, alongside with the *Reddy* and *Farahmand* presented above.

3 COMPOSITIONALITY DATASETS

As we have seen in Chapter 2, MWE compositionality can be modeled in terms of the contribution of meaning of each element toward the meaning of the whole. Some numerical datasets have been proposed in the literature, but they are restricted to English and German MWEs. Moreover, in the former language, only one relatively small dataset contains non-categorical compositionality scores.

In this chapter, we describe the construction of three new datasets of human-annotated compositionality scores for nominal compounds (NCs). These datasets are necessary for our evaluations of compositionality prediction models (reported in Chapter 5). The resources encompass: 180 French nominal compounds (*FR-comp*); 180 Brazilian Portuguese nominal compounds (*PT-comp*); an extension of the English-language *Reddy* dataset with 90 additional compounds ($\textit{EN-comp}_{90}$), for a total of 180 English compounds ($\textit{Reddy}^{++}$).

The work presented in this chapter has also been described in three published papers (RAMISCH et al., 2016; CORDEIRO; RAMISCH; VILLAVICENCIO, 2016a; WILKENS et al., 2017). For French and Portuguese, this is the first human-rated dataset of nominal compound compositionality.

3.1 Data collection

For each of the 3 target languages (English, French and Portuguese), quantitative measures for the level of compositionality of the nominal compounds in the dataset were collected through crowdsourcing. Non-expert participants were asked to judge each compound in the context of three sentences where the compound displayed the same sense, followed by an evaluation of the degree to which the meaning of the compound is related to the meaning of its individual parts. This follows from the assumption that a fully compositional expression will have an interpretation whose meaning comes from both words (e.g. *lime tree*, which is effectively a *tree* of *limes*), while a fully idiomatic compound will have a meaning that is unrelated to its components (e.g. *nut case*, which refers to an eccentric person and is not related to *nuts* or *cases*). This work follows the protocol from Reddy, McCarthy and Manandhar (2011), where the compositionality is explained in terms of the literality of the individual parts. This type of indirect annotation of compositionality is less specialized, and does not require expert linguistic knowledge, while still providing reliable data, as will be shown later.

For each language, data collection involved the following four steps: compound selection; sentence selection; questionnaire design; and questionnaire application.

3.1.1 Compound selection

The initial set of idiomatic and partially compositional candidates was constructed by introspection, independently for each language. This list of compounds was complemented by selecting entries from lists of frequent *adjective+noun* and *noun+noun* pairs. These were automatically extracted through POS-sequence queries using the mwetoolkit (RAMISCH, 2015). The source corpora were ukWaC (BARONI et al., 2009), frWaC and brWaC (BOOS; PRESTES; VILLAVICENCIO, 2014), each containing between 1.5 and 2.5 billion tokens.

We avoided selecting compounds in which the head was not necessarily a noun (e.g. FR *aller simple* ‘one-way ticket’ (lit. *going simple*), as *aller* doubles as the noun *going* and the infinitive of the verb *to go*). We also avoided selecting compounds whose literal sense was very common in the corpus (e.g. EN *low blow*). For PT and FR, we additionally discarded the compounds in which the complement was not an adjective (e.g. PT noun–noun *abelha-rainha* ‘queen bee’ (lit. *bee-queen*)), as these constructions are often seen as exocentric (no head/modifier distinction can be made between the compound elements).

For each language, a balanced set of 60 idiomatic, 60 partially compositional and 60 fully compositional compounds was selected by means of a coarse-grained manual pre-annotation.We eschewed any attempts at selecting equivalent compounds for all three languages. A compound in a given language may correspond to a single word in the other languages, and even when it does translate as another compound, its pattern of POS-tags and its level of compositionality may be widely different.

3.1.2 Sentence selection

For each compound, we selected 3 sentences from the WaC corpus where the compound is used with the same meaning. We sorted them by sentence length, in order to favor shorter sentences, and manually selected 3 examples that satisfy these criteria:

- The occurrence of the compound must have the same meaning in all three sentences.

- Each sentence must contain enough context to enable a clear disambiguation of the compound.
- There must be enough inter-sentence variability, so as to provide a higher amount of disambiguating contexts.

The goal of these sentences was to be used as disambiguating context for the annotators. For example, for the compound *benign tumour*, we present the following disambiguating sentences: (1) “Prince came onboard to have a large *benign tumor* removed from his head”; (2) “We were told at that time it was a slow growing *benign tumor* and to watch and wait”; (3) “Completely *benign tumor* is oncocytoma (it represents about 5 % of all kidney tumors)”.

3.1.3 Questionnaire design

The questionnaires were presented as online webpages, and followed the same structure for each compound. The questionnaire starts with a set of instructions that briefly describe the task and direct participants to fill an external identification form. This form collects demographics about the annotators, and ensures that they are native speakers of the target language, following Reddy, McCarthy and Manandhar (2011). This form also presents some example questions with annotated answers for training. After filling in the identification form, users could start working on the task itself. The questionnaire was structured in 5 subtasks, presented to the annotators through these instructions:

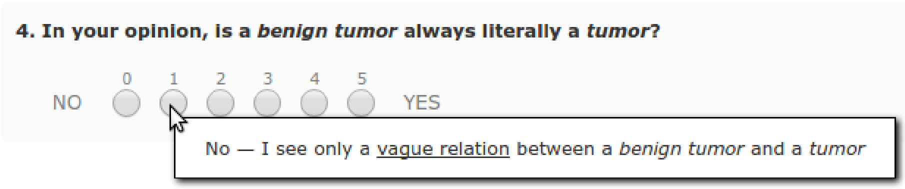
1. Read the compound itself.
2. Read 3 sentences containing the compound.
3. Provide 2 to 3 synonym expressions for the target compound seen in the sentences, preferably involving one of the words in the compound. We ask annotators to prioritize short expressions, with 1 to 3 words each, and to try to include the words from the nominal compound in their reply (eliciting a paraphrase).
4. Using a Likert scale from 0 (completely disagree) to 5 (completely agree), judge how much of the meaning of the compound comes from modifier and head separately.

Figure 3.1 shows an example for the judgment of the literality of the head (*benign*) in the compound *benign tumor*.

5. Using a Likert scale from 0 (completely disagree) to 5 (completely agree), judge how much of the meaning of the compound comes from both of its components (head and modifier). This judgment is requested through a question that paraphrases the compound: "would you say that a *benign tumor* is always literally a *tumor* that is *benign*?".

We have been consciously careful about requiring answers in an even-numbered scale (0–5 makes for 6 reply categories), as otherwise, undecided annotators could be biased towards the middle score. As an additional help for the annotators, when the mouse hovers over a reply to a multiple-choice question, we present a guiding tooltip, as in Figure 3.1. We avoid incomplete replies by making Subtasks 3–5 mandatory.

Figure 3.1: Excerpt from the questionnaire of the compound *benign tumor*, evaluating compositionality regarding the head of the compound.



4. In your opinion, is a *benign tumor* always literally a *tumor*?

No — I see only a vague relation between a *benign tumor* and a *tumor*

The order of subtasks has also been taken into account. During a pilot test, we found that presenting the multiple-choice questions (Subtasks 4–5) before asking for synonyms (Subtask 3) yielded lower agreement, as users were often less self-consistent in the multiple-choice questions (e.g. replying that "*benign tumor* is not a *tumor*" in Subtask 4 while replying that "*benign tumor* is a *tumor* that is *benign*" in Subtask 5). This behavior was observed even when they later carefully selected their synonyms. Asking for synonyms in Subtask 3, prior to the multiple-choice questions, prompts the user focus on the target meaning for the compound and also have more examples (the synonyms) when considering the semantic contribution of each element of the compound. In this work, the synonyms were only used to motivate annotators to think about the meaning of the compound. In the future, this information could be exploited for compositionality prediction, but also for lexical substitution tasks (WILKENS et al., 2017).

3.1.4 Judgment collection

Annotators participated via online questionnaires, with one webpage per compound. For `EN` and `FR`, annotators were recruited and paid through Amazon Mechanical Turk (AMT). For `PT`, we developed a standalone web interface that simulates AMT, as Portuguese speakers were rare in that platform. Annotators for `PT` were undergraduate and graduate students of Computer Science, Linguistics and Psychology. For each compound, we have collected judgments from around 15 annotators.[1]

For each compound, the response from all annotators were gathered up into an average compound score. We obtained the following variables:

- $c_{\mathbf{H}}$: The contribution of the *head* to the meaning of the compound (e.g. is a *busy bee* literally a *bee*?), with standard deviation $\sigma_{\mathbf{H}}$.
- $c_{\mathbf{M}}$: The contribution of the *modifier* to the meaning of the compound (e.g. is a *busy bee* literally *busy*?), with standard deviation $\sigma_{\mathbf{M}}$.
- $c_{\mathbf{WC}}$: The degree to which the *whole compound* can be interpreted as a combination of its parts (e.g. is a *busy bee* a *bee* that is *busy*?), with standard deviation $\sigma_{\mathbf{WC}}$.

The average c scores provide absolute judgments on the compositionality of a compound, ranging from 0 (non-literal or idiomatic) to 5 (literal or compositional). All datasets are freely available online.[2] For a complete list of all compounds, along with their translation, glosses and collected compositionality scores, we refer to Appendix A. (*EN-comp*$_{90}$), Appendix B. (*FR-comp*), and Appendix C. (*PT-comp*).

3.2 Dataset analysis

In this section, we analyze some properties of the datasets. These are performed on the 180 compounds of each language. We also present some graphics focusing on the 90 compounds collected in *EN-comp*$_{90}$, for comparison against the whole dataset of 180 compounds in *Reddy*$^{++}$. In some cases, in order to perform a cross-language analysis of the data, we group the compositionality scores of the three datasets into a single dataset *ALL-comp* with all 3 × 180 compounds.

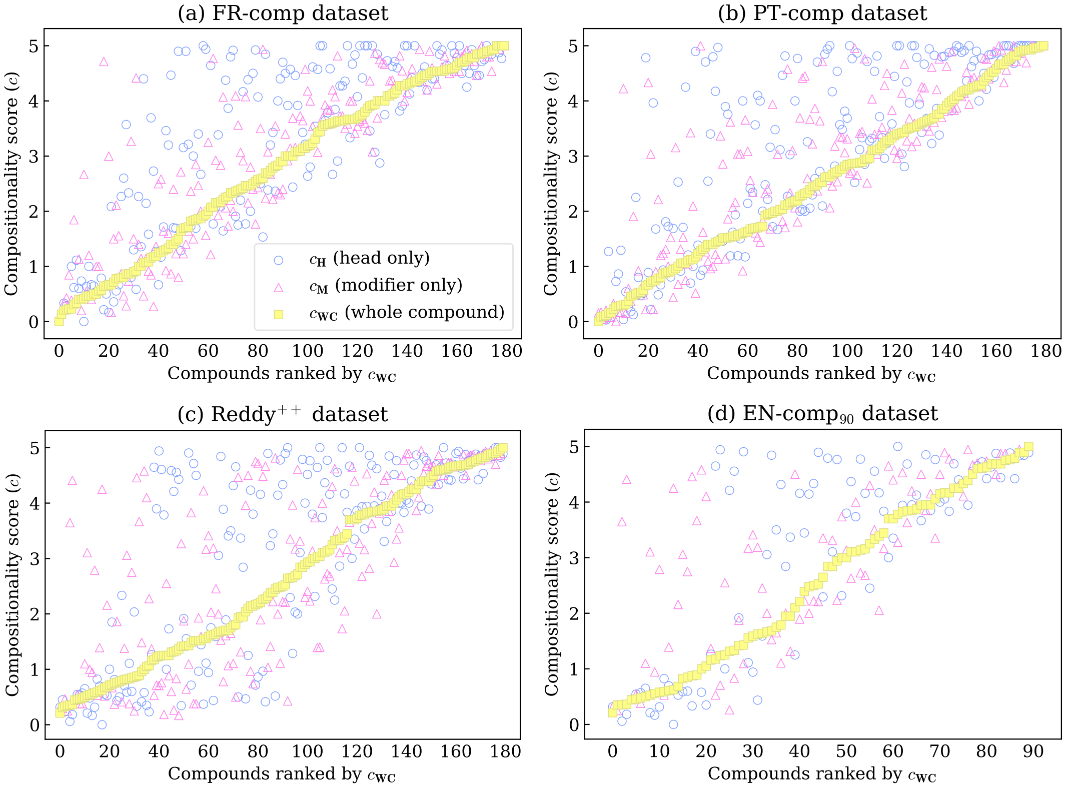
[1] `EN` includes the 90 compounds from Reddy, McCarthy and Manandhar (2011), which are compatible with the other 90 compounds collected for the dataset.

[2] <http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/compounds>

3.2.1 Score distribution

Figure 3.2 presents the average scores for the 180 compounds for each language. Each of the 4 graphs is ordered in the x-axis based on the whole-compound compositionality scores (rank-based $c_{\mathbf{WC}}$). Values in the y-axis then present the average score of each compound (value-based $c_{\mathbf{H}}$, $c_{\mathbf{M}}$ and $c_{\mathbf{WC}}$). The average human judgments confirm that the three datasets are balanced in terms of compound idiomaticity, with the whole-compound scores growing at a mostly linear rate (the correlation between $c_{\mathbf{WC}}$ and the list of numbers 1..180 is statistically significant, with Pearson $r > 0.99$ for all four graphs). Moreover, there seems to be a greater agreement between the score for the compound and that of its head/modifier for the two extremes (totally idiomatic and fully compositional), with a greater dispersion of head/modifier scores for partially idiomatic compounds.

Figure 3.2: Average compositionality ($c_{\mathbf{H}}$, $c_{\mathbf{M}}$ and $c_{\mathbf{WC}}$) per compound.



3.2.2 Difficulty of annotation

For each compound, the difficulty of annotation can be estimated as the standard deviation (σ) among the compositionality scores provided by multiple human raters. Ideally, if all annotators agreed on a compositionality score, σ should be low. Following Reddy, McCarthy and Manandhar (2011), we calculated for each language the number of compounds that had standard deviation greater than 1.5. The results are shown in Table 3.1. The largest deviations happened for modifiers, which suggests that adjectives may be harder for humans to judge than nouns. Indeed, if we consider the average of all standard deviations in *Reddy*$^{++}$, we obtain $\overline{\sigma_{\mathbf{H}}} = 0.97$ and noun-based $\overline{\sigma_{\mathbf{M}=\text{noun}}} = 0.97$ (with 132 cases), but adjective-based $\overline{\sigma_{\mathbf{M}=\text{adj}}} = 1.30$ (with 48 cases). This is in line with the average standard deviation found for the other two languages, where every modifier in the dataset is an adjective. For *FR-comp*, $\overline{\sigma_{\mathbf{H}}} = 1.01$ and $\overline{\sigma_{\mathbf{M}}} = 1.18$; while for *PT-comp*, $\overline{\sigma_{\mathbf{H}}} = 0.84$ and $\overline{\sigma_{\mathbf{M}}} = 0.98$.

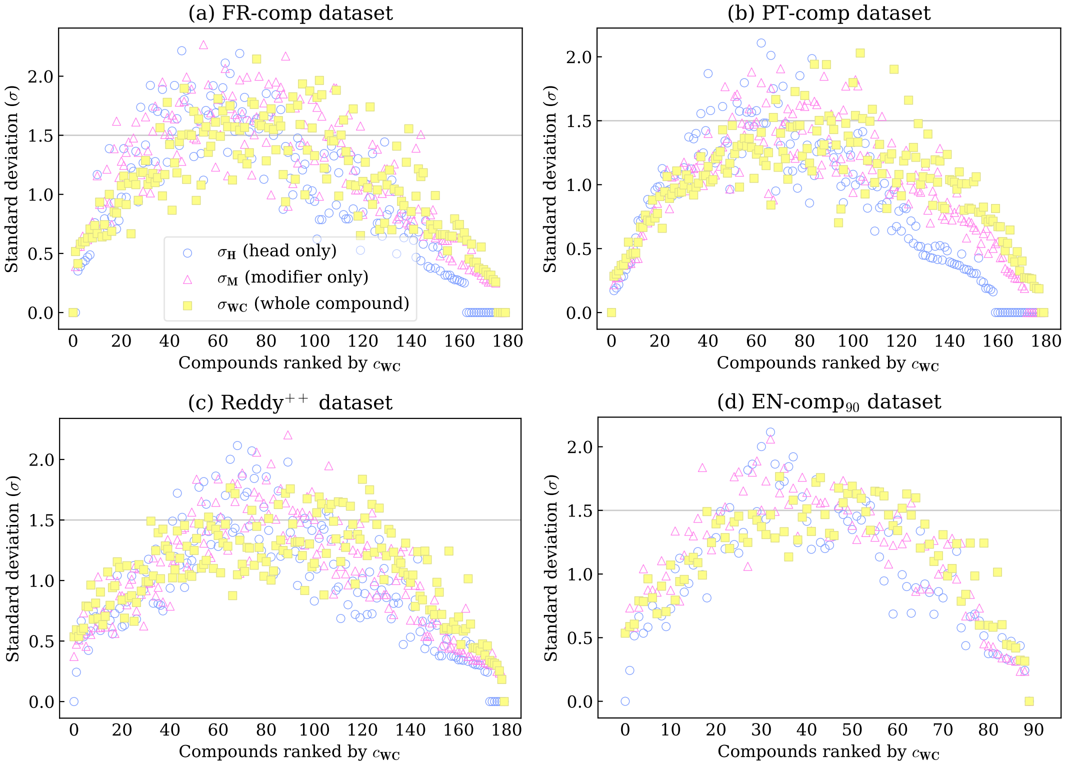
Table 3.1: Number of cases of high standard-deviation σ.

	EN	FR	PT
Compounds with $\sigma_{\mathbf{WC}} > 1.5$	22	41	30
Compounds with $\sigma_{\mathbf{H}} > 1.5$	23	44	33
Compounds with $\sigma_{\mathbf{M}} > 1.5$	35	55	34

Figure 3.3 presents the standard deviation scores of every compound as a function of its average compositionality score. Just as the head/modifier-only scores were closer in value to the whole-compound score in the extremities (e.g. highly idiomatic cases with compositionality $c_{\mathbf{WC}} < 20$ and highly compositional cases with $c_{\mathbf{WC}} > 160$), so are all standard deviations lower in these extremes. This phenomenon may be related to purely statistical effects of extremity values, or may indicate that more extreme judgments are easier for humans to produce. One consistent property seen among all datasets is that the peak of standard deviation occurs in the left side of the graphs (in particular for $\sigma_{\mathbf{H}}$ and $\sigma_{\mathbf{M}}$), suggesting that idiomatic compounds are slightly harder for humans to judge consistently.

The difficulty of annotation can also be measured through inter-rater agreement measures (described in Section 2.2.4). For the English and French datasets, most participants only provided a small amount of annotations, making these measures unfeasible. For the Portuguese dataset, 3 of the participants annotated a large subset of 119 compounds.

Figure 3.3: Standard deviation ($\sigma_{\mathbf{H}}$, $\sigma_{\mathbf{M}}$ and $\sigma_{\mathbf{WC}}$) per compound.



For this subset, the pairwise kappa values range from $\kappa = .28$ to $\kappa = .58$ depending on the question (head-only, modifier-only or whole-compound) and on the annotator pair. In the case of α, there was an agreement of $\alpha = .52$ for head-only, $\alpha = .36$ for modifier-only and $\alpha = .42$ for whole-compound compositionality scores. We also calculated the α between an expert annotator and himself some weeks later. The agreement rate ranges from $\alpha = .59$ for whole-compound and modifier-only, to $\alpha = .69$ for head-only compositionality scores.

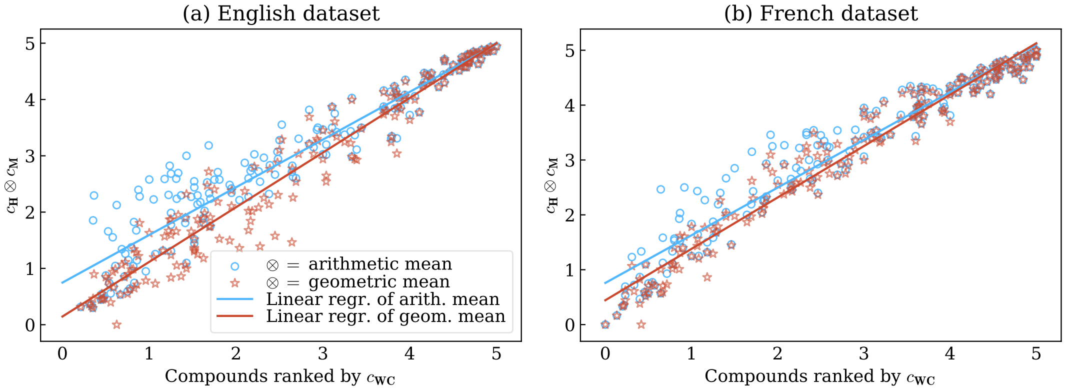
3.2.3 Estimating whole-compound from head/modifier

A careful analysis of the plot presented in Figure 3.2 suggests that the whole-compound score is lower-bounded by the head-only and modifier-only scores, i.e. $c_{\mathbf{WC}} \approx \min(c_{\mathbf{H}}, c_{\mathbf{M}})$. This would mean that the whole-compound compositionality scores is estimated by human raters based on the literality of its elements. We thus evaluated if it was possible to predict the compositionality score of the whole compound from the scores of its parts. To quantify this relation, we used two models: the arithmetic and geometric mean of the head-only and modifier-only scores for that compound.

Figure 3.4 shows the linear regression for both measures in the *Reddy*$^{++}$ and *FR-*

comp datasets. The goodness of fit results for *Reddy*$^{++}$ were $r^2_{\text{arith}} = .90$ for the arithmetic mean, and $r^2_{\text{geom}} = .96$ for the geometric mean, with the latter being a better predictor of the whole-compound compositionality. Similar results were achieved for *FR-comp* ($r^2_{\text{arith}} = .93$, $r^2_{\text{geom}} = .96$), for *PT-comp* ($r^2_{\text{arith}} = .91$, $r^2_{\text{geom}} = .96$) and for *EN-comp*$_{90}$ ($r^2_{\text{arith}} = .90$, $r^2_{\text{geom}} = .96$). This means that, whenever annotators judged an element of the compound as highly non-literal, they have also rated the whole compound as highly idiomatic. Estimating based on the min operation itself yields similar R^2 values as r^2_{geom} for *PT-comp* and *FR-comp*. For *Reddy*$^{++}$, $r^2_{\text{min}} = .90$, indicating that the geometric mean is actually a better estimator than the min function itself. The results of this analysis inspired the proposal of the *geom* compositionality prediction strategy, described in Chapter 4.

Figure 3.4: Distribution of $c_{\mathbf{H}} \otimes c_{\mathbf{M}}$ according to $c_{\mathbf{WC}}$ of each compound.



3.2.4 Correlation with distributional variables

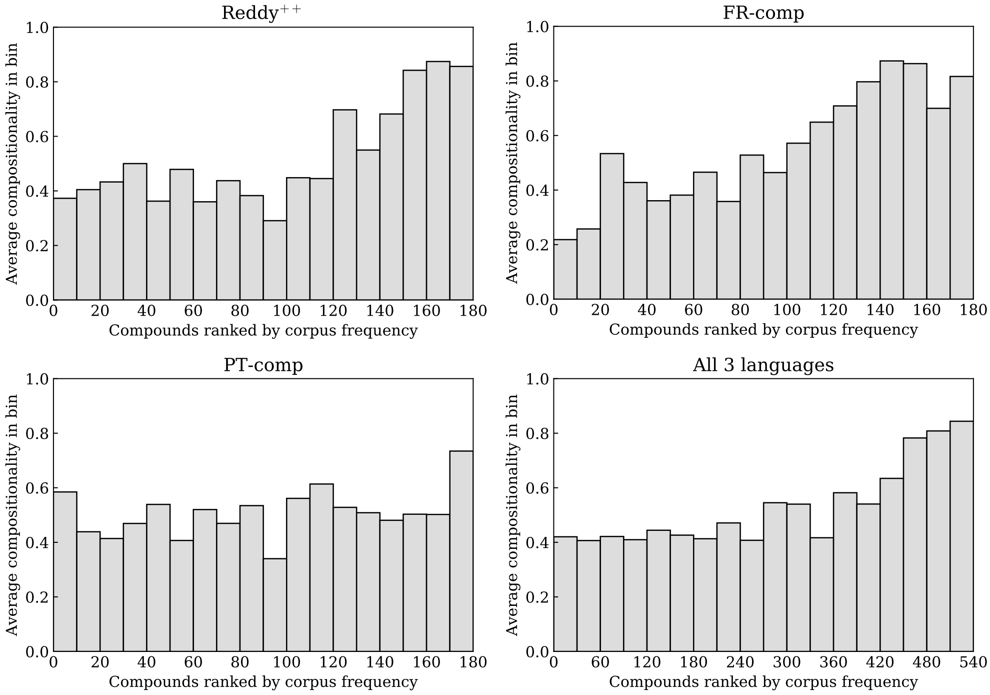
The hypothesis $\text{h}_{\text{idiom} \approx \text{distr.freq}}$ suggests that idiomatic MWEs should occur more frequently than compositional ones in general human communication. As a result, there should be a negative correlation between the compositionality score and the frequency of each compound in a sufficiently general corpus. We thus calculate the correlation between the compositionality score $c_{\mathbf{WC}}$ of each compound and its frequency in the WaC corpora[3]. The result is a statistically significant Spearman correlation of $\rho = .46$ for *Reddy*$^{++}$ and $\rho = .60$ for *FR-comp* (with p-value $p < 10^{-10}$ in both cases). In the case of *PT-comp*, no significant correlation was found (p-value $p > 0.1$). Figure 3.5 presents each dataset

[3] We used the same WaC corpora as in Section 3.1.1.

where compounds were ordered by frequency and grouped into 18 bins of 10 compounds each. The height of each bar indicates the average of the $c_{\mathbf{WC}}$ score assigned by humans to the 10 compounds in the bin. We can see that, in the case of *FR-comp* and *Reddy*$^{++}$, the compounds that are more frequent tend to be assigned higher compositionality scores by humans. These results go against the hypothesis $\mathrm{h}_{\mathrm{idiom} \approx \mathrm{distr.freq}}$, which proposed that idiomatic compounds should be overall rather frequent, so as to permit the assimilation of their meaning (PINKER, 1995)

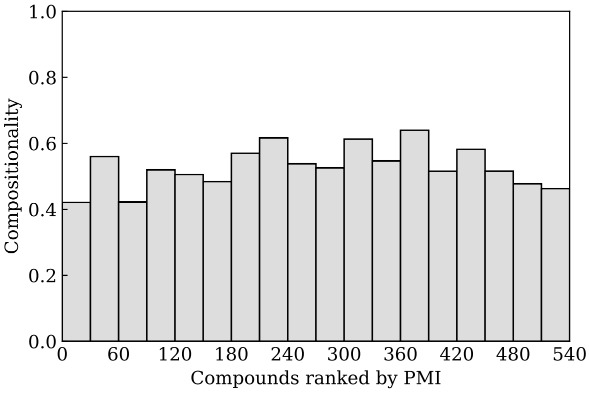
Figure 3.5 also presents a graph where the three datasets were combined so as to form a single set of 3×180 compounds. The height of each bin indicates the average score assigned by humans to the 30 compounds therein. As in the case of English and French data above, this combined dataset presents a statistically significant positive correlation of $\rho = .41$ (with $p < 10^{-24}$) between compositionality scores and corpus frequency. This correlation does not mean that all of these idiomatic compounds are rare in an absolute sense, but it does mean that, among the most frequent compounds in the datasets, the majority of entries is more compositional than the average.

Figure 3.5: Compositionality for compounds under different frequency bins.



We have similarly analyzed the correlation between the compositionality score of each expression and the level of conventionalization (estimated through the PMI). According to the $\mathrm{h}_{\mathrm{idiom} \approx \mathrm{distr.convent}}$ hypothesis, the level of idiomaticity of an MWE should be positively correlated with the PMI. Figure 3.6 presents all 3×180 compounds ordered by PMI and grouped under 18 bins of 30 compounds each. The height of each bin indicates the average score assigned by humans to the 30 compounds therein. As can be seen, there is no clear pattern of correlation between the two variables. Indeed, differently from the case of the frequency, we found no statistically significant correlation between the compositionality scores and the PMI (this holds true for each dataset by itself, as well as when the 3 datasets are combined). This stands in contrast with the fact that many works in the literature rely on association measures as estimators for compositionality (e.g. using PMI in the discovery of idioms) (FAZLY; STEVENSON, 2006; BU; ZHU; LI, 2010; GURRUTXAGA; ALEGRIA, 2013; MAAROUF; OAKES, 2015). Given the lack of correlation between these two variables, we do not recommend the use of PMI as an estimator of compositionality.[4]

Figure 3.6: Compositionality for noun compounds under different PMI bins.



3.3 Summary

In this chapter, we have presented a multilingual group of datasets containing human judgments about the compositionality of nominal compounds. It contains 180 compounds for each of the 3 target languages: English, French and Portuguese. Annotations were collected through crowdsourcing. Since the task was performed by native speakers who may not have a background in linguistics, it needs to be appropriately constrained not to require expert knowledge, and this section has described the methodology that were employed towards that goal.

[4] We leave the investigation of other association measures for future work.

An analysis of the resulting resource confirmed that the compounds are uniformly distributed across different ranges of compositionality scores. The 3 datasets are comparable regarding their difficulty of annotation, with partially compositional and adjective-based compounds consistently posing a higher level of difficulty for the human raters. Head-only and modifier-only judgments have also been compared to whole-compound compositionality judgments, with the latter scores behaving as the geometric mean of the former two. Compositionality scores showed no correlation with a measure of conventionalization, differently from what was predicted by the $\mathrm{h}_{\mathrm{idiom} \approx \mathrm{distr.convent}}$ hypothesis. Finally, compositionality scores consistently showed a positive correlation with compound frequency in a corpus, refuting the common intuition that the most frequent compounds tend to be idiomatic (hypothesis $\mathrm{h}_{\mathrm{idiom} \approx \mathrm{distr.freq}}$).

The datasets presented in this chapter can be used to evaluate applications and tasks requiring some degree of semantic processing, such as lexical substitution and text simplification. For the cases where the numerical judgments alone are not enough for a given task, the datasets also provide sets of paraphrases, which serve as a symbolic counterpart to those scores. These datasets have also been described and analysed in dedicated publications (RAMISCH et al., 2016; CORDEIRO; RAMISCH; VILLAVICENCIO, 2016a; WILKENS et al., 2017). The complete resource is freely available online.[5]

[5] <http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/compounds>

4 COMPOSITIONALITY PREDICTION

Multiword expressions exist in a wide spectrum of idiomaticity, ranging from mere statistically-idiosyncratic compositional combinations of words (such as *beach towel*, which refers to an actual towel), to completely opaque idioms (such as *eager beaver*, which refers to an enthusiastic person). Precision-oriented NLP systems must distinguish between the different levels of compositionality in order to appropriately handle these different kinds of MWEs. The level of MWE compositionality has been measured through a numerical representation in multiple datasets. However, the coverage of these datasets is limited by the availability of human resources.

This chapter focuses on the task of *compositionality prediction*, which consists in automatically identifying the level of compositionality of MWEs without the input of human raters. The core of this thesis consists in the evaluation of a compositionality prediction model under multiple DSMs and with a variety of parameters. Section 4.1 presents the related work on compositionality prediction. Section 4.2 then presents the compositionality prediction model that was proposed and implemented for this thesis. The remainder of this chapter describes the organization of the experiments for the evaluation of this model: corpus preprocessing (Section 4.3), DSMs (Section 4.4), parameters (Section 4.5), and the evaluation setup (Section 4.6). The evaluation of compositionality prediction models can be performed intrinsically or extrinsically:

- Intrinsic evaluation requires the existence of a dataset in which each MWE is associated with a compositionality score (e.g. the datasets presented in Section 2.5, as well as the resources developed in Section 3), serving as a gold standard. The compositionality prediction model is used to predict those scores, which are then directly compared to the gold standard using a correlation measure (such as those described in Section 2.2.5). This is the approach followed in Chapter 5, which analyses the results of compositionality prediction for nominal compounds under thousands of experimental setups.

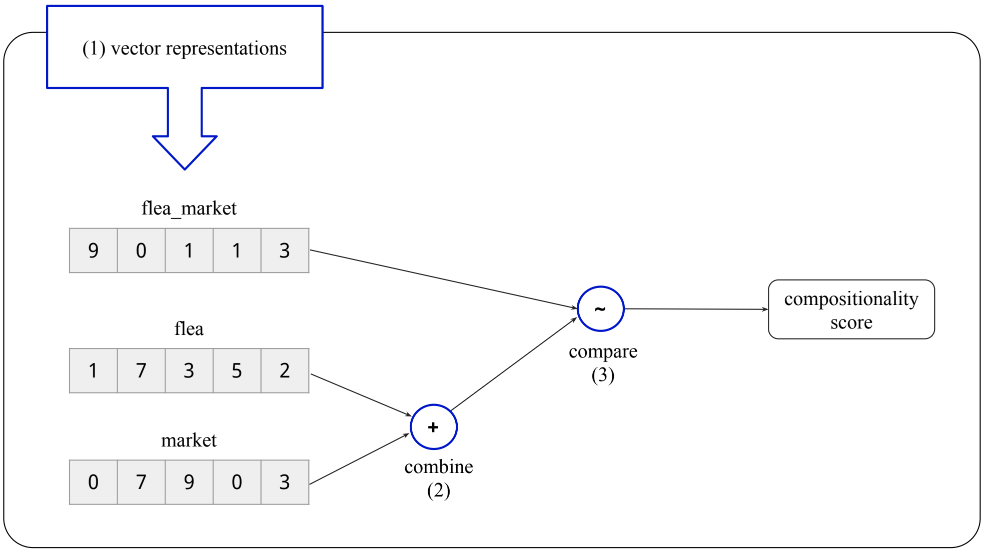
- In extrinsic evaluation, predicted compositionality scores can be used to decide how an MWE should be treated in NLP systems. For example, in an application such as machine translation, idiomatic MWEs should be identified and translated as an atomic unit. As a consequence, an evaluation of machine translation quality focusing on MWEs would indirectly reflect the ability of the system to predict compositionality (CAP et al., 2015; STYMNE; CANCEDDA; AHRENBERG, 2013;

SALEHI et al., 2015). A less application-focused alternative would be the evaluation of the usefulness of predicted compositionality scores in a task of identifying idiomatic MWE occurrences in a corpus. The predicted compositionality scores would then be used as a feature in the underlying MWE identification system. The evaluation of this system would then compare the automatically identified MWEs with a gold standard corpus, resulting in an indirect evaluation of the compositionality prediction scores. This is the approach followed in Chapter 6.

4.1 Related work

Compositionality prediction techniques usually involve measuring the extent to which the meaning of an expression is constructed from a combination of the meaning of its parts. One of the most common setups requires three ingredients: (1) vector representations of single word meanings, such as those built using DSMs; (2) a mathematical model of how the compositional meaning of a phrase should be calculated, as a combination of the single-word meaning of its parts; and (3) a measure of similarity, used to compare the compositionally-constructed meaning of a phrase and its own meaning derived from corpora.[1] Figure 4.1 presents this compositionality prediction architecture, along with the three main ingredients. For each MWE (e.g. *flea market*), the DSM vec-

Figure 4.1: Common compositionality prediction architecture



[1] These setups often assume that each word corresponds to a single meaning (i.e. no ambiguity is taken into account).

tors of its elements are combined into a single vector, which is then compared against the vector for the MWE built from its occurrences in the corpus. For each of the three ingredients, there are a number of different alternatives that can be seen employed in the literature. Throughout this thesis, we will refer to a specific choice of the three ingredients as a *compositionality prediction model*.

The 1^{st} ingredient applies to any kind of numerical representation of word semantics. All of the representations discussed in Section 2.4.2 can be equivalently used in this architecture, and different works in the literature will focus on different representations (e.g. Reddy, McCarthy and Manandhar (2011) use a co-occurrence matrix with a limited vocabulary V comprising $|V| = 10\,000$ words, while Yazdani, Farahmand and Henderson (2015) use word embeddings with a reduced number of dimensions). In most works, only a single DSM system is evaluated, under a limited set of parameters. This thesis will consider multiple DSMs under a variety of corpus and DSM parameters.

The 2^{nd} ingredient concerns the mathematical model of meaning combination. One of the most natural choices is the additive model, in which the compositional meaning of an MWE $w_1 w_2 \dots w_N$ is predicted as a linear combination of the word vectors of its components: $\sum_i \beta_i \mathbf{v}(w_i)$, where the β coefficients assign different weights for the representation of each word, and $\mathbf{v}(w_i)$ is a D-dimensional word vector for word w_i, with $D \leq |V|$ (REDDY; MCCARTHY; MANANDHAR, 2011; WALDE; MüLLER; ROLLER, 2013; SALEHI; COOK; BALDWIN, 2015). These different weights can capture asymmetric contributions by each of the components (BANNARD; BALDWIN; LASCARIDES, 2003; REDDY; MCCARTHY; MANANDHAR, 2011). For example, in the expression *couch potato* (which refers to a stereotypical person who spends a lot of time sitting down and watching television), it is the first word that has a clear contribution to the word meaning, and the highest weight should be in β_1. In the MWE *flea market*; it is the second word that contributes the most, and the highest weight should thus be in β_2.

The additive model of composition can be generalized so as to use a matrix of multiplicative coefficients, which can be estimated through linear regression (GUEVARA, 2011). This model can be further modified so as to learn polynomial projections of higher degree, with quadratic projections yielding particularly promising results (YAZDANI; FARAHMAND; HENDERSON, 2015). These models come with the caveat of being supervised approaches, thus requiring some amount of pre-annotated data in the target language. Due to these requirements, most works focus on unsupervised compositionality prediction methods only, based exclusively on large monolingual unannotated corpora. The latter is also the approach adopted in this thesis.

Alternatives to the linear models include the multiplicative model and its variants (MITCHELL; LAPATA, 2008). However, results suggest that this representation yields inferior results when compared to the predictions obtained through the additive model (REDDY; MCCARTHY; MANANDHAR, 2011; SALEHI; COOK; BALDWIN, 2015). Recent work on predicting intra-MWE semantics also supports the hypothesis that additive models tend to yield better results (HARTUNG et al., 2017). This thesis evaluates different variants of compositionality prediction models (see the prediction strategies in Section 4.2).

The 3$^{\text{rd}}$ ingredient is a measure of similarity, used to compare the MWE with the sum of its parts. Most works in the literature rely on cosine similarity (SCHONE; JURAFSKY, 2001; MITCHELL; LAPATA, 2008; FARAHMAND; SMITH; NIVRE, 2015) but it also can be calculated in terms of the overlap between the profiles of word distribution (MCCARTHY; KELLER; CARROLL, 2003), assuming that compositional expressions are more similar or share more semantic neighbors with their components than idiomatic ones. In this thesis, the cosine similarity will be used for all evaluations.

Compositionality prediction can also be achieved through an estimation of the likelihood of an MWE (e.g. *red-blood cell*) being replaced by single-word terms in a corpus (e.g. *erythrocyte*). MWEs that can be replaced by many single-word terms are then deemed idiomatic (RIEDL; BIEMANN, 2015). An alternative method would be to compare an MWE and its constituents across multiple translations of a text. If the MWE is translated literally, it is predicted as compositional, while non-literal translations are interpreted as an indication of idiomaticity (SALEHI; COOK; BALDWIN, 2014). The number of possible translations has also been used as an indicator of idiomaticity (CAP, 2017).

The level of compositionality of MWEs may also be predicted in the context of MWE-annotated sentences. This is particularly beneficial in the presence of ambiguous MWEs, whose degree of compositionality depends on the context (SPORLEDER; LI, 2009). One such type of ambiguity arises from the possibility of literal and non-literal interpretations for the same lexical unit (e.g. in the expression *spill the beans*). For instance, Köper and Walde (2016) evaluate the impact of different features on the prediction of the literality of German verb-particle constructions. Their features range from the use of a bag-of-words model to distributional statistical scores. The experiments in this thesis focus on non-ambiguous MWEs, so we do not present results for context-dependent compositionality prediction.

4.2 Proposed model

The compositionality principle assumes that the meaning of phrases and sentences can be derived from a combination of the meaning of their components. While this may hold for compositional MWEs, for idiomatic cases we expect the opposite to be true: by combining the semantic representations of the parts of an MWE, we should obtain a representation that is *different* from the representation of the MWE derived directly from corpora. This behavior can be exploited in the construction of the *compositionality prediction model* that was implemented for this thesis and which we evaluate in Chapters 5 and 6.

For each MWE (e.g. *flea market*), let the unitized representation $\mathbf{v}_{\text{u}}$ be the vector representation built for the MWE as a whole, as seen in the corpus[2]:

$$\mathbf{v}_{\text{u}}(w_i w_2 \dots w_N) = \mathbf{v}(w_i\_w_2\_\dots\_w_N).$$

Define the combined vector $\mathbf{v}_\beta$ as a function of the individual meaning of the MWE elements (e.g. *flea* and *market*). This vector is calculated through an additive operation of vector composition[3]:

$$\mathbf{v}_\beta(w_1 w_2 \dots w_N) = \beta_1 \frac{\mathbf{v}(w_1)}{||\mathbf{v}(w_1)||} + \beta_2 \frac{\mathbf{v}(w_2)}{||\mathbf{v}(w_2)||} + \dots + \beta_N \frac{\mathbf{v}(w_N)}{||\mathbf{v}(w_N)||}.$$

where $\beta_i \in [0, 1]$ is a parameter that controls the relative importance of each word for the combined representation, with $\sum_{i=1}^{N} \beta_i = 1$. The compositionality prediction model may then calculate the compositionality score CS as the cosine similarity between the unitized representation $\mathbf{v}_{\text{u}}$ and the combined representation $\mathbf{v}_\beta$:

$$\text{CS}_\beta(w_1 w_2 \dots w_N) = \cos(\mathbf{v}_{\text{u}}(w_1 w_2 \dots w_N),\ \mathbf{v}_\beta(w_1 w_2 \dots w_N)).$$

[2] In the corpus preprocessing stage, the components of the target MWEs are linked by "_" to be treated as a single token; e.g. $w_1\_w_2$ = *flea_market* (see Section 4.3).

[3] DSM vector length is usually not considered meaningful, and is implicitly normalized during the calculation of the cosine. We do explicitly normalize vector length for the computation of $\mathbf{v}_\beta$ to properly apply the β weights.

The above definition of the compositionality score leaves the precise values of β unspecified. In this thesis, we will consider the following composition strategies:

- Uniform, which defines $\beta_i = \frac{1}{N}$ (e.g. $\beta_1 = \beta_2 = 0.5$ for a 2-word MWE). Equal weights are assigned to every word in the MWE, assuming that they all contribute equally to the meaning of the MWE (as in the MWE *access road*). This is the most commonly used prediction strategy in the literature (MITCHELL; LAPATA, 2010), and most of the results in this thesis will focus on the scores obtained through this weighting strategy.
- Head, for 2-word MWEs, which defines $\beta_{\text{head}} = 1$ and $\beta_{\text{mod}} = 0$, i.e. the modifier is considered to make no contribution to the semantics of the MWE, and the meaning comes from the head alone (as in *crocodile tears*).
- Mod, for 2-word MWEs, which defines $\beta_{\text{head}} = 0$ and $\beta_{\text{mod}} = 1$, i.e. the meaning of the MWE is assumed to come from the modifier, while the head is assumed to make no contribution towards the meaning of the whole (as in the expression *night owl*).
- Maxsim, where the set of weights β_i in the construction of $\mathbf{v}_\beta$ is defined so as to maximize the value of CS, i.e.:

  $$\beta = \underset{\text{X}}{\text{argmax}} \ \text{CS}_{\beta=\text{X}}(w_1 w_2 \ldots w_N).$$

  As a consequence, *maxsim* is capable of expressing the 3 previous prediction strategies as a combination of weights. This model has been developed for the purpose of this thesis. The underlying hypothesis ($\text{h}_{\text{strat.maxsim}}$) is that this model is a better predictor of compositionality scores for compositional MWEs, as it constructs a vector with weights that are optimal for a compositional reading.

  Note that, in the case of $N = 2$, a closed formula can be derived for the calculation of β values, which avoids an exhaustive search of the parameter space (and is what we used for the experiments in Chapter 5). Let $\beta_2 = 1 - \beta_1$. We want to perform the following maximization:

  $$\beta_1 = \underset{y}{\text{argmax}} \cos\left(\mathbf{v}_{\text{u}}(w_1 w_2),\ y\frac{\mathbf{v}(w_1)}{||\mathbf{v}(w_1)||} + (1-y)\frac{\mathbf{v}(w_2)}{||\mathbf{v}(w_2)||}\right).$$

This can be achieved by differentiating the right side of the equation:

$$\frac{\mathrm{d}}{\mathrm{d}\beta_1} \cos\Big(\mathbf{v}_{\mathrm{u}}(w_1 w_2),\ \ \beta_1 \frac{\mathbf{v}(w_1)}{||\mathbf{v}(w_1)||} + (1-\beta_1)\frac{\mathbf{v}(w_2)}{||\mathbf{v}(w_2)||}\Big) \ = \ 0.$$

Replacing the cosine by the definition based on the dot product (see page 46) and solving for β_1, we obtain the closed-form solution:

$$\beta_1 = \frac{\cos(\mathbf{v}(w_1), \mathbf{v}_{\mathrm{u}}(w_1 w_2)) \ - \ \cos(\mathbf{v}(w_1), \mathbf{v}(w_2)) \ \cdot \ \cos(\mathbf{v}(w_2), \mathbf{v}_{\mathrm{u}}(w_1 w_2))}{\Big(1 - \cos(\mathbf{v}(w_1), \mathbf{v}(w_2))\Big) \ \cdot \ \Big(\cos(\mathbf{v}(w_1), \mathbf{v}_{\mathrm{u}}(w_1 w_2)) + \cos(\mathbf{v}(w_2), \mathbf{v}_{\mathrm{u}}(w_1 w_2))\Big)}.$$

We additionally consider two variations on the compositionality score that do not rely on the construction of a combined representation $\mathbf{v}_\beta$. They compare $\mathbf{v}_{\mathrm{u}}$ directly to the individual word vectors instead. The two relevant composition strategies are:

- Arith, which calculates the cosine between the MWE and each component individually, and yields the arithmetic mean of the cosines as the compositionality score; i.e.:

  $$\mathrm{CS}_{\mathrm{A}}(w_1 w_2 \ldots w_N) = \frac{1}{N}\Big(\sum_{i=1}^{N} \cos\Big(\mathbf{v}_{\mathrm{u}}(w_1 w_2 \ldots w_N),\ \mathbf{v}(w_i)\Big)\Big).$$

  While some works in the literature seem to favor the *uniform* model, some results have been published for *arith* (REDDY; MCCARTHY; MANANDHAR, 2011; WALDE; MüLLER; ROLLER, 2013; SALEHI; COOK; BALDWIN, 2015). To date, no work has compared the behavior of these two models. Given that both strategies represent an additive interpretation of the vectors, our hypothesis ($\mathrm{h}_{\text{strat.arith} \approx \text{ strat.uni}}$ is that the highest-ranking configurations of both models should obtain similar scores.

- Geom, which calculates the cosine between the MWE and each component individually, and yields the geometric mean of the cosines as the compositionality score; i.e.:

  $$\mathrm{CS}_{\mathrm{G}}(w_1 w_2 \ldots w_N) = \Big(\prod_{i=1}^{N} \cos\Big(\mathbf{v}_{\mathrm{u}}(w_1 w_2 \ldots w_N),\ \mathbf{v}(w_i)\Big)\Big)^{\frac{1}{N}}.$$

  This model is inspired by results found in Section 3.2, which suggest that humans interpret whole-MWE composition as the geometric mean of the composition of its parts. While *maxsim* optimizes for higher scores of compositionality, and should thus yield better results for compositional MWEs, we hypothesize ($\mathrm{h}_{\text{strat.geom}}$) that *geom* should obtain higher scores for idiomatic MWEs, due to its tendency to predict lower scores if either of the components is judged as non-compositional regarding the whole.

Other optimized functions such as the ones proposed by Yazdani, Farahmand and Henderson (2015) could also be verified, but are out of the scope of this thesis, as they are based on supervised learning. A multiplicative version of *uniform* has also been considered in early experiments, but it did not present promising results, in particular in the case of sparse representations such as *PPMI–thresh*, in which the product of any two vectors tends to have an impractical number of non-zero dimensions.

The compositionality prediction model proposed in this thesis was implemented as part of the mwetoolkit[4]. The code is publicly available, and contains an internal module for the interpretation of multiple DSM formats, including sparse-context representations (such as the output of *minimantics*[5]) and dense representations (such as the output of *word2vec*[6]). Internally, the set of word vectors is represented as a sparse mapping from $(\text{target}, \text{context})$ word-form pairs to a real number; i.e. $(V, V) \rightarrow \mathbb{R}$. In the case of dense (fixed-length) input vectors, where there is no clear semantics attached to each dimension, we generate artificial identifiers for the context (c_0, c_1, $c_2 \ldots c_{D-1}$). This allows a unified view of all types of vector representations.

In the implementation of the model, the sparse mapping from target–context pairs to a numerical representation is instantiated as a hash-table. Only non-zero mappings are explicitly represented, and thus all missing $(\text{target}, \text{context})$ pairs are assumed to map to 0. When launching the module, a parameter can be specified to chose from among the aforementioned composition strategies. The compositionality scores are then predicted according to the model. Several types of output format can be specified, including a comprehensive XML format and a more lightweight CSV output. A full description of the compositionality prediction model and the associated tool has been published as Cordeiro, Ramisch and Villavicencio (2016b).

4.3 Corpus preprocessing

Experiments in this thesis are based on distributional models built for English, French and Portuguese. The construction of these models uses the lemmatized and POS-

[4] <http://mwetoolkit.sf.net>
[5] <https://github.com/ceramisch/minimantics>
[6] <https://code.google.com/archive/p/word2vec>

tagged versions of the following corpora:

- For English, the ukWaC (BARONI et al., 2009), with 2.25 billion tokens, parsed with MaltParser (NIVRE; HALL; NILSSON, 2006).
- For French, the frWaC, with 1.61 billion tokens preprocessed with TreeTagger (SCHMID, 1995).
- For Portuguese, a combination of brWaC (BOOS; PRESTES; VILLAVICENCIO, 2014), Corpus Brasileiro[7] and all Wikipedia articles[8], with a total of 1.91 billion tokens. This corpus was obtained in raw form, and parsed with PALAVRAS (BICK, 2000) for the specific purpose of this thesis.

Target MWEs in these corpora are re-tokenized so as to be represented by a single token, with its components joined by an underscore character (e.g. the surface form `EN` *monkey business* → *monkey_business* and `FR` *belle-mère* → *belle_mère*).

During initial experiments, we noticed an inconsistency in the POS tags of MWE occurrences (e.g. the joined token *sitting_duck* had most of its occurrences tagged as `VERB_NOUN` instead of `ADJ_NOUN`). To handle such errors, we also re-tag every annotated occurrence of an MWE with a global manually selected POS tag.[9]

All forms are then lowercased (surface forms, lemmas and POS tags); and noisy tokens, with special characters, numbers or punctuation, are removed. Additionally, ligatures are normalized for French (e.g. *œ* → *oe*) and a spellchecker[10] is applied to normalize words across English spelling variants (e.g. *color* → *colour*). Additionally, proper nouns are replaced by a placeholder to reduce data sparsity.

To evaluate the influence of preprocessing in model accuracy (see Section 5.3.1), we generated four versions of each corpus, with decreasing levels of specificity in the informational content of each token:

1. *surface*$^{+}$: the surface-level forms of every word in the original corpus, with only the preprocessing described above. Example:

```
she is not interested in your fake crocodile_tears !
```

[7] <http://corpusbrasileiro.pucsp.br/cb/Inicial.html>

[8] Wikipedia articles downloaded on June 2016.

[9] For simplicity, our work assumes that every annotated occurrence is an instance of an MWE. We do not account for literal readings, such as cases of *sitting duck* that refer to an actual duck that is sitting.

[10] <https://hunspell.github.io>

2. *surface*: stopword removal[11]; generating a corpus of surface forms for content words only (i.e. nouns, adjectives, adverbs and verbs). Example:

   ```
   is not interested fake crocodile_tears
   ```

3. *lemma*$_{PoS}$: stopword removal, lemmatization[12] and POS tagging; generating a corpus of content words distinguished by POS tags, encoded in the format *lemma/tag*. This conflates the multiple inflectional forms of a word while maintaining the information of its grammatical category. Example:

   ```
   be/VERB not/ADV interested/ADJ fake/ADJ crocodile_tear/N_N
   ```

4. *lemma*: stopword removal and lemmatization without POS tagging; generating a corpus containing only lemmas of content words. This conflates identically-spelled words of different grammatical categories. Example:

   ```
   be not interested fake crocodile_tear
   ```

4.4 DSMs

One of the goals of this thesis is to evaluate the proposed model of compositionality prediction under a variety of distributional settings. In particular, we verify the impact of different types of DSMs (previously described in Section 2.4) in the predictive abilities of the model. For reproducibility, we present below the fixed parameters that were used in the DSM instantiations:[13]

PPMI We consider three DSMs based on positive pointwise mutual information (PPMI). In all cases, the representation of a target word is a vector containing the PPMI association scores between the target and its contexts. The contexts are nouns and verbs, selected in a symmetric sliding window of w words to the left/right and weighted linearly according to their distance d to the target (LEVY; GOLDBERG; DAGAN, 2015). We consider three models that differ in how the contexts are selected:

[11] Stopword removal reduces the size of the corpus. Given that only nouns and verbs are used as contexts, the resulting co-occurrence matrices for *surface* will be less sparse than the matrices for *surface*$^+$, for a given window size.

[12] In the lemmatized corpora, the lemmas of proper names are replaced by placeholders.

[13] These parameters were selected with the goal of homogenizing the configurations across DSMs, and to follow the original paper's recommendations in the cases where the default differs.

- *PPMI–thresh*, where the vectors are $|V|$-dimensional but only the top δ local contexts with highest PPMI for each target word have non-zero values (PADRÓ et al., 2014).

- *PPMI–TopK*, where the vectors are k-dimensional, with a fixed global list of k words to be considered as context. We have defined k as the 1000 most frequent words in the corpus after removing the top 50 most frequent words, replicating the setup from Salehi, Cook and Baldwin (2015).

- *PPMI–SVD*, where SVD is used to factorize the PPMI matrix and reduce its dimensionality from $|V|$ to D.[14] We use a Context Distribution Smoothing of 0.75 and negative sampling of 5 for the SVD (LEVY; GOLDBERG; DAGAN, 2015).

w2v We perform experiments on both variants of word2vec (MIKOLOV et al., 2013): continuous bag-of-words (*w2v–cbow*) and skip-gram (*w2v–sg*). The models are built with default configurations, except for the following: no hierarchical softmax; negative sampling of 25; frequent-word downsampling weight of 10^{-6}; execution of 15 training iterations. We use the default minimum word count threshold of 5.

glove GloVe implements a factorization of the co-occurrence count matrix (PENNINGTON; SOCHER; MANNING, 2014). We use its default configurations, except for the following: internal cutoff parameter $x_{max} = 75$; co-occurrence matrix is built in 15 iterations. For lemma-based models, we use the minimum word count threshold of 5. Due to the large vocabulary size, we use a threshold[15] of 15 for *surface* and 20 for *surface*$^+$.

lexvec The lexvec model (SALLE; VILLAVICENCIO; IDIART, 2016) factorizes the PPMI matrices, strongly penalizing prediction errors on frequent words. We use default configurations, except for the following: negative sampling of 25; subsampling threshold of 10^{-6}; processes the corpus for 15 iterations. Due to the large vocabulary size, we use a minimum word count threshold of 10 for lemma-based models and 100 for *surface* and *surface*$^+$.[16]

[14] We use the hyperwords toolkit: <https://bitbucket.org/omerlevy/hyperwords>

[15] Thresholds were selected so as to not use more than 128 GB of RAM during the construction of a DSM instance.

[16] This is in line with the authors' threshold suggestion in their paper.

4.5 Parameters

For every DSM, we construct multiple distributional models under different sets of configurations. In particular, we exhaustively evaluate the influence of the following variables:

- WordForm: One of the four word-form and stopword removal variants when representing a corpus: *surface*$^+$, *surface*, *lemma*, and *lemma*$_{PoS}$ (see Section 4.3). These variants were selected so as to represent different levels of specificity in the informational content of the tokens.
- WindowSize: Indicates the number of context words that will be considered to the left/right of the target word when searching for target-context co-occurrence pairs. We evaluate the behavior of compositionality prediction when the underlying DSM model is built with context window sizes of 1+1, 4+4, and 8+8.[17]
- Dimension: We generate models with 250, 500 and 750 dimensions. The underlying hypothesis is that, the higher the number of dimensions, the more accurate the representation of the context is going to be.

Table 4.1 presents the set of all possible parameter configurations. These combinations produce a total of 228 models per language (12 models for *PPMI–TopK*, 36 models for each of the other 6 DSMs). Throughout the thesis, when referring to a particular model configuration, an abbreviated notation will be used. For example, WordForm=*lemma*, with WindowSize=4+4 and Dimension=250 will be represented as *lemma*.w$_4$.d$_{250}$.

Table 4.1: 228 parameter combinations across all DSMs.

DSM	Dimension	WordForm	WindowSize
PPMI–TopK	$D = 1000$	*surface*$^+$, *surface*, *lemma*, *lemma*$_{PoS}$	1+1, 4+4, 8+8
PPMI–thresh	$D = \lvert V \rvert$, $\delta \in \{250, 500, 750\}$		
PPMI–SVD *w2v–cbow* *w2v–sg* *glove* *lexvec*	$D \in \{250, 500, 750\}$		

[17] Most works in the literature choose a window of size between the extremes 1+1 and 10+10, with a few works considering higher window sizes such as 16+16 or 20+20 (KIELA; CLARK, 2014; LAPESA; EVERT, 2014).

4.6 Evaluation setup

For the intrinsic evaluation of the compositionality prediction model in Chapter 5, we calculate the compositionality scores for every MWE in a dataset and compare them to the human-rated scores. The following datasets are evaluated:

- For English: *Reddy*, *Reddy*$^{++}$, *EN-comp*$_{90}$ and *Farahmand*;
- For French: *FR-comp*;
- For Portuguese: *PT-comp*.

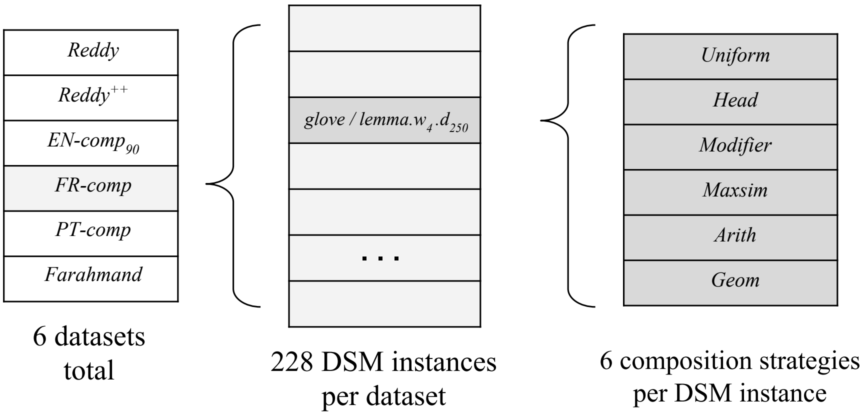
The datasets *Reddy* and *Farahmand* were described in Section 2.5.3. The other datasets were constructed as part of this thesis, and were described in Chapter 3.

For most datasets, we report Spearman's ρ correlation between the ranking provided by humans and those calculated from the models (as explained in Section 2.2.5). Exceptionally for the *Farahmand* dataset, due to the binary nature of its compositionality scores, we follow Yazdani, Farahmand and Henderson (2015) and report the best F_1 score (BF1), obtained by calculating the F_1 score for the top k MWEs classified as positive (non-compositional), for all possible values of k (described in Section 2.2.6).

Evaluation metrics were calculated for a total of more than 8 thousand models (see Figure 4.2). Given the high number of experiments performed, we report the best performance of each model parameter. For instance, the performances reported for *w2v–cbow* using different values of WindowSize are the best configurations across all possible values of other parameters (i.e. Dimension and WordForm). This avoids reporting local maxima that can arise if one fixes all other parameters when evaluating a given one (LAPESA; EVERT, 2014).

For English datasets, we distinguish between *strict* and *fallback* evaluation. Strict evaluation corresponds to the performance of the model only on those MWEs that have a vector representation in all underlying DSMs: 89 (out of 90) for *Reddy*, 86 (out of 90) for *EN-comp*$_{90}$, 175 (out of 180) for *Reddy*$^{++}$, and 913 (out of 1042) for *Farahmand*. Fallback evaluation considers the full dataset, using a fallback strategy for the imputation of missing values, assigning the average of other compositionality scores to MWEs in which one of the vectors $\mathbf{v}(w_1)$, $\mathbf{v}(w_1)$ or $\mathbf{v}_{\mathrm{u}}$ has not been built due to a lack of occurrences in the corpus (SALEHI; COOK; BALDWIN, 2015). This distinction is particularly important

Figure 4.2: Number of compositionality prediction models evaluated: $6 \times 228 \times 6 = 8\,208$.



in the case of *Farahmand*, which contains more rare MWEs[18] such as *universe human* and *mankind instruction*, so that 129 MWEs do not occur often enough in the English corpus. Strict evaluation allows us to properly evaluate the quality of the predictive method itself, while fallback evaluation allows us to evaluate the quality of corpus + method, and is the better alternative when comparing to state-of-the-art results (as it considers the whole dataset). Only strict evaluation is reported for *FR-comp* and *PT-comp*, as all MWEs are frequent enough in their respective corpora.

To determine whether the results for different DSM configurations are statistically different from each other, we present results from Wilcoxon's sign-rank test (REY; NEUHÄUSER, 2011).

[18] Partly due to Wikipedia tokenization errors.

5 INTRINSIC EVALUATION OF COMPOSITIONALITY PREDICTION

This chapter presents an extensive intrinsic evaluation of the compositionality prediction framework presented in Chapter 4, using the three datasets whose construction was presented in Chapter 3. We construct DSM instances under multiple configurations for the three target languages. For each configuration, we generate compositionality predictions for all nominal compounds (NCs), which we then compare with the compositionality scores provided by humans.

This chapter is organized as follows: Section 5.1 presents our findings on the accuracy of compositionality prediction models using state-of-the-art DSMs for the representation of word semantics. Section 5.2 investigates the impact of DSM-specific parameters related to the size of the context window and the number of dimensions used to represent context. Section 5.3 examines the impact of corpus parameters related to corpus size and to the degree of corpus preprocessing adopted. Section 5.4 extends the evaluations performed on *uniform* prediction so as to encompass five other prediction strategies. Section 5.5 performs a variety of sanity checks involving other model-specific parameters. Section 5.6 presents an error analysis comparing predicted compositionality and different variables associated with the compounds. Finally, Section 5.7 summarizes the results from this chapter.

5.1 Overall highest results per DSM

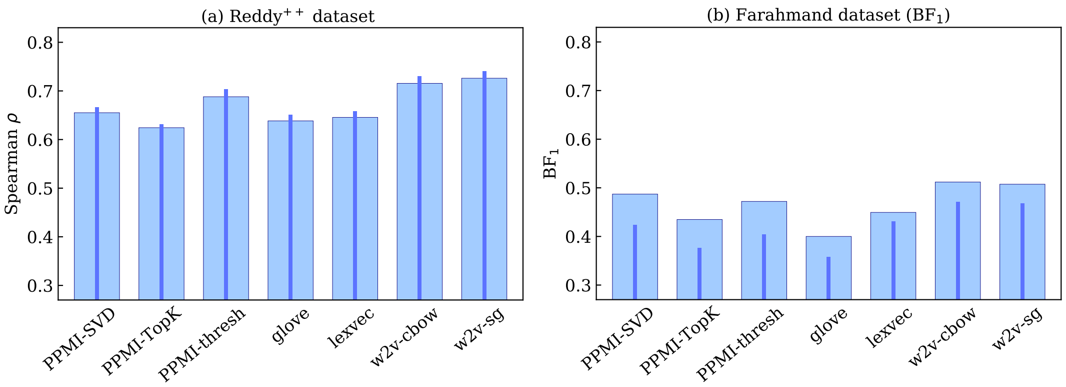
This first evaluation aims at verifying whether some DSMs, independently of their specific parameters, are more suitable for a given dataset/language than others. We perform a language-based analysis, evaluating the highest-scoring parameter combination of each DSM. All evaluations reported here use the *uniform* composition strategy (described in Section 4.2).

5.1.1 English

Figure 5.1 presents the best scores achieved under each DSM for the English datasets. The predictions for the *Reddy*$^{++}$ dataset were evaluated through Spearman ρ, while *Farahmand* predictions were evaluated with BF_1 (both described in Section 2.2).

Each of the wide bars in these graphs represents the highest score obtained from the set of 36 different configurations[1], using different combinations of WordForm, WindowSize, and Dimension. Similarly, each of the narrow inner bar represents the highest score among 36 configurations using fallback evaluation. While the fallback evaluation is responsible for slightly higher results in *Reddy*$^{++}$, its pessimistic approach is detrimental when evaluating the predictive model on the *Farahmand* dataset, which contains a considerable number of NCs that do not appear in the corpus. In both cases, these two types of evaluation produce similar rankings among the different DSMs, and we will henceforth focus on the highest results of strict evaluation (outer bars).

Figure 5.1: Overall highest results per DSM on English datasets.



For the *Reddy*$^{++}$ dataset, the highest results are found for the word-embedding models of *w2v*: the highest-Spearman *w2v–sg* model has a $\rho = .741$, while the best *w2v–cbow* has a $\rho = .730$. These results are followed by *PPMI–thresh*, with $\rho = .704$. Other models offer progressively inferior results: *PPMI–SVD* ($\rho = .666$), *lexvec* ($\rho = .658$), *glove* ($\rho = .651$) and *PPMI–TopK* ($\rho = .632$). We performed Wilcoxon's sign-rank test between all possible pairs of highest-Spearman configurations. The distributions of *w2v–cbow*, *w2v–sg* and *glove* were not deemed to be different from one another pairwise. Moreover, *glove* was not deemed different from *lexvec* or from *PPMI–SVD*. This is somewhat surprising, given the difference in scores between *glove* and the other DSMs (especially *w2v*). All other model pairs were deemed statistically different from each other ($p < 0.05$).

The *Reddy*$^{++}$ dataset combines all NCs from *Reddy* and *EN-comp*$_{90}$. If the *Reddy* dataset is considered by itself, we see the same trends as in *Reddy*$^{++}$. The overall best

[1] Only 12 configurations for *PPMI–TopK*, as the number of dimensions is fixed at 1000.

performance is found for *w2v–sg* (ρ = .812). The performance of the models *PPMI–thresh* (ρ = .803) and *w2v–cbow* (ρ = .796) closely follow the first place. If we isolate the *EN-comp*$_{90}$ NCs, a similar pattern emerges: the highest performance is achieved by *w2v–sg* (ρ = .669) and *w2v–cbow* (ρ = .665). Note that the highest results for *EN-comp*$_{90}$ are inferior to the ones obtained for *Reddy*. As we will see later, the best results for the French and Portuguese datasets are also in the same range as *EN-comp*$_{90}$. This difference in performance might be caused by the fact that these 3 datasets contain a higher amount of adjective+noun pairs than *Reddy*. As suggested in Section 3.2.2, humans had more difficultly judging the compositionality of adjectives than judging the compositionality of nouns. This could imply that noun+adjective scores are less reliable than noun+noun scores, and thus automatic methods should also obtain lower scores when predicting the compositionality of adjectives.

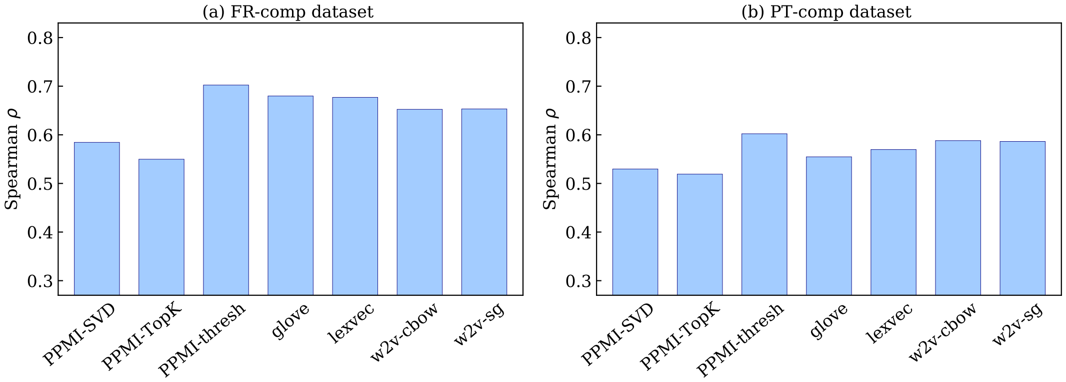
Similarly to the *Reddy*$^{++}$ dataset, evaluation on *Farahmand* yields the best results for the models *w2v–sg* (strict BF_1 = .498, fallback BF_1 = .455) and *w2v–cbow* (strict BF_1 = .501, fallback BF_1 = .471). The highest results are comparable to the BF_1 = .487 reported by Yazdani, Farahmand and Henderson (2015), while avoiding their use of functions whose parameters must be tuned through supervised learning for the prediction of compositionality.

5.1.2 French

As shown in Figure 5.2(a), overall *FR-comp* results in terms of Spearman correlation are reasonably different from *Reddy*$^{++}$ results. One of the most striking differences is the fact that the *w2v* models have lower quality. They are notably surpassed by *PPMI–thresh*, which rises to the first place with ρ = .702. This result is followed by word-embedding models: *glove* has ρ = .680 and *lexvec* has ρ = .677. Only then do we see the neural-network *w2v* models: *w2v–sg* (ρ = .653) and *w2v–cbow* (ρ = .652). Other PPMI-based models have a lower quality, dropping to the lowest ρ = .550 for *PPMI–TopK*.

As in the case of *Reddy*$^{++}$, we performed Wilcoxon's sign-rank test between all model pairs ($p < 0.05$), and both *w2v* models were deemed equivalent (i.e. we could not reject the hypothesis that they followed the same distribution). The *glove* model, however, was deemed different from both *w2v* configurations. The *PPMI–SVD* model was deemed equivalent to all other models except for *lexvec*. All other model pairs were deemed different from each other. Particularly in the case of the highest-Spearman DSM, *PPMI–thresh*, these results confirm that its best configuration is responsible for compositionality predictions that are statistically different form the predictions of other models.

Figure 5.2: Overall highest results per DSM on French and Portuguese datasets.



5.1.3 Portuguese

Figure 5.2(b) presents the overall highest Spearman correlations for the *PT-comp* dataset. As in the case of *FR-comp*, the *PPMI–thresh* model leads with the highest score ($\rho = .602$). This result is followed closely by word-embedding models: *w2v–cbow* has $\rho = .588$ and *w2v–sg* has $\rho = .586$. They are followed by *lexvec* with $\rho = .570$ and *glove* with $\rho = .555$. As with the French dataset, the other PPMI methods had the lowest scores: *PPMI–SVD* has $\rho = .530$ and *PPMI–TopK* has $\rho = .519$.

We performed Wilcoxon's sign-rank test between all model pairs ($p < 0.05$). Just like for *Reddy*$^{++}$, the distributions of *w2v–cbow*, *w2v–sg* and *glove* were deemed equivalent to each other pairwise. The *PPMI–SVD* and *glove* models were also deemed equivalent. Moreover, *PPMI–TopK* was deemed equivalent to all other models except *glove* and *PPMI–SVD*. Other model pairs were deemed different from each other ($p < 0.05$). As in the case of French results, *PPMI–thresh* had the highest scores, and Wilcoxon's test has confirmed that its predictions are statistically different form the predictions of other models.

5.1.4 Cross-language analysis

Table 5.1 presents, for each dataset, the Spearman correlation score of the highest-Spearman configuration for every DSM (in strict/fallback format for English datasets), with the top strict score highlighted in bold. In most of the cases, the best score obtained under fallback evaluation is comparable to the best strict score. Fallback results are considerably lower in the case of *Farahmand*. This reflects the fact that this dataset is not balanced with regards to compositionality: most of its NCs are compositional, leading to a higher average compositionality score that may not be suitable for the missing NCs, as these tend to be idiomatic.

Table 5.1: Highest Spearman ρ for all datasets (strict/fallback format for English datasets). The *Farahmand* dataset uses the highest BF_1 instead.

Dataset	*PPMI–SVD*	*PPMI–TopK*	*PPMI–thresh*	*glove*	*lexvec*	*w2v–cbow*	*w2v–sg*
FR-comp	.58	.55	**.70**	.68	.68	.65	.65
PT-comp	.53	.52	**.60**	.55	.57	.59	.59
EN-comp$_{90}$	.59/.60	.56/.56	.59/.60	.52/.54	.55/.57	.65/.67	**.65**/.67
Reddy^{++}	.66/.67	.62/.63	.69/.70	.64/.65	.65/.66	.72/.73	**.73**/.74
Reddy	.74/.74	.71/.72	.79/.80	.75/.76	.77/.77	.80/.80	**.81**/.81
Farahmand	.49/.42	.43/.38	.47/.40	.40/.36	.45/.43	.51/.47	**.51**/.47

Focusing on strict evaluation results, some interesting patterns can be observed in a comparison of the three languages. In all of the collected datasets, the predictions from *w2v–cbow* and *w2v–sg* follow the same distribution (as per a Wilcoxon sign-rank test), which is reflected in the fact that their scores are almost identical. Concerning the highest Spearman score out of all DSMs, the *w2v* models yield the best results for the two English datasets, while the highest scores for the two Romance languages were attained instead by *PPMI–thresh*. On the other side of this scale, the *PPMI–TopK* model is consistently ranked among the worst results. The best *PPMI–SVD* configuration presents a similar behavior, with consistently low results for all datasets but *Farahmand*.

5.2 DSM parameters

In this section, we investigate the hypothesis $\mathrm{h}_{\text{accur} \leftarrow \text{DSM}}$, which affirms that the accuracy of the model depends on DSM-specific parameters. We consider two parameters that can be independently tuned in every DSM: context-window size and number of dimensions in the output vectors.

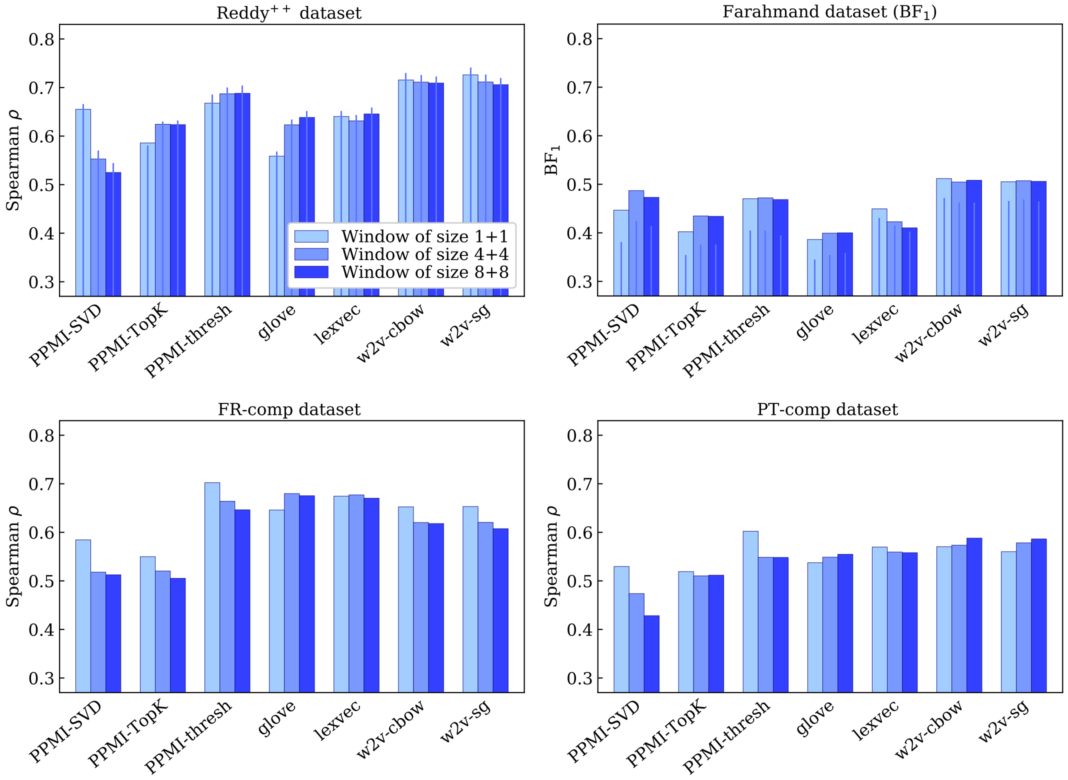
5.2.1 Context-window size

DSMs build the representation of every word based on the frequency of other words that appear in its context. A very simple way of defining such a context is through a window, whereby the context of a word with e.g. WindowSize=4+4 would consist in the previous four words and the following four words in the text. Most works in the literature construct DSMs with window sizes between the extremes 1+1 and 10+10, with a few works considering larger window sizes such as 16+16 or 20+20 (KIELA; CLARK, 2014; LAPESA; EVERT, 2014). We evaluate the behavior of compositionality prediction when the underlying DSM model is built with the commonly-used context-window sizes of 1+1, 4+4, and 8+8.[2] Our hypothesis ($\mathrm{h}_{\text{accur} \leftarrow \text{DSM.window}}$) is that the highest scores should be obtained by window sizes of 8+8, as the extra amount of data would lead to a better representation of the word-level semantics.

As can be seen in Figure 5.3, the performance with different context-window sizes is mostly DSM-dependent. In the case of *PPMI–SVD*, a window of size 1+1 yields better results for all datasets, with the exception of *Farahmand* evaluations. The *glove* model exhibits the opposite behavior: windows of size 1+1 are consistently worse than the windows of size 4+4 or 8+8. These results seem to be related to the manner with which the weights decay for different models. In the case of *PPMI–SVD* with WindowSize=8+8, a context word at distance d from its target word is weighted as $\frac{8-d}{8}$. In the case of *glove*, the decay happens much faster, with a weight of $\frac{8}{d}$, which allows the model to look farther away without being affected by the extra noise associated with the more distant contexts. For the other DSMs, window size is not a visible predictor of performance. In the exceptional case of *PPMI–thresh*, the results were language-dependent instead: French and Portuguese data can be better approximated through smaller windows, while English data

[2] Section 5.5.3 also performs some sanity checks for windows of size 2+2.

Figure 5.3: Best Spearman's ρ per DSM and WindowSize.



(including *Reddy* and *Reddy*$^{++}$) displays a weaker preference for larger window sizes. The appropriate choice of window size has been shown to be task-specific (LAPESA; EVERT, 2017), and the results above suggest that, in the task of compositionality prediction, even the choice of DSM may interact with this parameter.

5.2.2 Number of dimensions

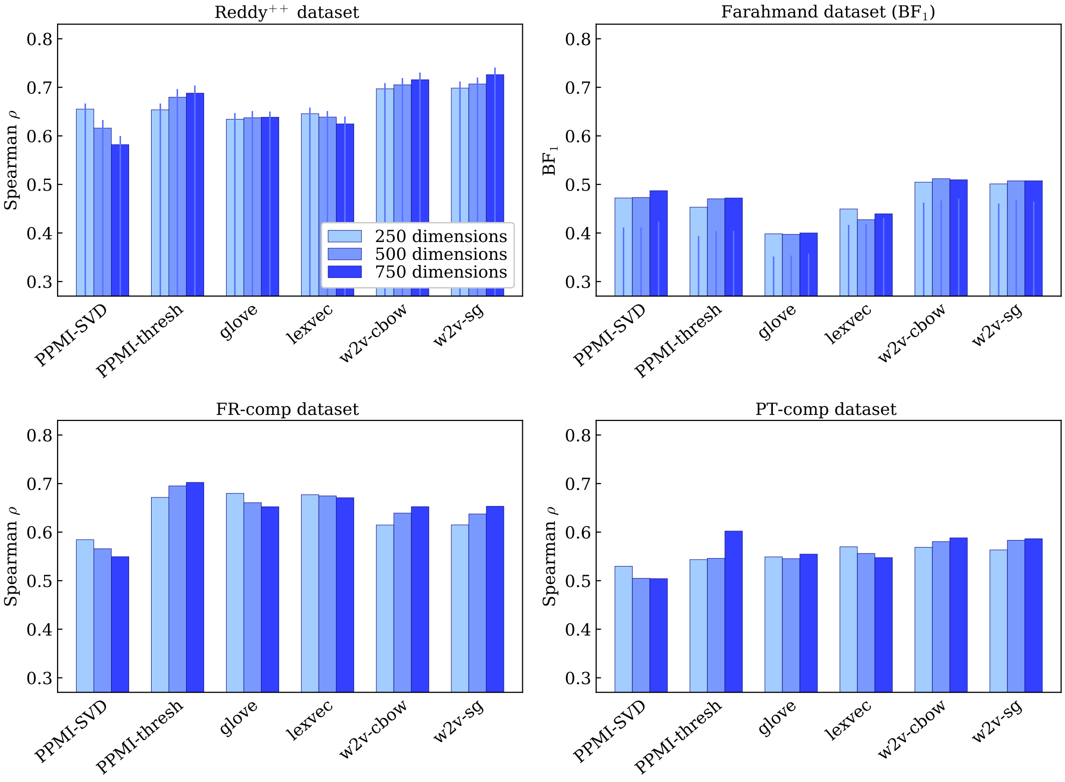
When instantiating a DSM, there is a trade-off in the number of vector dimensions. Models that have lower amount of dimensions will correspondingly have a smaller memory footprint[3], while models that have a larger number of dimensions eschew any memory concerns so as to be able to represent more fine-grained patterns of co-occurrence. The question is whether these extra dimensions can be put to good use by state-of-the-art DSMs. Most works in the literature build distributional models whose vectors contain between 200 and 900 dimensions (BARONI; DINU; KRUSZEWSKI, 2014; LAPESA; EVERT, 2014). We evaluate different DSMs by using Dimension=250, which approximates the common value used in the literature. We additionally present results for two of its

[3] Memory usage grows linearly with the number of dimensions.

multiples: 500 and 750 dimensions. Our hypothesis ($\text{h}_{\text{accur}\,\leftarrow\,\text{DSM.dims}}$) is that the highest scores should be obtained by DSMs where the vectors contain a higher number of dimensions.

As Figure 5.4 shows, for most DSMs, an increase in the number of dimensions causes a moderate increase in the quality of the predictive model, reflecting the additional information that can be used to perform the compositionality predictions. This is particularly true in the case of the DSMs that obtain the highest Spearman scores. This behavior is however inverted for *PPMI–SVD*, in which the highest results can be obtained by building lower-dimension models. Moreover, two DSMs seem to be particularly unaffected by the number of dimensions: *glove* and *lexvec* models seem to have around the same predictive power regardless of the number of dimensions. Overall, these results suggest that, across all DSMs, the best scores can be obtained by configurations involving a higher number of dimensions, as hypothesized.

Figure 5.4: Best Spearman's ρ per DSM and Dimension.



5.3 Corpus parameters

According to the hypothesis $\text{h}_{\text{accur} \leftarrow \text{corpus}}$, the accuracy of the semantic representation in a DSM is dependent on the quality of the input representation. In this section, we analyze the impact of different types of corpus preprocessing, corpus size variation, as well as the use of parallel sub-corpora in the prediction of NC compositionality.

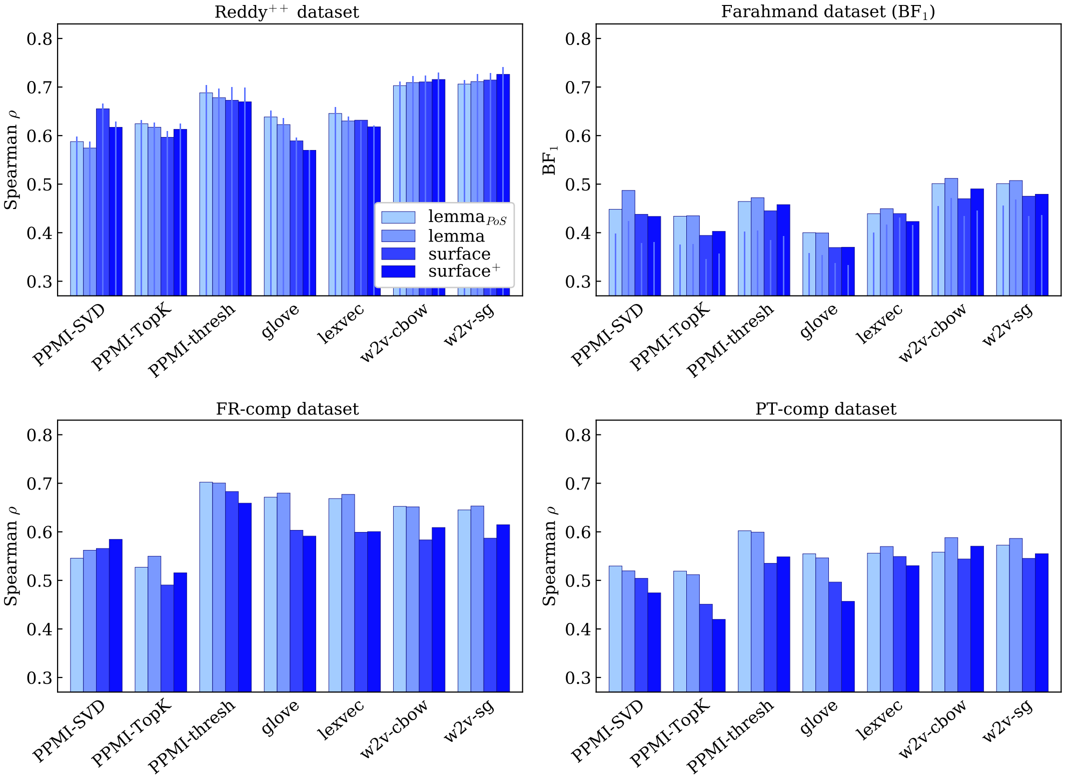
5.3.1 Type of preprocessing

Most works in the literature on distributional models focus on the analysis and tuning of different statistical parameters, reducing the preprocessing to tokenization, along with simple procedures such as lower-casing and rare word removal (MIKOLOV et al., 2013; PENNINGTON; SOCHER; MANNING, 2014; LEVY; GOLDBERG; DAGAN, 2015; SALLE; VILLAVICENCIO; IDIART, 2016). Works that rely on DSMs in semantic tasks tend to consider other preprocessing techniques, such as lemmatization, stemming, POS tagging, and stopword removal (BULLINARIA; LEVY, 2012; KIELA; CLARK, 2014). The goal of such procedures is to increase the quality of the word representations, by conflating different uses of the same word into a single canonical form, by allowing the disambiguation of homonyms based on their syntactic function, and through the elimination of random noise from the corpora.

As described in Section 4.3, we consider four different levels of corpus preprocessing: WordForm=*surface*$^{+}$, *surface*, *lemma*$_{PoS}$ and *lemma*. Under each of these configurations, there is a difference in how much information is condensed into each token in the corpus, with *surface*$^{+}$ being the most specific (every word in the corpus is considered verbatim) and *lemma* being the most general (where only the lemmas of content words are used in the representation of each token). Our hypothesis ($\text{h}_{\text{accur} \leftarrow \text{corpus.wordform}}$) is that the less specific configurations present a less sparse view of the data, contributing to higher-quality DSM representations and thus achieving higher scores.

Figure 5.5 presents the impact of different types of corpus preprocessing on the quality of the compositionality prediction model. The results seem to be language-dependent: The results for the English-language datasets are quite heterogeneous, while for the other two languages, the lemma-based word representations consistently allow a better prediction of compositionality scores. This phenomenon may be explained by the fact that French and Portuguese are morphologically richer than English. For the former languages, lemma-based representations reduce the sparsity in the data and allow more information to be gathered from the same amount of data. In the case of English, lemmatization has a reduced effect, and in particular for *PPMI–SVD*, it visibly reduces the quality of predictions.

Figure 5.5: Best Spearman's ρ per DSM and WordForm



The results for *lemma$_{PoS}$* represent the addition of POS tags to every word in the corpus. This extra information does not show any improvement over *lemma*. This suggests that words that share the same lemma are semantically close enough that any gains from disambiguation are compensated by the sparsity of a higher vocabulary size.

The results obtained with *surface*$^+$ are surprisingly similar to *surface*, which confirms previous suggestions that stopword removal does not significantly affect the data (BULLINARIA; LEVY, 2012). These results are achieved even though *surface*$^+$ contains stopwords, which one might expect would dilute the DSM representation (due to their low level of association with most other words), which might then reduce the accuracy of predictions. A possible explanation could be that the stopwords in *surface*$^+$ effectively contribute to a reduced window size of content words[4], which is shown in Section 5.2.1 to

[4] This could be investigated in future work with a WordForm that only uses the content words inside a content+stopword window.

consistently yield better results. Indeed, when looking at the highest *surface*$^+$ scores across all models and datasets, the majority of the configurations involve WindowSize=1+1, further highlighting the role of the context-window size in the best-performing models.

5.3.2 Corpus size

In this thesis, we have presented results for compositionality prediction based on three similarly-sized corpora. The general intuition gathered from the literature is that predictions based on larger corpora should obtain higher scores, as rarer contexts would also be taken into account, thus improving the quality of the vector representation of these words. We express this intuition through the hypothesis $\mathrm{h}_{\mathrm{accur} \leftarrow \mathrm{corpus.size}}$, which predicts higher scores for DSMs instantiated for larger-sized corpora.

This section performs a quantitative analysis of the impact of different corpus sizes on the quality of the predictions. For each of the *Reddy*, *FR-comp* and *PT-comp* datasets, we consider the highest-Spearman *PPMI–thresh* and *w2v–sg* configuration obtained thus far for the full-size corpus.[5] We then build new models under the same configuration, but using corpus fragments of size varying from 1% to 100% of the whole corpus, increasing by steps of 1/100 at a time.

Figure 5.6(a) presents three graphs of the scores obtained by building models for the best *PPMI–thresh* configuration of each dataset. The 100 positions in the x-axis correspond to the corpus sizes (1% to 100%). Eight different samplings of corpus fragments were performed (for a total of 800 models per language), with each y-axis data point presenting the average of the 8 Spearman scores obtained from those samplings. Each data point also presents the sample standard deviation for those 8 executions. Points to the left of the vertical bar have at least one sampling with missing compounds, while points to the right have 100% of the compounds in all 8 samplings. The results suggest that, for the three languages, a corpus size of around 800 million to 1 billion tokens (40% of the whole corpus size) is large enough to obtain the best results, with further increases in the amount of data available only contributing marginally to the overall quality of the predictions.

[5] Results for corpus size do not consider variations in the configurations, as the best configuration for the full-size corpus is used for every corpus size.

Figure 5.6: Spearman's ρ for different corpus sizes, running *PPMI–thresh* (left) and *w2v–sg* (right).

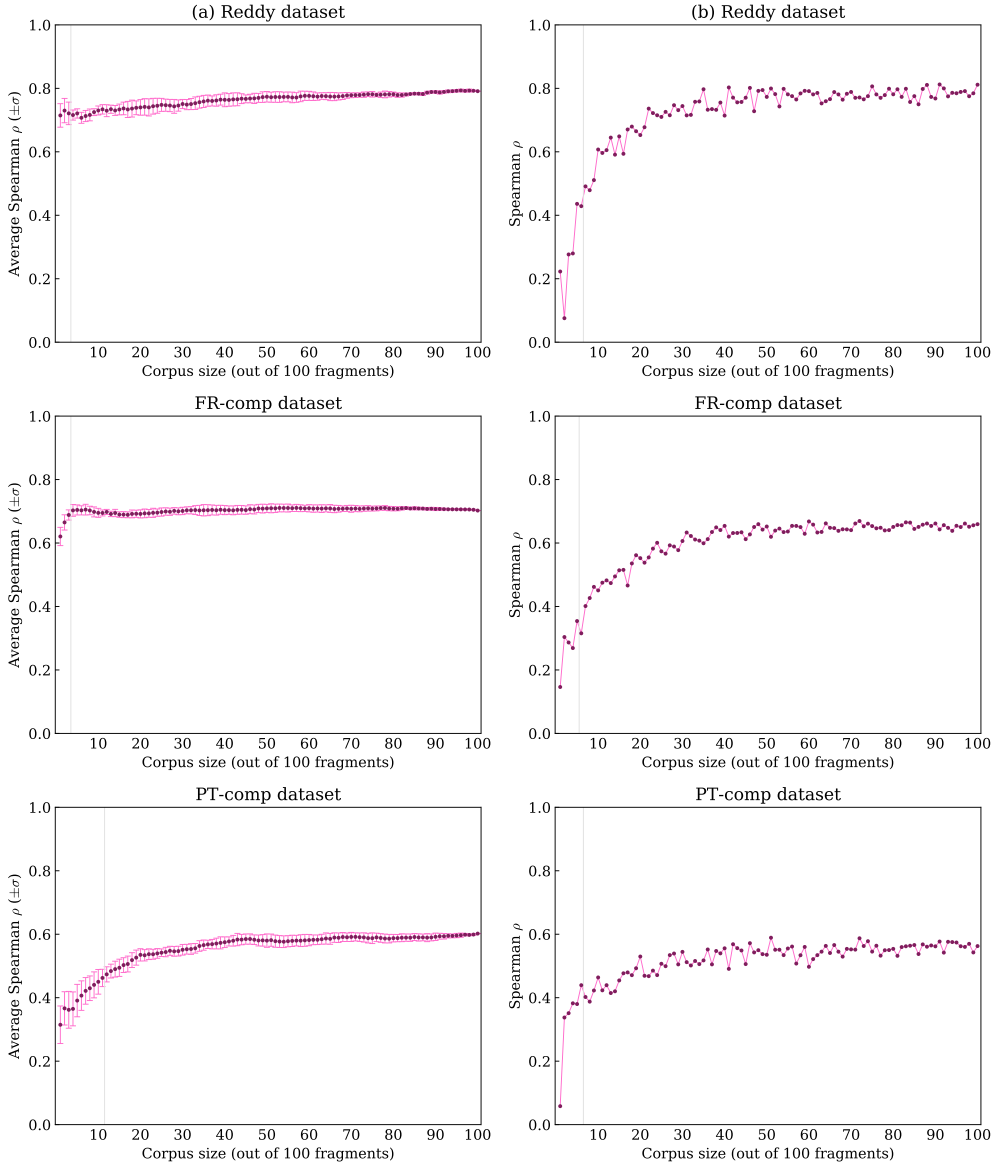


Figure 5.6(b) presents three graphs with the scores obtained by the best *w2v–sg* configuration for each dataset. Due to the fact that *w2v–sg* is much more time-consuming than *PPMI–thresh*, a single sampling was used, and thus only one execution was performed for each datapoint (for a total of 100 vector models per language). Similarly to *PPMI–thresh*, a corpus fragment of around 40% size (800 million to 1 billion tokens) was already large enough for the results to stabilize close to the score obtained by the fragment of size 100%. This suggests that corpus size does strongly affect the quality of the underlying DSM representation, but that it reaches a plateau around a billion tokens. Even higher corpus sizes would presumably only offer a minor improvement in DSM representation quality. Future work should investigate whether a similar plateau

can also be observed for other methods of compositionality prediction.

5.3.3 Parallel predictions

One idea that has been employed in the literature is that of *ensemble methods*, in which the predictions from multiple methods (e.g. multiple DSM instances built from the same data) are combined into a single prediction that outperforms all other methods (ZHOU, 2012). We leave ensemble experiments for future work, but we focus on an approach that is inspired by its success: *parallel predictions*.

The results obtained for different corpus sizes above imply that a subset of the corpus can lead to results that are equivalent to the ones obtained for the whole corpus. In fact, even when a smaller fraction of the corpus is considered, such as 20% of the total corpus size, the resulting model can still yield reasonably good predictions of compositionality. We hypothesize ($\mathrm{h}_{\mathrm{accur} \leftarrow \mathrm{corpus.parallel}}$) that, just as an ensemble of *methods* instantiated from a single corpus may complement each other and achieve higher scores, so can a single DSM method instantiated multiple times in parallel from an ensemble of *corpora* be combined so as to achieve accurate DSM representations.

We thus propose a technique in which the whole corpus is divided in M parallel fragments (c_1, c_2, $\ldots c_M$). We then instantiate the same DSM M times in parallel, each one based on a different corpus fragment. These DSMs can then be used to produce a set of M compositionality predictions per compound, using the *uniform* strategy ($\mathrm{CS}_{\beta(1)}$, $\mathrm{CS}_{\beta(2)}$, $\ldots \mathrm{CS}_{\beta(M)}$). The M parallel predictions for each compound are then combined through the arithmetic average into a single compositionality score CS_{P}. We can then evaluate the set of CS_{P} for every compound by comparing them with the reference dataset.

In order to verify the hypothesis that the parallel predictions can yield results that are comparable to the ones obtained on the whole corpus, we ran an experiment in which the whole corpus was divided in $M = 5$ fragments of equal size, each corresponding to 20% of the whole corpus. Table 5.2 presents the Spearman scores of whole-corpus ($\rho_{100\%}$) and parallel prediction ($\rho_{5\times20\%}$) for two models: *PPMI–thresh* and *w2v–sg*. For the latter, we have considered smaller subsampling sizes rates of 10^{-3} and 10^{-4}, to account for the smaller corpus sizes. The results suggest that the parallel prediction on smaller corpus fragments can be as effective as a single prediction generated from the whole corpus.

Table 5.2: Results for whole-corpus and parallel predictions.

Model (*Reddy*$^{++}$)	$\rho_{100\%}$	$\rho_{5\times20\%}$	Difference (%)	Worst $\rho_{20\%}$	Best $\rho_{20\%}$
PPMI–thresh	**.699**	.680	(−1.9)	.626	.678
w2v–sg (10^{-3})	**.731**	.719	(−1.2)	.668	.708
w2v–sg (10^{-4})	**.731**	.717	(−1.4)	.667	.707
Model (*FR-comp*)	$\rho_{100\%}$	$\rho_{5\times20\%}$	Difference (%)	Worst $\rho_{20\%}$	Best $\rho_{20\%}$
PPMI–thresh	.702	**.709**	(+0.7)	.686	.714
w2v–sg (10^{-3})	.672	**.685**	(+1.3)	.654	.688
w2v–sg (10^{-4})	.672	**.688**	(+1.6)	.671	.693
Model (*PT-comp*)	$\rho_{100\%}$	$\rho_{5\times20\%}$	Difference (%)	Worst $\rho_{20\%}$	Best $\rho_{20\%}$
PPMI–thresh	**.602**	.572	(−3.1)	.496	.549
w2v–sg (10^{-3})	**.586**	.581	(−0.5)	.528	.569
w2v–sg (10^{-4})	**.586**	.581	(−0.6)	.520	.566

Table 5.2 also indicates the Spearman scores obtained for the highest and lowest-ranking corpus fragments (Worst $\rho_{20\%}$ and Best $\rho_{20\%}$, each using only 20% of the corpus). In the case of *Reddy*$^{++}$ and *PT-comp*, in all configurations, we can see that the average score obtained from the 5 fragments ($\rho_{5\times20\%}$) is slightly *higher* than the score obtained by the best fragment. This suggests that the technique of parallel predictions is actually able to combine the results from the different fragments, confirming the underlying hypothesis. We leave it for future work the investigation of different ways of combining the parallel predictions into a single score.

One of the greatest advantages of parallel prediction is its potential for scalability. While a standard DSM-based predictive model requires the whole corpus to be processed at once, a parallel model allows the computation of such predictions in a distributed fashion. This reduces the total execution time, allows the better utilization of distributed resources (such as computer clusters), and bypasses memory limitations of a single machine.

5.4 Prediction strategy

Now that we have evaluated the impact of DSM and corpus parameters on the predicted compositionality scores, we turn to the underlying prediction strategy itself. As we have seen in Chapter 3, the elements of an MWE may vary in terms of the semantic contribution of each element to the MWE as a whole, and this may have an impact on the success of the composition model adopted for deriving the vector space representation of the MWE (hypothesis $\mathrm{h_{strat}}$). For instance, adopting a uniform (50%:50%) composition for the elements of a compound might not accurately capture a faithful representation of compounds whose meaning is more semantically related to one of the components than to the other (as in the case of the compound *crocodile tears*, regarding its head, and *night owl*, regarding its modifier).

We compare below six different compositionality prediction strategies (all described in Section 4.2). Some of these strategies consider different variations of weights on the compound elements themselves: *uniform* uses a 50%:50% scheme, while two other strategies (*head* and *mod*) use a 0%:100% scheme We also evaluate a proposed new model of additive composition, *maxsim*, which dynamically determines weights so as to assign an optimal proximity of the compound to each of its single-word elements. Additionally, we consider the *arith* and *geom* prediction strategies, in which the representation of the compound is independently compared to the representation of each component, and with the resulting score being the (arithmetic or geometric) mean of the comparison scores.

Table 5.3 presents the scores obtained for all strategies on the configurations in which *uniform* obtains its highest scores (i.e. using the best configuration for each DSM as reported up to now). In most of the cases, the score obtained for the *uniform* prediction is higher than both the *head* and *mod* scores when taken separately, which is in line with the hypothesis $\mathrm{h}_{\text{strat.partial-info}}$ that these two strategies are somewhat limited due to the fact that they only consider half of the available distributional information.[6] However, this difference is only moderate, with *mod* predictions in particular attaining results that are quite close to *uniform*.

Results for other strategies are slightly worse than *uniform*, suggesting that these approaches are either subpar, or that they improve in configurations that differ from the best *uniform* configurations.

Table 5.3: Spearman scores for best *uniform* model of each dataset, using different prediction strategies. The *Farahmand* dataset uses BF_1.

Dataset	DSM	configuration	*uniform*	*maxsim*	*geom*	*arith*	*head*	*mod*
Reddy	*w2v–sg*	*surface*.w_1.d_{750}	**.812**	.802	.756	.805	.635	.752
EN-comp$_{90}$	*w2v–cbow*	*lemma*.w_4.d_{500}	**.653**	.651	.600	.647	.463	.613
Reddy^{++}	*w2v–sg*	*surface*$^{+}$.w_1.d_{750}	.726	**.730**	.657	.718	.524	.677
FR-comp	*PPMI–thresh*	*lemma*$_{PoS}$.w_1.d_{750}	**.702**	.688	.668	.698	.605	.603
PT-comp	*PPMI–thresh*	*lemma*$_{PoS}$.w_1.d_{750}	**.602**	.577	.524	.595	.524	.413
Farahmand	*w2v–cbow*	*lemma*$_{PoS}$.w_8.d_{250}	.501	.484	.523	.517	.402	**.534**

[6] Future work should investigate why the BF_1 scores of *Farahmand* are higher for *mod* than for all other strategies.

To verify whether the other prediction strategies improve models that differ from the highest-Spearman *uniform* configurations, we have evaluated every strategy on all DSM instances. Table 5.4 presents a summary of the highest scores obtainable for each prediction strategy individually (i.e. each score represents the best configuration for a given strategy evaluated on a given dataset). Every best score is statistically different from all other scores in its row ($p < 0.05$). Similarly to the above results, the score obtained for the *uniform* prediction is higher than the one obtained for both *head* and *mod* strategies (hypothesis $\text{h}_{\text{strat.partial-info}}$), which further suggests that the quality of *uniform* predictions is derived from the combination of the vector representation of the two words (in particular from *mod*).

The *arith* strategy obtains performance results that are very similar to the ones of *uniform*, reflecting the fact that both methods rely on an additive model of composition. Indeed, if we consider the average (across the 7 DSMs) of the Pearson correlation between the 180 predictions for the highest-Spearman configuration of *arith* and *uniform*, we obtain $r = .972$ for *Reddy*$^{++}$, .991 for *FR-comp* and .969 for *PT-comp*, confirming that these models produce very similar predictions. Moreover, these two strategies behave very similarly when we consider the Spearman scores obtained for the 228 DSM instances in each language: if we consider the Pearson correlation between the 228 pairs of Spearman scores associated with both strategies, we obtain $r = .981$ for *Reddy*$^{++}$, .985 for *FR-comp* and .944 *PT-comp*.

Table 5.4: Highest Spearman score for each prediction strategy individually. *Farahmand* uses BF_1.

Dataset	*uniform*	*maxsim*	*geom*	*arith*	*head*	*mod*
Reddy	.812	**.814**	.797	.805	.654	.776
EN-comp$_{90}$	.653	**.659**	.600	.647	.483	.615
Reddy^{++}	.726	**.730**	.677	.718	.555	.677
FR-comp	.702	.693	.699	**.703**	.617	.645
PT-comp	**.602**	.590	.580	.598	.558	.486
Farahmand	.501	.487	**.529**	.518	.422	.528

When one considers the highest-Spearman configuration for each strategy, results for *maxsim* are competitive with the results for *uniform*. While *maxsim* fares slightly better on English continuous-score datasets, *uniform* obtains slightly higher scores on the other two languages. Implicit to the calculation of *maxsim* is the assignment of weights for the components of every NC, and we have considered whether the assigned weights actually differ from the 50:50 assignment of *uniform*. This does seem to be the case, as some NCs have a weight prediction that is much closer to human head-only and modifier-only scores than a 50:50 prediction. For example, in the *w2v–sg* prediction for *Reddy*$^{++}$, some NCs were weighted more heavily in favor of the head (e.g. *silver screen* had weight 11:89), while others had more weight in the modifier (e.g. *spelling bee* with weight 94:06), and with many intermediary NCs in between (e.g. *dirty word* with weight 47:53).

One of the side effects of calculating these weights is that they also reveal any bias in the semantic influence of the head or the modifier of each compound, and we consider whether this bias may be affecting the results on each dataset. Table 5.5 presents the highest-Spearman *maxsim* model for each dataset, along with the average of the weights assigned to head and modifier for every NC in the dataset. The results are extremely stable: while the weights that optimize for compositionality are fairly similar for the English datasets, they are highly discrepant for both *FR-comp* and *PT-comp*, in which the weight of the head is disproportionately higher than the weight of the modifier.

The fact that *FR-comp* and *PT-comp* compounds have the potential for higher compositional interpretation in the head than in the modifier could be elucidated by the consideration that all of the modifiers in these datasets are adjectives, while English-language modifiers may also be nouns. Therefore, the contribution of adjectives to the overall meaning could be lower due to some linguistic phenomenon. For example, some of the adjectives used in these compounds are highly polysemous, and could be seen contributing to some specific meaning is not found on isolated occurrences of the adjective itself (e.g. `FR` *beau* (lit. *beautiful*) is used in the translation of most *in-law* family members, such as *beau-frère* 'brother-in-law' (lit. *beautiful-brother*)).

Table 5.5: Average weight of highest-Spearman *maxsim* model for each dataset.

Dataset	DSM	configuration	maxsim	weight$_{\text{head}}$	weight$_{\text{mod}}$
Reddy	*w2v–sg*	*surface*$^{+}$.w$_1$.d$_{750}$	.814	53	47
EN-comp$_{90}$	*w2v–cbow*	*lemma*.w$_8$.d$_{750}$	.659	54	46
Reddy^{++}	*w2v–sg*	*surface*$^{+}$.w$_1$.d$_{750}$	.730	55	45
FR-comp	*PPMI–thresh*	*lemma*.w$_1$.d$_{750}$	.693	68	32
PT-comp	*w2v–sg*	*lemma*.w$_8$.d$_{750}$	.590	68	32

Up to this point, it is still unclear whether it is true that the *maxsim* strategy is able to more aptly capture the semantics of compositional MWEs (hypothesis h$_{\text{strat.maxsim}}$). In order to better understand the behavior of *maxsim* with regards to *uniform*, we rank the compounds in *ALL-comp* according to three possible sets of scores: (a) the compositionality score assigned by human annotators; (b) the highest-Spearman *maxsim* prediction; and (c) the highest-Spearman *uniform* prediction. Each compound is then assigned three corresponding ranks (positive integers): rk_{human}, rk_{maxsim}, $\text{rk}_{uniform}$. We then calculate the improvement score of each compound as:

$$\text{improv}_{maxsim} = |\text{rk}_{uniform} - \text{rk}_{human}| - |\text{rk}_{maxsim} - \text{rk}_{human}|.$$

Figure 5.7 presents the distribution of rank improvement scores for the highest-scoring *PPMI–thresh* and *w2v–sg* configurations.[7] Each graph presents the improvement score for NCs from the three languages, ranked according to rk_{human}. It can be seen that, for most NCs, there is only a light variation in the rank compositionality of *maxsim*. For the NCs that have a more drastic variation in rank, positive improvements are associated with higher human-ranked compositionality (right side of the graph), while negative improvement scores are associated with idiomatic NCs. This confirms the hypothesis that *maxsim* can better capture the semantics of compositional MWEs, albeit this only applies to some outlier cases.

Figure 5.7: Distribution of improv_{maxsim} as a function of human judgments.

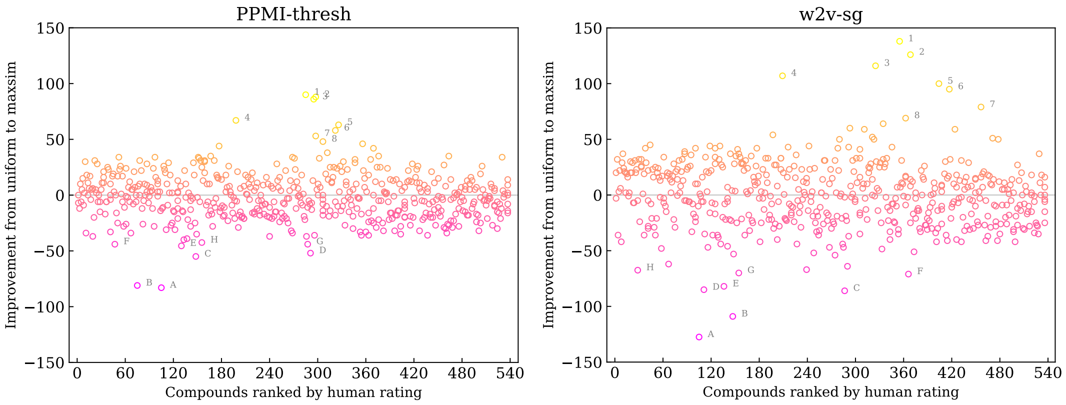
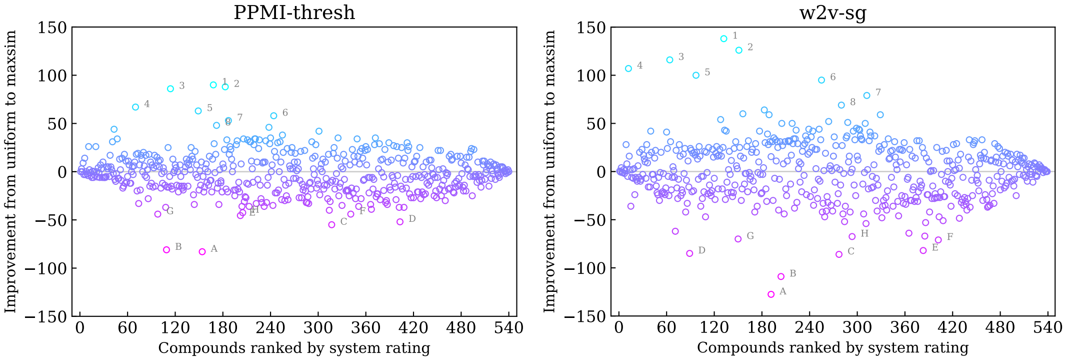


Figure 5.8 presents the distribution of rank improvements for all NCs, ranked according to $\text{rk}_{uniform}$ instead. Differently from above, NCs with the highest variation in rank are found on the left side of the graph, indicating that they were all initially judged as idiomatic. This indicates that *maxsim* tends to improve the score of NCs that humans considered more compositional, but that the *uniform* system considered more idiomatic. On the other hand, NCs that are correctly classified as idiomatic by the *uniform* prediction

[7] We focus on one representative of PPMI-based DSMs and one representative of word-embedding ones. Similar results were observed for the highest-Spearman configuration of other DSMs.

are somewhat under-estimated by *maxsim*. The positive and negative improvements are somewhat balanced, which explains why *maxsim* predictions fare as well as *uniform*.

Figure 5.8: Distribution of improv_{maxsim} as a function of *uniform* scores.



Figures 5.7 and 5.8 also indicate the outlier NCs with highest most improvement (numbers from 1 to 8), as well as the NCs with lowest improvement scores (letters from A to H). Table 5.6 presents these outlier NCs along with their improvement scores (see the Appendices A., B. and C. for the translation, glosses and human scores associated with these compounds). We can see that there is a disproportionate amount of outlier NCs for Portuguese and French (particularly the former), suggesting that *maxsim* has a stronger impact on those languages than on English. It is also noticeable that some NCs had a similar improvement score under both DSMs, with e.g. high improvement for PT *caixa forte* and low improvement scores for PT *coração partido*. It is further remarkable that equivalent NCs in different languages are similarly impacted by *maxsim*, as in the case of PT *caixa forte* and FR *coffre fort*. Nevertheless, *maxsim* does not present a considerable overall impact on the rank of the predictions, obtaining an average improvement of $\overline{\text{improv}_{maxsim}} = +0.41$.

As in the case of *maxsim*, we also consider the rank improvement of *geom* predictions over *uniform*. We rank the compounds in *ALL-comp* according to three possible sets of scores: (a) the compositionality score assigned by human annotators; (b) the highest-Spearman *geom* prediction; and (c) the highest-Spearman *uniform* prediction. Each compound is then assigned three corresponding ranks (positive integers): rk_{human}, rk_{geom}, $\text{rk}_{uniform}$. We then calculate the improvement score of each compound as:

$$\text{improv}_{geom} = |\text{rk}_{uniform} - \text{rk}_{human}| - |\text{rk}_{geom} - \text{rk}_{human}|.$$

Table 5.6: Outliers regarding positive/negative *maxsim* improvement.

ID	improv	*PPMI–thresh*	improv	*w2v–sg*
1	(+90)	`FR` *premier plan*	(+138)	`PT` *cerca viva*
2	(+88)	`FR` *matière première*	(+126)	`FR` *coffre fort*
3	(+86)	`PT` *amigo oculto*	(+116)	`PT` *caixa forte*
4	(+67)	`FR` *première dame*	(+107)	`PT` *golpe baixo*
5	(+63)	`PT` *caixa forte*	(+100)	`PT` *primeira necessidade*
6	(+58)	`PT` *prato feito*	(+95)	`EN` *role model*
7	(+53)	`FR` *idée reçue*	(+79)	`FR` *bonne pratique*
8	(+48)	`FR` *marée noire*	(+69)	`PT` *carta aberta*
H	(−42)	`PT` *alta costura*	(−68)	`FR` *bras droit*
G	(−44)	`EN` *half sister*	(−70)	`PT` *alta costura*
F	(−44)	`EN` *melting pot*	(−71)	`PT` *carne vermelha*
E	(−46)	`FR` *berger allemand*	(−82)	`PT` *alto mar*
D	(−52)	`PT` *mar aberto*	(−85)	`PT` *mesa redonda*
C	(−55)	`PT` *febre amarela*	(−86)	`EN` *half sister*
B	(−81)	`PT` *livro aberto*	(−109)	`PT` *febre amarela*
A	(−83)	`PT` *coração partido*	(−128)	`PT` *coração partido*

The hypothesis ($\text{h}_{\text{strat.geom}}$) is that *geom* should more accurately represent the semantics of idiomatic NCs, and this would be reflected in the improvement scores. Figure 5.9 presents the distribution of rank improvements for NCs in *ALL-comp*, ranked according to rk_{human}. It can be seen that, for most NCs, there is only a slight variation in the rank compositionality of *geom.* For the NCs that have a more drastic variation in rank, positive improvements are slightly more associated with lower human-ranked compositionality (left side of the graph), while negative improvement scores are visibly associated with compositional NCs (right side of the graph). This is the opposite of what was observed for *maxsim*, and confirms the interpretation that these models optimize for opposite extremes of compositionality (with *geom* focusing on idiomatic NCs at the expense of more compositional ones). As in the case of *maxsim*, this behavior is only observed for the outlier cases.

Figure 5.10 presents the distribution of rank improvements for all NCs in the highest-Spearman configuration, ranked according to $\text{rk}_{uniform}$ instead. Here again, the behavior of the *geom* strategy is the opposite of what was observed for *maxsim*: NCs with the highest variation in rank are found on the right side of the graph, indicating that they were all initially judged as compositional. This indicates that *geom* tends to improve the score of NCs that humans considered more idiomatic, but that the *uniform* system considered more compositional. On the other hand, NCs that are correctly classified as compositional by the *uniform* prediction are somewhat pessimized by *geom.*

Figure 5.9: Distribution of improv$_{geom}$ as a function of human judgments.

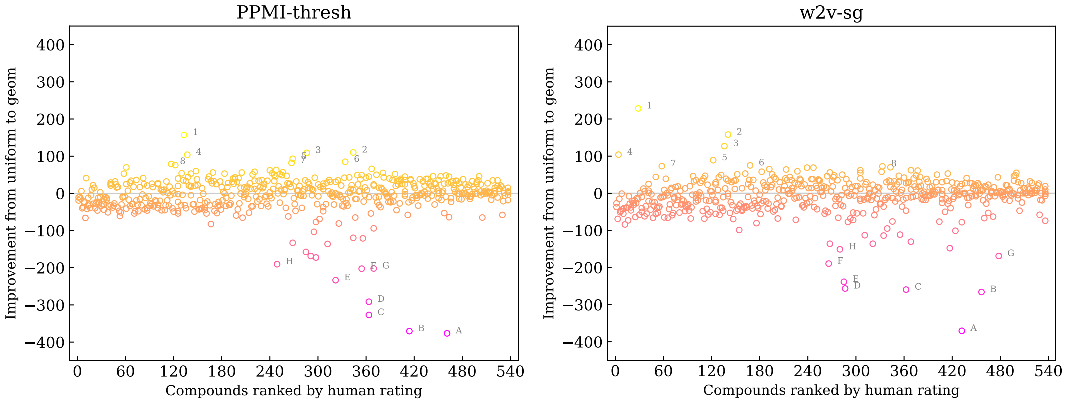
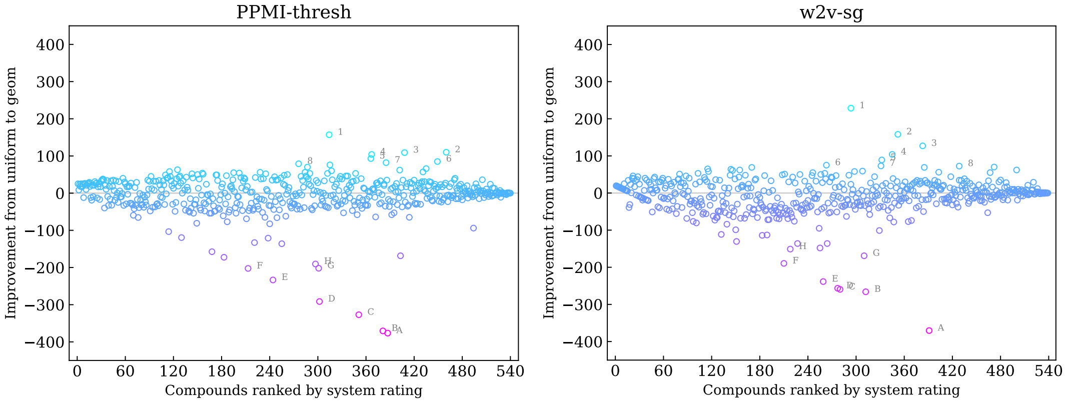


Figure 5.10: Distribution of improv$_{geom}$ as a function of *uniform* scores.



Figures 5.9 and 5.10 also indicate the outlier NCs with the highest and lowest improvement scores (numbers and letters, respectively). Table 5.7 presents these outlier NCs along with their improvement scores. As in the case of *maxsim*, the majority of the outliers belong to the Portuguese dataset. Some of the NCs that were found as outliers in *maxsim* re-appear as outliers for *geom* with inverted polarity in the improvement score, e.g. FR *bras droit* scores predicted by *PPMI–thresh* ($\text{improv}_{maxsim} = +58$, $\text{improv}_{geom} = -234$) and PT *prato feito* as predicted by *w2v–sg* ($\text{improv}_{maxsim} = -68$, $\text{improv}_{geom} = +228$). This suggests that future work should consider combining both approaches into a single prediction strategy that decides which sub-strategy to use as a function of the *uniform* prediction for each NC. As it stands, however, the *geom* strategy has a mild negative influence on the rank of the predictions, obtaining an average improvement score of $\overline{\text{improv}_{geom}} = -7.87$.

Table 5.7: Outliers regarding positive/negative *geom* improvement.

ID	improv	*PPMI–thresh*	improv	*w2v–sg*
1	(+157)	EN *snail mail*	(+228)	FR *bras droit*
2	(+110)	FR *guerre civile*	(+158)	PT *lua nova*
3	(+109)	FR *disque dur*	(+127)	PT *alto mar*
4	(+104)	PT *alto mar*	(+104)	PT *pé direito*
5	(+93)	PT *ônibus executivo*	(+89)	EN *carpet bombing*
6	(+85)	EN *search engine*	(+75)	PT *lista negra*
7	(+82)	PT *carro forte*	(+73)	PT *arma branca*
8	(+79)	EN *noble gas*	(+72)	EN *search engine*
H	(−190)	PT *ar condicionado*	(−151)	PT *disco rígido*
G	(−202)	FR *coffre fort*	(−169)	EN *subway system*
F	(−202)	FR *bon sens*	(−190)	PT *carro forte*
E	(−234)	PT *prato feito*	(−238)	FR *disque dur*
D	(−292)	FR *baie vitrée*	(−256)	EN *half sister*
C	(−327)	PT *carta aberta*	(−260)	PT *carta aberta*
B	(−370)	PT *vinho tinto*	(−266)	FR *bonne pratique*
A	(−376)	PT *circuito integrado*	(−370)	EN *end user*

5.5 Sanity checks

The number of possible DSM configurations grows exponentially with the number of internal variables in a DSM, forestalling the possibility of an exhaustive search for every possible parameter. We have evaluated above the set of variables that are most often manually tuned in the literature, but a reasonable question would be whether these results can be further improved through the modification of some other often-ignored model-specific parameters. We thus perform some sanity checks through a local search of such parameters around the highest-Spearman configuration of each DSM.

Section 5.5.1 evaluates the number of DSM iterations. Section 5.5.2 evaluates the minimum word-count threshold in the DSM. Section 5.5.3 considers a WindowSize=2+2. Section 5.5.4 considers higher numbers of DSM vector dimensions. Section 5.5.5 evaluates the non-determinism of DSMs through multiple random initializations. Finally, Section 5.5.6 considers whether the filtering of dataset annotations could improve its quality as well as the accuracy of predictions.

5.5.1 Number of iterations

Some of the DSMs in consideration on this chapter are iterative: they re-read and re-process the same corpus multiple times. For those DSMs, we present the results of running their best configuration, but using a higher number of iterations. This higher number of iterations is inspired by the models found in parts of the literature, where e.g. the number of *glove* iterations can be as high as 50 (SALLE; VILLAVICENCIO; IDIART, 2016) or even 100 (PENNINGTON; SOCHER; MANNING, 2014). The intuition is that most models will lose some information (due to their probabilistic sampling), which could be regained at the cost of a higher number of iterations.

Table 5.8 presents a comparison between the baseline ρ for 15 iterations and the ρ obtained when 100 iterations are performed. For all DSMs, we see that the increase in the number of iterations does not improve the quality of the vectors, with the relatively small number of 15 iterations yielding better results. This may suggest that a small number of iterations can already sample enough distributional information, with further iterations accruing additional noise from low-frequency words. The extra number of iterations could also be responsible for overfitting of the DSM to represent particularities of the corpus, which would reduce the quality of the underlying vectors. Given the extra cost of running more iterations[8], we refrain from building further models with as many iterations in this thesis.

Table 5.8: Results using a higher number of iterations.

Model (*FR-comp*)	ρ_{base}	$\rho_{\text{iter=100}}$	Difference (%)
w2v–cbow	**.660**	.640	(−2.0)
w2v–sg	**.672**	.636	(−3.7)
glove	**.680**	.677	(−0.3)
lexvec	**.677**	.671	(−0.6)
Model (*Reddy*)	ρ_{base}	$\rho_{\text{iter=100}}$	Difference (%)
w2v–cbow	**.809**	.766	(−4.3)
w2v–sg	**.821**	.777	(−4.4)
glove	**.764**	.746	(−1.8)
lexvec	**.774**	.757	(−1.7)
Model (*PT-comp*)	ρ_{base}	$\rho_{\text{iter=100}}$	Difference (%)
w2v–cbow	**.588**	.558	(−3.0)
w2v–sg	**.586**	.551	(−3.6)
glove	**.555**	.464	(−9.1)
lexvec	**.570**	.561	(−0.9)

5.5.2 Minimum count threshold

Minimum-count thresholds are often neglected in the literature, where a default configuration of 0, 1 or 5 being presumably used by most authors. An exception to this trend is the threshold of 100 occurrences used by Levy, Goldberg and Dagan (2015), whose toolkit we use in *PPMI–SVD*. No explicit justification has been found for this higher word-count threshold. A reasonable hypothesis would be that higher thresholds improve the quality of the data, as it filters rare words more aggressively.

Table 5.9 presents the result from the highest-Spearman configurations alongside the results for an identical configuration with a higher occurrence threshold of 50. The results unanimously agree that a higher threshold does not contribute to the removal of any extra noise. In particular, for *PPMI–SVD*, it seems to discard enough useful information to considerably reduce the quality of the compositionality prediction measure. The results strongly contradict the default configuration used for *PPMI–SVD*, suggesting that a lower word-count threshold might yield better results for this task.

[8] The running time grows linearly with the number of iterations.

Table 5.9: Results for a higher minimum threshold of word count.

Model (*FR-comp*)	ρ_{base}	$\rho_{\text{mincount=50}}$	Difference (%)
w2v–cbow	**.660**	.610	(−5.0)
w2v–sg	**.672**	.613	(−5.9)
glove	**.680**	.673	(−0.7)
PPMI–SVD	**.584**	.258	(−32.6)
lexvec	**.677**	.653	(−2.4)
Model (*Reddy*)	ρ_{base}	$\rho_{\text{mincount=50}}$	Difference (%)
w2v–cbow	**.809**	.778	(−3.1)
w2v–sg	**.821**	.776	(−4.5)
glove	**.764**	.672	(−9.2)
PPMI–SVD	**.743**	.515	(−22.8)
lexvec	**.774**	.738	(−3.6)
Model (*PT-comp*)	ρ_{base}	$\rho_{\text{mincount=50}}$	Difference (%)
w2v–cbow	**.588**	.580	(−0.8)
w2v–sg	**.586**	.575	(−1.1)
glove	**.555**	.540	(−1.5)
PPMI–SVD	**.530**	.418	(−11.1)
lexvec	**.570**	.566	(−0.4)

5.5.3 Windows of size 2+2

For many models, the best window size found was either WindowSize=1+1 or WindowSize=4+4 (see Section 5.2.1). It is possible that a higher score could obtained by a configuration in between. While a full exhaustive search would be the ideal solution, a useful approximation of the best 2+2 configuration could be obtained by running the experiments on the highest-Spearman configurations, with the window size replaced by 2+2.

Results in Table 5.10 for a window size of 2+2 are consistently worse than the base model, indicating that the optimal configuration is likely the one that was obtained with window size of 1+1 or 4+4. This is further confirmed by the fact that most DSMs had the best configuration with window size of 1+1 or 8+8, with few cases of 4+4 as best model, which suggests that the quality of most configurations in the space of models is either monotonically increasing or decreasing with regards to these window sizes, favoring thus the configurations with more extreme WindowSize parameters.

Table 5.10: Results using a window of size 2+2.

Model (*FR-comp*)	ρ_{base}	$\rho_{\text{win=2+2}}$	Difference (%)
PPMI–SVD	**.584**	.397	(−18.7)
PPMI–thresh	**.702**	.678	(−2.4)
glove	**.680**	.657	(−2.3)
lexvec	**.677**	.671	(−0.6)
w2v–cbow	**.660**	.644	(−1.6)
w2v–sg	**.672**	.639	(−3.3)
Model (*Reddy*)	ρ_{base}	$\rho_{\text{win=2+2}}$	Difference (%)
PPMI–SVD	**.743**	.583	(−16.0)
lexvec	**.774**	.757	(−1.7)
w2v–cbow	**.809**	.777	(−3.2)
w2v–sg	**.821**	.784	(−3.7)
Model (*PT-comp*)	ρ_{base}	$\rho_{\text{win=2+2}}$	Difference (%)
PPMI–SVD	**.530**	.446	(−8.4)
PPMI–thresh	**.602**	.561	(−4.1)
lexvec	**.570**	.564	(−0.6)

5.5.4 Higher number of dimensions

As seen in Section 5.2.2, some DSMs obtain better results when moving from 250 to 500 dimensions, and this trend continues when moving to 750 dimensions. This behavior is notably stronger for *PPMI–thresh*, which suggests that an even higher number of dimensions could have better predictive power.

Table 5.11 presents the result of running *PPMI–thresh* for increasing values of of the Dimension parameter. The baseline configuration (indicated as * in Table 5.11) was the highest-scoring configuration found in Section 5.2.2: $\textit{lemma}_{PoS}$.w_1.d_{750} for *PT-comp* and *FR-comp*, and *surface*.w_8.d_{750} for *Reddy*. As seen in Section 5.2.2, results for 250 and 500 dimensions have lower scores than the results for 750 dimensions. Results for 1000 dimensions were mixed: they are slightly worse for *FR-comp* and $\textit{Reddy}^{++}$, and slightly better for *PT-comp*. Increasing the number of dimensions generates models that are progressively worse. These results suggests that the maximum vector quality is achieved between 750 and 1000 dimensions.

Table 5.11: Results for higher numbers of dimensions (*PPMI–thresh*).

Model (*FR-comp*)	$\rho_{\text{dim=X}}$	Difference (%)
dim = 250	.671	(−3.1)
dim = 500	.695	(−0.7)
dim = 750	**.702***	(0.0)
dim = 1000	.694	(−0.8)
dim = 2000	.645	(−5.8)
dim = 5000	.636	(−6.7)
dim = 30000	.552	(−15.1)
dim = 999999	.539	(−16.3)
Model (*Reddy*)	$\rho_{\text{dim=X}}$	Difference (%)
dim = 250	.764	(−2.7)
dim = 500	.782	(−1.0)
dim = 750	**.791***	(0.0)
dim = 1000	.784	(−0.7)
dim = 2000	.760	(−3.1)
dim = 5000	.744	(−4.7)
dim = 30000	.700	(−9.1)
dim = 999999	.566	(−22.5)
Model (*PT-comp*)	$\rho_{\text{dim=X}}$	Difference (%)
dim = 250	.543	(−5.9)
dim = 500	.546	(−5.6)
dim = 750	.602*	(0.0)
dim = 1000	**.609**	(+0.7)
dim = 2000	.601	(−0.1)
dim = 5000	.505	(−9.7)
dim = 30000	.532	(−7.0)
dim = 999999	.500	(−10.2)

5.5.5 Random initialization

The word vectors generated by the *glove* and *w2v* models have some level of non-determinism caused by random initialization and random sampling techniques. A reasonable concern would be whether the results presented for different parameter variations are close enough to the scores obtained by an average model. To assess the variability of these models, we evaluated 3 different runs of every DSM configuration (the original execution ρ_1, used elsewhere in this thesis, along with two other executions ρ_2 and ρ_3) for *glove*, *w2v–cbow* and *w2v–sg*. We then calculate the average ρ_{avg} of these 3 executions for every model.

Table 5.12 reports the highest-Spearman configurations of ρ_{avg} for the *Reddy* and *Reddy*$^{++}$ datasets. When comparing ρ_{avg} to the results of the original execution ρ_1, we see that the variability in the different executions of the same configuration is minimal.

This is further confirmed by the low sample standard deviation[9] obtained from the scores of the 3 executions. Given the high stability of these models, results in the rest of the thesis were calculated and reported as ρ_1 for all datasets.

Table 5.12: Configurations with highest ρ_{avg} for non-deterministic models.

Dataset	DSM	configuration	ρ_1	ρ_2	ρ_3	ρ_{avg}	stddev
Reddy	*glove*	$\textit{lemma}_{PoS}.\text{w}_8.\text{d}_{250}$	.759	.760	.753	.757	.004
	w2v–cbow	$\textit{surface}.\text{w}_1.\text{d}_{500}$	.796	.807	.799	.801	.006
	w2v–sg	$\textit{surface}.\text{w}_1.\text{d}_{750}$	.812	.788	.812	.804	.014
Reddy^{++}	*glove*	$\textit{lemma}_{PoS}.\text{w}_8.\text{d}_{500}$	.651	.646	.650	.649	.003
	w2v–cbow	$\textit{surface}^{+}.\text{w}_1.\text{d}_{750}$	.730	.732	.728	.730	.002
	w2v–sg	$\textit{surface}^{+}.\text{w}_1.\text{d}_{750}$	.741	.732	.721	.731	.010

5.5.6 Data filtering

Along with the verification of parameters, we also evaluate whether dataset variations could yield better results. In particular, we consider the use of filtering techniques, which are used in the literature as a method of guaranteeing dataset quality. As per Roller, Walde and Scheible (2013), we consider two strategies of data removal: (1) removing individual outlier compositionality judgments through z-score filtering; and (2) removing all annotations from outlier human judges. A compositionality judgment is considered an outlier if it stands at more than z standard deviations away from the mean; a human judge is deemed an outlier if its Spearman correlation to the average of the others ρ_{oth} is lower than a given threshold R[10]. These methods allow us to remove accidentally erroneous annotations, as well as annotators whose response deviated too much form the mean (in particular spammers and non-native speakers).

Table 5.13 presents the evaluation of raw and filtered datasets regarding two quality measures: the average of the standard deviations for all NCs ($\overline{\sigma_{\mathbf{WC}}}$); and the proportion of NCs in the dataset whose standard deviation is higher than 1.5 ($P_{\sigma>1.5}$), as per Reddy, McCarthy and Manandhar (2011). The results suggest that filtering techniques can improve the overall quality of the datasets, as seen in the reduction of the proportion of NCs with high standard deviation, as well as in the reduction of the average standard

[9] The low standard deviation is not a unique property of high-ranking configurations: The average of deviations for all models was .004 for *Reddy*$^{++}$ and .006 for *Reddy*.

[10] The judgment threshold we adopted was $z = 2.2$ for *EN-comp*$_{90}$, $z = 2.2$ for *PT-comp* and $z = 2.5$ for *FR-comp*. The human judge threshold was $R = 0.5$.

deviation itself. We additionally present the data retention rate (DRR), which is the proportion of NCs that remained in the dataset after filtering. While the DRR does indicate a reduction in the amount of data, this reduction may be considered acceptable in light of the improvement suggested by the quality measures.

Table 5.13: Intrinsic quality measures for the raw and filtered datasets

Dataset	$\overline{\sigma_{\mathbf{WC}}}$		$P_{\sigma>1.5}$		DRR
	raw	filtered	raw	filtered	
FR-comp	1.15	0.94	22.78%	13.89%	87.34%
PT-comp	1.22	1.00	14.44%	6.11%	87.81%
EN-comp$_{90}$	1.17	0.87	18.89%	3.33%	83.61%
Reddy	0.99	—	5.56%	—	—

On a more detailed analysis, we have verified that the improvement in these quality measures is heavily tied to the use of z-score filtering, with similar results obtained when it is considered alone. The application of R-filtering by itself, on the other hand, did not show any noticeable improvement in the quality measures for reasonable amounts of DRR. This is the opposite from what was found by Roller, Walde and Scheible (2013) on their German dataset, where only R-filtering was found to improve results under these quality measures. We present our findings in more detail in Cordeiro, Ramisch and Villavicencio (2016a).

We then consider whether filtering can have an impact on on the performance of predicted compositionality scores. As z-score filtering was responsible for improvement in quality measures above, we consider For each of the 228 model configurations that were constructed for each language, we launched an evaluation on the filtered *EN-comp*$_{90}$, *FR-comp* and *PT-comp* datasets (use use z-score filtering only, as it was responsible for most of the improvement in quality measures). Overall, no improvement was observed in the results of the prediction (values of Spearman ρ) when we compare raw and filtered datasets. Looking more specifically at the best configurations for each DSM (Table 5.14), we can see that most results do not significantly change when the evaluation is performed on the raw or filtered datasets. This suggests that the amount of judgments collected for each compound greatly offsets any irregularity caused by outliers, making the use of filtering techniques superfluous.

Table 5.14: Extrinsic quality measures for the raw and filtered datasets

Dataset	*EN-comp*$_{90}$		*FR-comp*		*PT-comp*	
	raw	filtered	raw	filtered	raw	filtered
PPMI–SVD	**.604**	.601	**.584**	.579	**.530**	.526
PPMI–TopK	.564	**.571**	**.550**	.545	**.519**	.516
PPMI–thresh	.602	**.607**	**.702**	.700	**.602**	.601
glove	.538	**.544**	**.680**	.676	**.555**	.552
lexvec	.567	**.572**	**.677**	.676	**.570**	.568
w2v–cbow	**.669**	.665	**.651**	**.651**	**.588**	.587
w2v–sg	**.665**	.661	.653	**.654**	**.586**	.584

5.6 Error analysis

In the previous sections, we have studied the performance of the compositionality prediction framework in terms of the correlation between system predictions and human judgments. We now investigate the system output with regards to other variables that may have an impact on results, such as corpus frequency and conventionalization. We also compare the predicted compositionality scores with some patterns we previously found in human scores (see Section 3.2).

5.6.1 Frequency and compositionality prediction

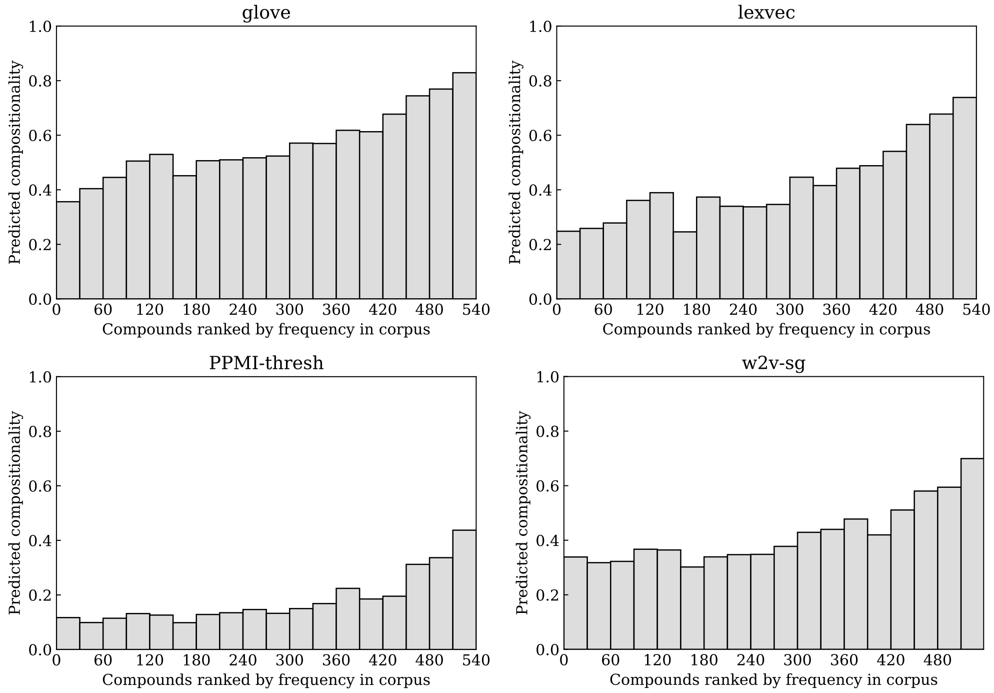
Results from an evaluation of the hypothesis $\mathrm{h}_{\mathrm{idiom} \approx \mathrm{distr.freq}}$ in Section 3.2.4 show that the frequency of NCs in large corpora is somewhat associated with the compositionality scores assigned by humans. We investigate whether this correlation also holds true to system predictions: are the most frequent NCs being predicted as more compositional?

In this experiment, we focus on a cross-language analysis with the *ALL-comp* dataset, which combines the $3 \times 180 = 540$ NCs from the three datasets presented in Chapter 3. Figure 5.11 presents the 540 NCs, ordered according to corpus frequency and grouped into 18 bins of 30 NCs each.[11] The height of each bin indicates the average of the scores predicted (using the *uniform* strategy) by a given system to the 30 NCs therein. There is a high variability in the level of correlation between the corpus frequency of compounds and the prediction of the models. The level of correlation ranged from $\rho = .28$ for *PPMI–TopK* (not shown here) to $\rho = .68$ for *glove*, with the intermediate results of $\rho = .36$ for *PPMI–SVD*, $\rho = .46$ for *PPMI–thresh*, $\rho = .50$ for *w2v–sg*, $\rho = .51$ for *w2v–*

[11] We use binning so as to smooth over the outliers.

cbow and $\rho = .54$ for *lexvec*. For every system, the correlation was significant ($p < 0.05$). This is in line with human judgments of compositionality, which also had a positive correlation with the frequency of the NCs.

Figure 5.11: Compositionality prediction under different frequency bins.



Another hypothesis ($\mathrm{h}_{\mathrm{accur} \leftarrow \mathrm{MWE.freq}}$) we test is whether higher-frequency NCs are easier to predict. A first intuition would be that this hypothesis is true, as a higher number of occurrences is also associated with a larger amount of data, from which more representative vectors could be built. To test this hypothesis, we calculated the correlation between NC frequency and the human–system difference $|h - s|$, where h is the human score and s is the system's predicted score for a given compound. Higher values of human–system difference indicate that an NC's compositionality is harder to predict. We found a weak (though statistically significant) correlation for some of the systems: *PPMI–TopK* had $\rho = .15$, *PPMI–SVD* had $\rho = .17$, and *PPMI–thresh* had $\rho = .22$ (all with $p < 0.05$). This correlation is positive, which means that the frequency is correlated with difficulty. This implies that the compositionality of rarer NCs was mildly *easier* to predict for these systems, suggesting that the hypothesis above is false. On the other hand, *glove* had an easier time predicting frequent NC, with negative correlation of $\rho = -.19$, favoring the aforementioned hypothesis. Moreover, the correlation was not statistically significant for *lexvec* and *w2v* models. These results are mixed, and either point to an overall lack of correlation between frequency and difficulty, or indicate mild DSM-specific behaviors, which should be investigated in further research.

5.6.2 Conventionalization and compositionality prediction

Section 2.2.3 has described the PMI as one well-known estimator of the level of MWE conventionalization. Many of the DSMs investigated on this thesis also rely on PMI as a way to estimate the strength of association between two words. This measure is then directly applied to target–context word pairs during the construction of the DSM, and the result becomes an internal matrix that is further processed to build the real-valued output vectors. In light of the results from the previous section, and given the reliance of most DSMs on the PMI for the construction of their word representation, one might expect similarly high correlations between compositionality and the PMI of compound elements. On the other hand, given the lack of correlation found between the conventionalization and human judgments of compositionality, good system predictions should ideally not correlate with measures of conventionalization such as the PMI. We thus evaluate whether our model really is predicting something different from conventionalization.

Figure 5.12: Compositionality prediction under different PMI bins.

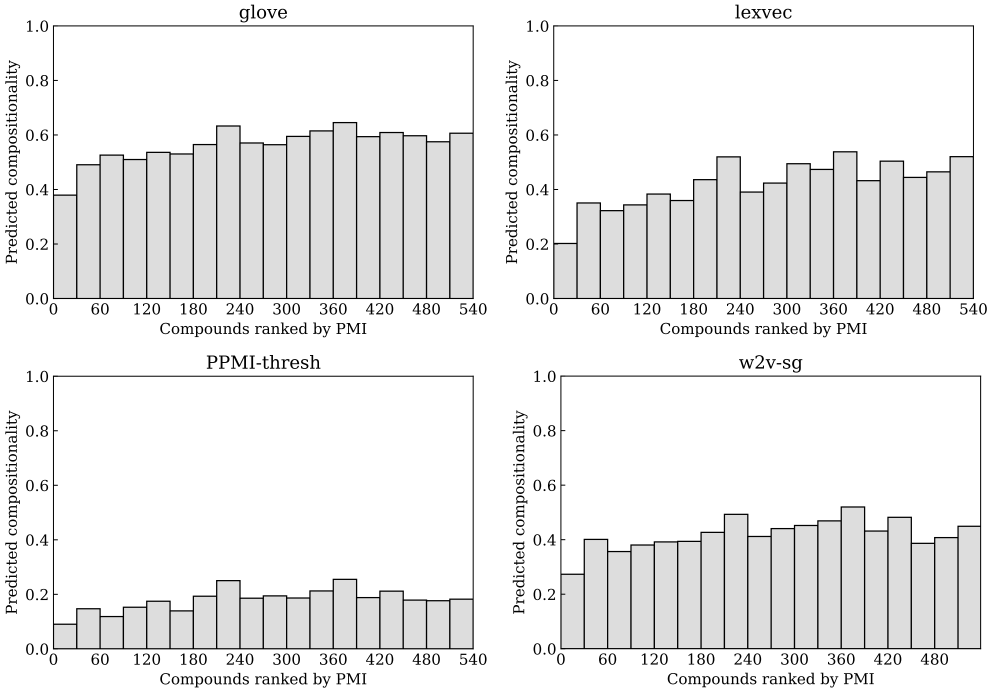


Figure 5.12 presents the 540 NCs of *ALL-comp*, ordered according to PMI and grouped under 18 bins of 30 NCs each. The height of each bin indicates the average of the scores predicted (using the *uniform* strategy) by a given system to the 30 NCs therein. The effects are milder than the ones seen for the frequency (in Section 5.6.1). Statistically significant correlations are $\rho = .13$ for *PPMI–thresh*, $\rho = .17$ for *w2v–cbow* and *w2v–sg*, $\rho = .26$ for *glove* and *lexvec*, and $\rho = .28$ for *PPMI–SVD*. No correlation was found for *PPMI–TopK*. Overall, these results suggest that the vector representations generated by these models preserve some level of information regarding the strength of association between words. Given that there was no correlation between PMI and human-rated compositionality when testing hypothesis $\mathrm{h}_{\text{idiom} \approx \text{distr.convent}}$ in Section 3.2.4, the systems that do keep this information are at a disadvantage. Particularly in the case of *w2v* models, this result is surprising, as it suggests that its high scores could be further improved by a method that did not keep as much of a correlation with the PMI in the word-embedding representation.

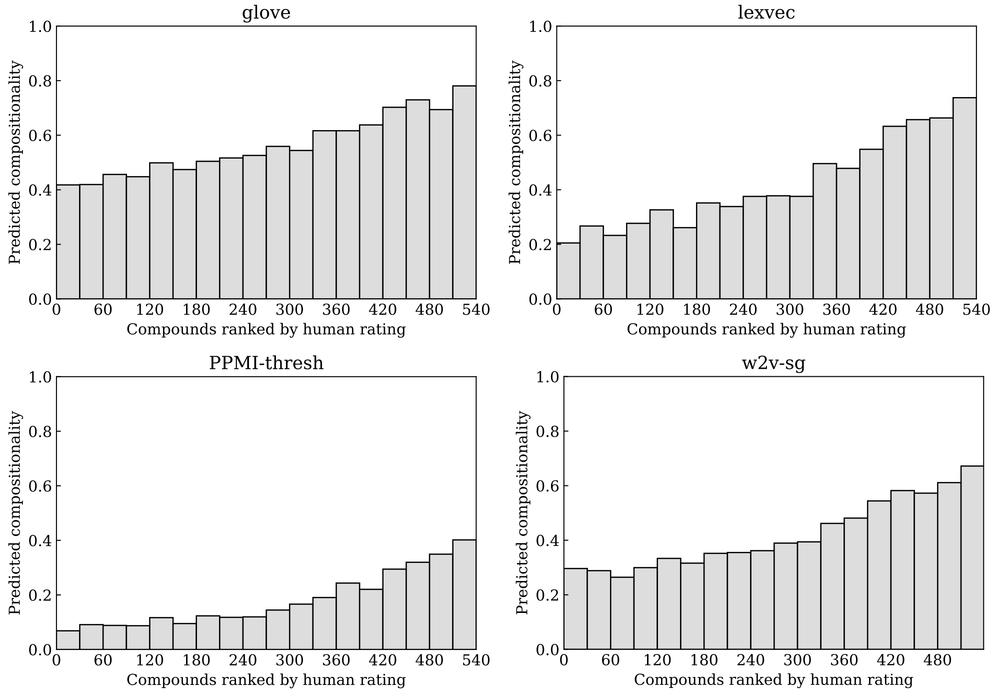
We also calculated the correlation between the PMI and the human–system difference, calculated as $|h - s|$, where h is the human score and s is the predicted system score for a given NC. The hypothesis ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE.convent}}$) is that the DSMs should have lower accuracy when dealing with less conventionalized NCs (and whose elements are not strongly associated through PMI), due to a lower amount of shared contexts. However, for almost all DSMs, the results obtained do not show a statistically significant correlation, suggesting that this hypothesis is not true. For *lexvec*, there was a minor negative correlation of $\rho = -.12$ ($p < 0.05$) between the PMI and the difficulty, indicating that NCs with higher PMI do have slightly more accurate internal representation than the others in this particular DSM. This differs from the results obtained when comparing the human–system difference with NC frequency (Section 5.6.1), in which *lexvec* did not show any statistically significant correlation, but most other models did. As in the case of frequency, the *w2v* models showed no correlation between difficulty of prediction and the PMI.

5.6.3 Human–system comparison

The general hypothesis $\mathrm{h}_{\text{pred-comp} \approx \text{comp}}$ predicted a correlation between human-rated NC compositionality and model predictions, and this has been extensively verified in the highest-Spearman predictions (e.g. in Section 5.1). In this section, we present a visual validation of this hypothesis, by considering the highest-Spearman predictions of 4 DSMs, with all datasets combined.

Figure 5.13 presents 4 graphs (one per DSM), with the predicted compositionality of the NCs in *ALL-comp* for the best configuration of each language. The NCs were ranked by human compositionality scores, and grouped under 18 bins of 30 NCs each. The height of each bin indicates the average of the scores predicted (using the *uniform* strategy) by a given system to the 30 NCs therein. The four systems present a behavior that is consistent with their Spearman scores (see Section 5.1), where system predictions grow along with the corresponding human ratings.

Figure 5.13: Compositionality prediction as a function of human judgments.



For all systems but *PPMI–thresh*, the pattern of predicted compositionality grows mostly linearly with respect to the human scores (and this includes *PPMI–TopK* and *PPMI–SVD*, not shown here). The *PPMI–thresh* system ratings behave unusually, with overall lower predicted scores and a super-linear pattern of predictions, suggesting that the model is quite capable of capturing different levels of compositionality for the most compositional NCs, but fails at capturing the compositionality on the idiomatic side of the spectrum. This pattern may be explained by the fact that *PPMI–thresh* uses a sparse context representation (without any kind of dimensionality reduction other than context filtering), which means that the intersection of two vectors is often a vector with zero in many dimensions, yielding overall lower scores, especially for more idiomatic cases.

5.6.4 Range-based analyses

The Spearman score assesses the performance of a given model by providing a single numerical value. This facilitates the comparison between different models, but it hides the internal behavior of the predictions. By splitting the datasets into different ranges, we obtain a more fine-grained view of the pattern that governs the prediction of each model.[12]

Figure 5.14 presents the highest-Spearman models (seen in Section 5.1), evaluated separately on 3 different sub-datasets of 60 NCs, split according to the standard deviation among human annotators (low, mid-range, and high values of $\sigma_{\mathbf{WC}}$)[13]. High values of standard deviation indicate disagreement among annotators, which can be regarded as an indicator that the annotation was difficult for humans. We can see that low-deviation NCs obtained considerably better system scores than the NCs for which humans disagreed among themselves. This can be taken as an evidence in favor of the hypothesis $\mathrm{h}_{\text{accur} \leftarrow \text{MWE.diffic}}$ that higher scores are achieved for NCs that were easier for humans to annotate (i.e. that had lower standard deviation of human ratings), and suggests that part of the difficulty of this task is related to the inability of humans to determine a consensual interpretation for each NC.

We have similarly evaluated the datasets based on three ranges of compositionality scores (low, mid-range and high values of $c_{\mathbf{WC}}$). The underlying hypothesis ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE.idiom}}$) was that compositional NCs would be more precisely classified by the model than idiomatic NCs, as the former have been more extensively considered in the literature (MITCHELL; LAPATA, 2010; MIKOLOV et al., 2013). Here, we consider the 540 NCs of *ALL-comp*, divided in three sub-datasets based on the level of human-rated compositionality, with 180 NCs in each sub-dataset[14]. Table 5.15 presents the Spearman score obtained on each sub-dataset for the highest-Spearman configuration of each DSM.[15] The results suggest that distinctions on the level of compositionality are easier to perform for compositional compounds than they are for idiomatic compounds. In all cases, however, the result for sub-datasets was far lower than the score obtained for the full dataset. This might be

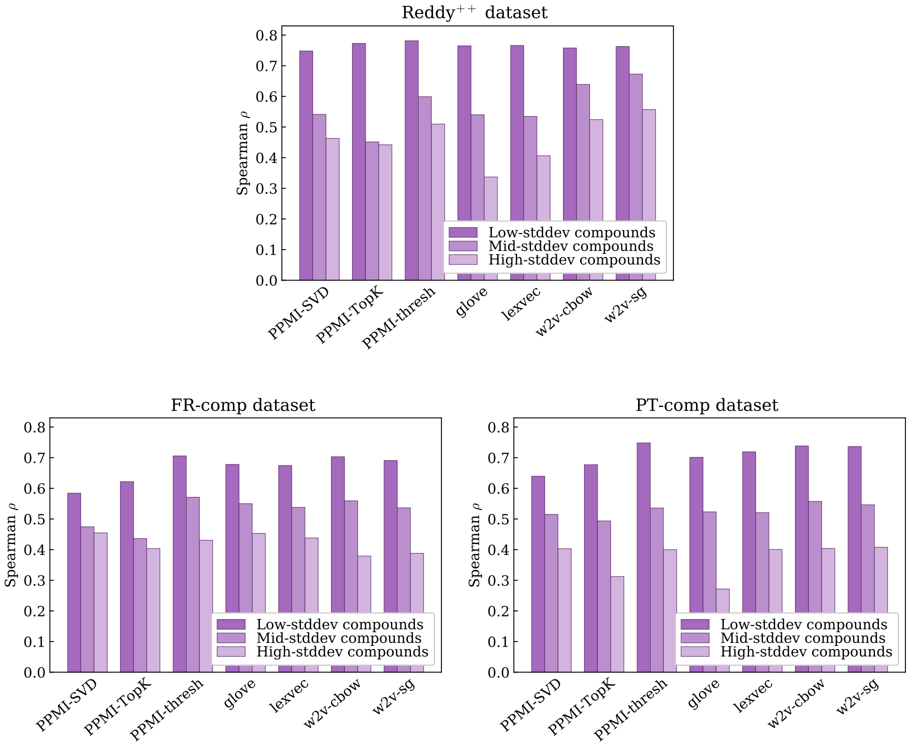
[12] The experiments in this section involve *Reddy*$^{++}$, *FR-comp* and *PT-comp*, but not *Farahmand*, as the latter dataset has binary judgments and thus cannot be easily split in ranges.

[13] All Spearman scores for sub-datasets had $p < 0.05$.

[14] Scores from compounds in different languages are mixed together in each sub-dataset.

[15] All Spearman scores for datasets and sub-datasets had $p < 0.05$.

Figure 5.14: Spearman of best *uniform* models, separated by $\sigma_{\mathbf{WC}}$ ranges.



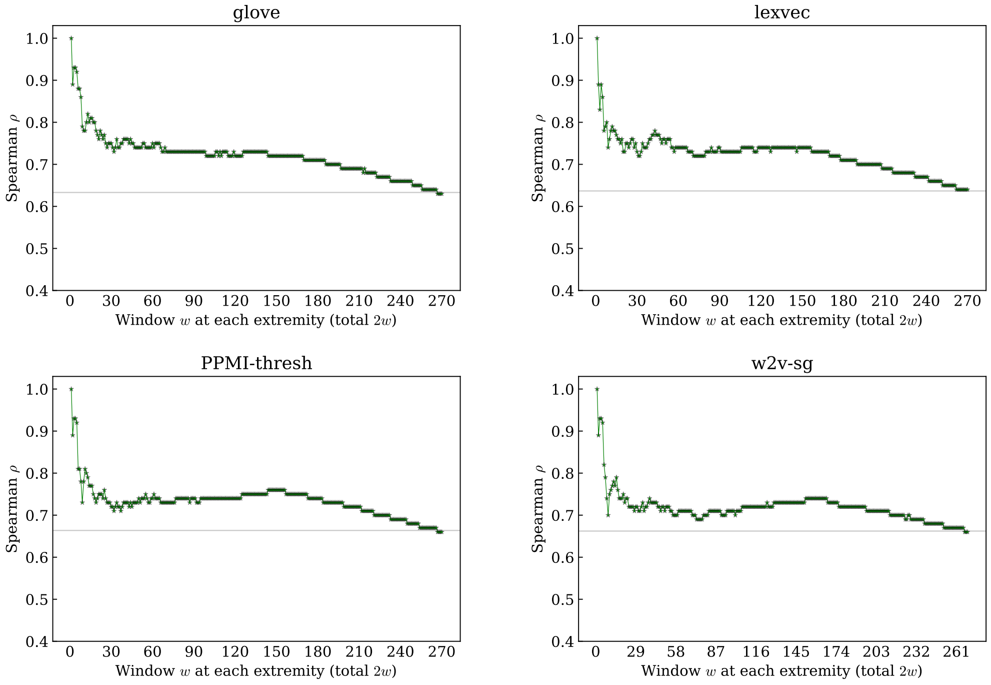
explained by the fact that it is harder to make fine-grained distinctions of compositionality, while inter-range distinctions are more straightforward. In other words, it is easier to distinguish between a idiomatic compound (such as *ivory tower*) and a compositional one (such as *access road*) than it is to distinguish between two compositional compounds (such as *access road* and *subway system*).

Table 5.15: Spearman of best *uniform* models, separated by $c_{\mathbf{WC}}$ ranges.

Model	full dataset	low	mid	high
PPMI–thresh	**0.66**	0.29	0.24	**0.37**
glove	**0.63**	0.27	0.26	**0.35**
lexvec	**0.64**	0.18	0.20	**0.37**
w2v–sg	**0.66**	0.16	0.24	**0.32**

The results above suggest that higher scores could be obtained by considering only the compounds with scores in the two extremities: lowest and highest compositionality. We evaluate this hypothesis for a given DSM by merging the predictions of its highest-Spearman configurations for *Reddy*$^{++}$, *FR-comp* and *PT-comp* (creating a single set of 540 compositionality predictions). We then consider different subsets of NC predictions in the extremities. In particular, for every window w from 1 to $270 = 540/2$, we consider the subset of w NC with lowest score prediction along with the subset of w NCs with the highest score prediction. We then calculate the Spearman ρ for this subset of $2w$ NCs, for different values of w. Figure 5.15 presents such results. As can be seen, for all 4 DSMs considered, low values of w consistently result in high Spearman scores, suggesting that the DSMs encode enough semantic information to make coarse-grained distinctions of compositionality. As we consider increasingly more cases of partially-compositional NCs (with higher values of w), we obtain increasingly lower results, until we arrive at the whole dataset of 540 NCs, where we get the lowest Spearman scores in every DSMs.

Figure 5.15: Compositionality sliding windows, evaluating top w + bottom w compounds, for different values of w.



We have additionally performed both standard-deviation and compositionality-range analyses for other prediction strategies than *uniform*. In the case of *arith*, the Spearman score for different sub-datasets followed very closely the results of *uniform*. In the case of *maxsim*, we hypothesized that its favoring of a compositional reading of every compound would optimize results for the compositional sub-dataset when compared to *uniform*. Nevertheless, the results fluctuated around the *uniform* scores, with no clear pattern of improvement for this model. As for *geom*, we previously hypothesized that their tendency to lowering the compositionality score would optimize the quality of prediction for idiomatic compounds. The results refuted this hypothesis. Most scores were similar to *uniform* scores, with improvements seen more often in the compositional range than in the idiomatic range. However, even then the differences were small and the pattern of improvement unclear.

5.7 Summary and discussion

In this chapter, we have described the results of a large-scale evaluation of parameter choices in a DSM-based framework of compositionality prediction. Evaluations were performed on six datasets, spanning across three languages. We have built 228 DSMs for each language, and evaluated more than 8 thousand prediction model configurations, examining the impact of DSM choice and various types of parameters.

The compositionality prediction model proposed in this thesis was implemented as part of the mwetoolkit, and is freely available online.[16] Given the large amount of experiments performed in this thesis, and in order to guarantee the reproducibility of results, we defined our experiments through a system of file dependencies. Every step of preprocessing (e.g. re-tokenization of compounds as a single unit, removal of stopwords) was defined in term of these dependencies, so that any modification in the code (e.g. bug fixes) would automatically invalidate experiment results. The results presented in this chapter were obtained under this system of dependencies.

Considering the experimental results in terms of DSMs, the *w2v* models performed better than *PPMI* for *Reddy*$^{++}$, both were in a tie for *Farahmand*, and *w2v* was outperformed by *PPMI–thresh* for *FR-comp* and *PT-comp*. The performance of *glove* on English datasets was underwhelming, and might be related to the lack of tuning of model-specific parameters. As previously argued by Salehi, Cook and Baldwin (2015), *PPMI–TopK* is not an appropriate DSM for this task, as it does not model relevant co-occurrence very well.

When comparing Dimension across languages and datasets, larger values often bring better performance, likely due to the possibility of representing more fine-grained semantic distinctions (in agreement with the hypothesis $\mathrm{h}_{\mathrm{accur} \leftarrow \mathrm{DSM.dims}}$). An upper limit of around 1000 dimensions has been verified, however, with even higher numbers of dimensions obtaining lower scores.

The most effective WindowSize depends on the model and language, but for the best models in all datasets, a window of 1+1 outperforms the others (which suggests that $\mathrm{h}_{\mathrm{accur} \leftarrow \mathrm{DSM.window}}$ is false). This may be a consequence of the fact that higher window

[16] <http://mwetoolkit.sf.net>

sizes are more likely to consider unrelated words as part of a target's context.

Regarding the WordForm, the *lemma* (i.e. stopword removal + lemmatization) seems to be the overall best type of preprocessing across languages (as predicted by $\text{h}_{\text{accur} \leftarrow \text{corpus.wordform}}$). The use of POS tags does not seem to improve on the results, which could indicate that the higher precision of grammatical category does not compensate for the added sparsity. In the case of English, the effects of both stopword removal and lemmatization are questionable, with plain surface-level word-forms producing slightly better models in some cases.

Corpus size seems to play a fundamental role in the quality of the constructed distributional models, as corpora with less than a billion tokens result in considerably weaker predictions ($\text{h}_{\text{accur} \leftarrow \text{corpus.size}}$). However, the improvement in prediction quality seems to be capped at around a threshold of one billion tokens: larger corpora do not result in better predictions of compositionality for nominal compounds. This threshold may be related to the minimum frequency necessary for rarer NCs so as to permit the calculation of cosine similarity with its components.

The technique of parallel predictions was shown to perform equivalently to whole-corpus predictions ($\text{h}_{\text{accur} \leftarrow \text{corpus.parallel}}$). While the use of this technique does not improve on the results obtained through whole-corpus models, it does permit a more flexible utilization of computational resources (e.g. clusters) in the construction of the underlying semantic representations.

Regarding the different compositionality prediction strategies, the *uniform* strategy produces predictions that are consistently among the best ones. The *maxsim* strategy does improve the prediction of compositional NCs, but only for outlier cases, contributing to random variation in most cases ($\text{h}_{\text{strat.maxsim}}$). While this does not improve on the results from *uniform*, it does consistently produce similarly good results. The *head* and *mod* strategies perform surprisingly well for all top models of every dataset, in spite of their reliance on incomplete information ($\text{h}_{\text{strat.partial-info}}$). The performance of *arith* is quite similar to *uniform*, reflecting the fact that both rely on an additive model of compositionality ($\text{h}_{\text{strat.arith} \approx \text{strat.uniform}}$). The *geom* strategy did optimize the scores of idiomatic NCs, but at the expense of a pessimization of scores for some compositional cases ($\text{h}_{\text{strat.geom}}$). A combination of the *geom* and *maxsim* strategies is left for future work.

Concerning the sanity checks, we found no advantage in the use of a higher number of iterations for the construction of DSMs. The minimum word-count has similarly been found to be a small value, with higher thresholds removing too much information. An evaluation of the random initialization used in some DSMs found no difference in the final results across multiple executions. Regarding the dataset scores, filtering techniques were also considered, but the results were comparable to the ones obtained on the unfiltered datasets.

This chapter has also performed an error analysis of the predicted compositionality scores. As in the case of human-rated scores, frequency was found to be positively correlated with compositionality ($\mathrm{h}_{\text{idiom} \approx \text{distr.freq}}$). This result disputes the hypothesis that idiomatic expressions are more frequent. In the case of PMI, while it was not correlated to human-rated scores, it did show a mild correlation with some system scores, suggesting that these systems could be improved by reducing their reliance on that measure ($\mathrm{h}_{\text{idiom} \approx \text{distr.convent}}$). Intra-NC standard deviation on human ratings has also been shown to be related to system scores: systems have difficulty on NCs that humans also find difficult ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE.diffic}}$). Moreover, system predictions were found to have higher quality in the case of compositional expressions ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE.idiom}}$). Further work would be required to improve score predictions of idiomatic NCs.

An overall recommendation for future work would be the use of large dimensions and small window sizes. Moreover, investing in preprocessing provides a good balance of a small vocabulary (of lemmas) and good accuracy. The underlying corpus size should contain at least 1 billion tokens. As for the underlying model, the simple *uniform* prediction strategy can achieve the highest-quality predictions.

Regarding the choice of DSM, the average Spearman's ρ for *Reddy* over all tested parameter configurations was 0.71 for both *w2v* models and 0.67 for *PPMI–thresh*, suggesting that both types of models can obtain good results. While *PPMI–thresh* is a simple, fast and inexpensive model to build, *w2v* has a free and push-button implementation, and requires less hyper-parameter tuning, as is it seems more robust to parameter variation.

More generally, the best results obtained are comparable and even outperform the state of the art. Table 5.16 compares the highest results in the literature for the *Reddy* dataset against the highest-Spearman and highest-Pearson configuration obtained for each DSM.[17] Reddy, McCarthy and Manandhar (2011) use a compositionality prediction model with a global set of contexts that resembles *PPMI–TopK*, and the results are correspondingly similar to the ones obtained for this DSM. Salehi, Cook and Baldwin (2014) also

[17] Due to space constraints, only the highest-Spearman configuration is shown.

use global contexts, but augment it with information obtained from translations, which improves the results (they are somewhat comparable to our highest-Pearson *PPMI–SVD* configuration). Salehi, Cook and Baldwin (2015) use a configuration that is similar to our highest-Pearson *w2v–cbow*. We obtain slightly better results due to our exploration of the space of DSM and corpus configurations.[18]

Table 5.16: Comparison of our best models with state-of-the-art results for *Reddy*. Results in parentheses for fallback evaluation.

Model & Parameters	Spearman ρ	Pearson r
Reddy, McCarthy and Manandhar (2011)	.714	—
Salehi, Cook and Baldwin (2014)	—	.744
Salehi, Cook and Baldwin (2015)	—	.796
Best *w2v–sg* [Spearman: *surface*.w$_1$.d$_{750}$]	**.812** (.812)	**.814** (.814)
Best *PPMI–thresh* [Spearman: *surface*.w$_8$.d$_{750}$]	.791 (.803)	.762 (.768)
Best *w2v–cbow* [Spearman: *surface*$^+$.w$_1$.d$_{500}$]	.796 (.796)	.803 (.798)
Best *lexvec* [Spearman: *surface*$^+$.w$_4$.d$_{500}$]	.774 (.773)	.787 (.787)
Best *glove* [Spearman: *lemma*$_{PoS}$.w$_8$.d$_{250}$]	.754 (.759)	.783 (.787)
Best *PPMI–SVD* [Spearman: *surface*$^+$.w$_1$.d$_{500}$]	.743 (.743)	.738 (.726)
Best *PPMI–TopK* [Spearman: *lemma*$_{PoS}$.w$_8$.d$_{1000}$]	.706 (.716)	.732 (.717)

Our results are also comparable to the state of the art regarding the *Farahmand* dataset, particularly when the fallback evaluation is adopted, as shown in Table 5.17. The predictive model of Yazdani, Farahmand and Henderson (2015) generalizes the linear combination of word representations (such as the one used on the *uniform* strategy) so as to allow for other polynomial projections, with quadratic projections on *w2v–cbow* obtaining the highest BF$_1$ score of .487. We show that a DSM and corpus parameter tuning can beat the use of these more complex functions, as our best configuration for *w2v–cbow* obtains a BF$_1$ of .512. Future work should investigate the joint use of quadratic projections and the recommended DSM configurations from this thesis.

[18] Note that the main goal was not to beat the state of the art, but to explore the space of configurations.

Table 5.17: Comparison of our best models with state-of-the-art BF1 for *Farahmand*. Results in parentheses for fallback evaluation.

Model & Parameters	BF_1
Yazdani, Farahmand and Henderson (2015)	.487
Best *w2v–cbow* [*lemma*.w_1.d_{750}]	**.512** (.471)
Best *w2v–sg* [*lemma*.w_4.d_{500}]	.507 (.468)
Best *lexvec* [*surface*.w_1.d_{750}]	.449 (.431)
Best *PPMI–SVD* [*lemma*.w_4.d_{750}]	.487 (.424)
Best *PPMI–thresh* [*lemma*.w_4.d_{750}]	.472 (.404)
Best *PPMI–TopK* [*lemma*.w_8.d_{1000}]	.435 (.376)
Best *glove* [$lemma_{PoS}$.w_8.d_{750}]	.400 (.358)

6 EXTRINSIC EVALUATION OF COMPOSITIONALITY PREDICTION

The accurate identification of MWEs in running text is a major challenge in the general pipeline of NLP applications. The set of all possible categories of MWEs in a language can be quite diverse (SCHNEIDER et al., 2014b; CONSTANT et al., 2017), and the often-employed method of identifying such expressions from a predetermined lexicon may not yield satisfactory results for productive MWE patterns (such as nominal compounds). MWE identification has notably been one of the goals of the SemEval 2016 task 10: DiMSUM (Detecting Minimal Semantic Units and their Meanings) (SCHNEIDER et al., 2016a). In this shared task, participants were expected to present a system that was able to detect and group MWEs, and to assign supersense tags to each semantic unit (MWE or single word).

In this chapter, we consider an extrinsic evaluation of predicted compositionality scores, which are adopted as features in a system of MWE identification. The hypothesis we want to evaluate is $\text{h}_{\text{pred-comp} \rightarrow \text{ident-accur}}$, which predicts that the task of MWE token identification should benefit from the use of compositionality scores. We focus on the identification of noun-based compounds (i.e. nominal compounds, including proper names and nominal compounds with prepositions, such as *chamber of commerce*). For the identification of other categories of MWEs (as well as our work on supersense tagging), we refer to the paper that describes our submission for the DiMSUM shared task (CORDEIRO; RAMISCH; VILLAVICENCIO, 2016c), as well as the paper on CRF-based detection of MWEs (SCHOLIVET; RAMISCH; CORDEIRO, 2017). Section 6.1 presents two methods of MWE identification. Section 6.2 describes the experimental setup for the extrinsic evaluation. Section 6.3 then presents the results obtained with and without compositionality scores. Finally, Section 6.4 concludes with the summary of the main findings from this chapter.

6.1 Proposed models of MWE identification

In the interest of validating the compositionality prediction model proposed in this thesis, we consider two methods of MWE token identification, both of which can be applied with or without the feature of compositionality scores. The task of MWE identification consists in taking a tokenized corpus as input and generating an extra layer in which every occurrence of an MWE is explicitly indicated.[1] We consider two techniques of MWE identification: a rule-based method and a probabilistic method.

6.1.1 Rule-based identification model

Rule-based methods identify MWE occurrences by projecting type-level representations from a lexicon onto a layer of MWE occurrences in a corpus (see Section 2.3.3 on MWE token identification). We propose a baseline model of rule-based MWE identification which identifies words in the corpus that correspond to MWE entries in the lexicon. This identification is based on lemmas and POS tags, and may be done on a preexisting lexicon or on a list of MWE candidates extracted through techniques of MWE type discovery (described in Section 2.3.2).

We perform MWE token identification using an augmented version of the mwetoolkit, including support for both type-level discovery and token-level identification of contiguous and non-contiguous MWEs based on some degree of customization (CORDEIRO; RAMISCH; VILLAVICENCIO, 2015). MWE type-level candidates are extracted from a training corpus through syntactic patterns, without losing track of their token-level occurrences, to guarantee that all the MWE occurrences learned from the training data can be projected onto the test corpus. These candidates can then be filtered based on a variety of conditions (in particular, whether their occurrences are always annotated in the training corpus). The resulting set of candidates can then be automatically projected onto a layer of corpus MWE occurrences. We will use this as a baseline model, and as such, the identification will be context-independent (identifying every possible occurrence as an MWE regardless of any contextual clues).

These are the main functionalities that we have developed and integrated into the mwetoolkit for experiments on MWE identification:

1. Different match distances:

    - Longest: Matches the longest possible candidate. Useful e.g. for nominal compounds, where we want to match the whole compound.
    - Shortest: Matches the shortest possible candidate. Useful e.g. for phrasal verbs, where we want to find only the closest particle.

[1] A full review of MWE identification methods is out of the scope of this work. We refer to the *MWE Identification* section of Constant et al. (2017) for a thorough survey of other methods of MWE identification.

   - All: Matches all possible candidates. Useful as a fallback when shortest and longest are too strict (post-processing is then required).

2. Different match modes:

   - Non-overlapping: Matches at most one MWE per word in the corpus.
   - Overlapping: Allows words to be part of more than one MWE. This can be used to find MWEs that occur inside the gap of another MWE, or MWE occurrences that share a token.

3. Source-based identification: When information is retrieved in MWE type discovery, we keep a detailed description of the source corpus and sentence. The identification step can then be quickly performed by projecting the MWEs back on the source corpus.

As an example, consider the following two MWE patterns described by regular expressions over POS tags:

- `NounCompound` $\rightarrow$ `Noun Noun`$^{+}$
- `PhrasalVerb` $\rightarrow$ `Verb (Ignored`*`) Particle`

Figure 6.1 presents the results of applying different matching combinations to these patterns. Consider an input such as the one in Figure 6.1(a). By applying a non-overlapping contiguous approach to the noun compound identification and a gappy approach to the verb-particle construction, we may automatically identify two MWE candidates in the sentence. If we use the *longest* match distance for both patterns, we capture the whole nominal compound, but we go too far for the verb-particle construction (Figure 6.1(b)). The opposite happens if we use *shortest* match distance for both patterns, which works well for the verb-particle construction but does not capture the whole nominal compound (Figure 6.1(c)). By using different configurations for each type of MWE, we are able to identify the correct occurrences in the text (Figure 6.1(d)).

Figure 6.1: MWE-annotated output with different match distances.

(a) You threw those lab rat tissue samples out without thinking ?
(b) You threw those **lab rat tissue samples** out without thinking ?
(c) You threw those **lab rat** tissue samples out without thinking ?
(d) You threw those **lab rat tissue samples** out without thinking ?

6.1.2 Probabilistic identification model

In addition to the rule-based approach of MWE identification, we also consider a probabilistic model, in the form of linear-chain conditional random fields (CRFs) (LAFFERTY; MCCALLUM; PEREIRA, 2001). Under this approach, we construct a classifier that tags each input token based on whether it is independent or a part of an MWE.[2] The CRF is trained based on a set of observations $T = T_1 \dots T_n$, in which each observed input token T_i is paired up with a tag y_i. When performing predictions, the probability of a given output tag y_i for an input token T_i depends on the tag of its neighbor token (y_{i-1}), and on a set of features of the input $\phi(T)$. The feature values can come from any position of the input sequence, including the current token T_i.

We represent MWE identification as a tagging problem through the use of the Begin-Inside-Outside (BIO) encoding (RAMSHAW; MARCUS, 1995). In a BIO representation, each token T_i in the training corpus is annotated with a corresponding tag `B` (beginning of the MWE), `I` (inside MWE) or `O` (independent token, outside any MWE). In this scheme, MWEs must all be contiguous, and overlaps cannot be represented.[3]

6.2 Experimental setup

We instantiate multiple variants of the rule-based and probabilistic methods of MWE identification. We present below the configuration that we use for each method, as well as the annotated corpora on which they are evaluated. For both methods, we consider the task identification with and without a *compositionality* feature, derived from a lexicon of predicted compositionality scores, which we also describe below.

[2] The CRF tagger was trained with CRFSuite (OKAZAKI, 2007).

[3] Note however that more complex schemes could account for some level of discontinuity and overlap, such as the two-layer BB̃IĨOÕ representation of Schneider (2014).

6.2.1 Reference corpora

We perform the evaluation of the MWE identification models on two MWE-annotated corpora.

- For the English language: training, development and test data are the ones provided for the DiMSUM shared task (SCHNEIDER et al., 2016a). The sentences originally come from multiple English corpora: a corpus of online reviews (STREUSLE), two corpora containing Twitter data (Ritter and Lowlands), and a corpus built from TED Talks transcripts[4]. The resulting training corpus contains 4 800 sentences, and the test corpus contains 1 000 sentences. Every MWE annotation was reviewed by at least two annotators. The authors do not report the annotator agreement (SCHNEIDER et al., 2014b; SCHNEIDER et al., 2016a).

- For the French language: training, development and test data come from an adaptation of the French Treebank (FTB) (ABEILLé; CLéMENT; TOUSSENEL, 2003) from the SPMRL shared task on Statistical Parsing of Morphologically-Rich Languages (SEDDAH et al., 2013). The corpus consists of a collection of newspaper entries (*Le Monde*) from multiple domains, with a total of around 1 million tokens. It contains manually-validated lemmas, POS-tag annotations, syntactic information (ignored in this work) and a layer of MWE occurrence annotations.

In both corpora, we keep only MWEs representing nominal compounds. This is done through a pattern-based filtering on the MWE layer, using the mwetoolkit. The goal of this step is to filter out all MWEs that do not contain a noun (e.g. *by and large*), as well as MWEs that contain verbs (e.g. *give birth*). The resulting English test corpus has 254 MWEs, and the resulting French test corpus has 849 MWEs.

6.2.2 Compositionality lexicons

In both rule-based and probabilistic methods, we consider the use of a *compositionality* feature, which we derive from a lexicon of predicted compositionality scores. The lexicon itself was constructed through a type-based extraction of MWE candidates from the reference corpora. The extraction used a language-specific pattern. For English,

[4] The train/dev corpora did not contain sentences from TED. Only the blind test data did.

we allow adjective+noun pairs (e.g. *red wine*) as well as noun+preposition+noun (e.g. *cup of tea*), and combinations thereof (e.g. *president of the United States*). For French, we consider noun+adjective pairs (e.g. *vin blanc* (lit. *wine white*)), adjective+noun pairs (e.g. *longue durée*, 'long-term' (lit. *long duration*)), as well as noun+preposition+noun expressions (e.g. *mise à jour* 'update' (lit. *put to day*)), including combinations of these (e.g. *Journal officiel de la République Française* 'Official gazette of the French Republic' (lit. *Newspaper official of the Republic French*)).

For each language, we projected the extracted MWE candidates onto WaC corpora[5], and then constructed two DSMs instances (*w2v–sg* and *PPMI–thresh*), with the same setup as in Section 4.4, using lemma.w_1.d_{750}. For each DSM, we calculated the predicted compositionality score for each MWE candidate under the *uniform* strategy. This resulted in a total of four lexicons of compositionality (varying between two DSMs and two languages).

6.2.3 Rule-based identification

For each language, our baseline rule-based MWE identification algorithm considers 7 different rule configurations. Two of these rules are directly based on data from the training corpus, two are based on an approach of MWE identification based on POS-tag patterns, and two are based on the previously described compositionality lexicons (described in Section 6.2.2).

For the rules based on training data, annotated MWEs are extracted from the training corpus and then filtered. We keep MWE candidates whose proportion of annotated instances with respect to all occurrences in the training corpus is above a threshold τ, discarding the rest. For the selection of thresholds, we refer to Cordeiro, Ramisch and Villavicencio (2016c), where we considered thresholds $\tau \in \{0\%, 10\%, 20\%, \ldots, 100\%\}$, obtaining the best results for $\tau = 40\%$ (contiguous MWEs) and $\tau = 70\%$ (gappy MWEs).

The last step of rule-based identification consists in projecting the filtered list of MWE candidates on the test data, that is, we segment as MWEs the test token sequences that are contained in the lexicon extracted from the training data. These configurations are:

- TRAIN$_{\text{CONTIG}}$: Contiguous MWEs annotated in the training corpus at least once are

[5] We use the same corpora as in Section 4.3.

extracted and filtered with a threshold of $\tau = 40\%$. That is, we create a lexicon containing all contiguous lemma+POS sequences for which at least 40% of the occurrences in the training corpus were annotated (e.g. we keep the expression *last minute*, as it was annotated in $^5/_6 = 83\%$ of its occurrences in the training data). The resulting lexicon is projected on the test corpus using this rule: an MWE is deemed to occur if its component words appear contiguously in a sentence.

- TRAIN$_{\text{GAPPY}}$: Non-contiguous MWEs are extracted from the training corpus and filtered with a threshold of $\tau = 70\%$. The resulting MWEs are projected on the test corpus using the following rule: an MWE is deemed to occur if its component words appear sequentially with at most a total of 3 gap words in between them. This method is not used for French, as only contiguous MWEs were annotated in the corpus.

We also identify MWEs in the test corpus based on POS-tag patterns:

- PATTERN$_{\text{NOUN}}$: We collect candidate nominal compounds from the test corpus that never appear in the training corpus, and project them back on the test corpus. For English, we focus on contiguous noun+noun sequences (e.g. *car wash*), as they are the most prevalent in the DiMSUM corpus. For French, we consider contiguous noun+adjective pairs.[6] As the French corpus does not distinguish common nouns from proper nouns, both are included as part of this method.
- PATTERN$_{\text{PROPN}}$: The English corpus distinguishes common nouns and proper nouns through their POS tag. In this method, we annotate sequences of two or more tokens POS tagged as proper nouns (`PROPN`), in an effort to identify named entities such as *New York City*. We do not consider any thresholds, as named entities are sparse and most occurrences from training do not appear in test.

For each language, we also consider two methods of MWE identification based on compositionality lexicons (Section 6.2.2). We annotate as MWE every contiguous occurrence of an entry in the compositionality lexicon, as long as its compositionality score is under a given threshold ($\text{CS}_\beta \leq$ threshold). We consider thresholds between 0 (most restrictive, eliminates almost all MWEs from the lexicon) and 1 (most permissive, keeps almost all MWEs in the lexicon).

[6] Syntactic structures involving combinations of nouns, adjectives and prepositions are rarely annotated in this corpus.

- COMP$_{\text{W2V}}$: Uses the compositionality lexicon built from *w2v–sg*.
- COMP$_{\text{PPMI}}$: Uses the compositionality lexicon built from *PPMI–thresh*.

6.2.4 Probabilistic identification

We consider the following sets of features $\phi(T)$:

- CTX: This is a set of contextual features which corresponds to the BEST$_2$ set from Scholivet, Ramisch and Cordeiro (2017), without the association measures. The feature set contains 21 single-token, 2-gram and 3-gram features (involving surface-form, lemma and POS tag of tokens). It also includes features indicating: whether the current token has a hyphen, whether it has a digit, and whether it is in upper-case form. We refer to the paper for an in-depth feature analysis, as well as a broader evaluation of this model for all categories of MWEs in the French corpus.
- AM: This set of features contains four association measures: PMI, MLE, Student's t and log-likelihood (see Section 2.2.3). These are the association measures that were found to be the most impactful in Scholivet, Ramisch and Cordeiro (2017), when evaluating corpora with multiple categories of MWEs.
- COMP: This set of features is based on the MWE scores from the compositionality lexicons. We designate the set of features derived from *w2v–sg* as COMP$_{\text{W2V}}$, and the one derived from *PPMI–thresh* as COMP$_{\text{PPMI}}$.

Different methods of CRF modeling may or may not be able to accurately represent continuous features. In this work, we circumvent possible limitations by quantizing every numerical score (obtained in AM and COMP) using a uniform distribution; i.e. we assign an equal number of MWEs to 5 different bins based on their numerical scores. We leave the evaluation of continuous CRF models for future work (HUANG; XU; YU, 2015).

6.3 Results

For each reference corpus, we evaluate the MWE identification models on the *development* part under a variety of setups, as described in Section 6.2. We present the results below.

6.3.1 Rule-based identification: baseline

We start with an analysis of the baseline results obtained by the rule-based MWE identifier for the English corpus. Table 6.1 presents the individual score obtained by each rule. At a first glance, the most promising rules seem to be TRAIN$_{\text{CONTIG}}$ and PATTERN$_{\text{PROPN}}$, both of which obtain a high level of precision. The rule TRAIN$_{\text{GAPPY}}$ also obtains a high precision, but it does not capture many occurrences of MWEs, obtaining low recall.

Table 6.1: Baseline results for rule-based MWE identifier (English dataset).

Rules	Precision	Recall	F_1
TRAIN$_{\text{CONTIG}}$	**.843**	.232	.364
TRAIN$_{\text{GAPPY}}$	.750	.012	.023
PATTERN$_{\text{PROPN}}$	.750	.272	**.399**
PATTERN$_{\text{NOUN}}$	.315	.181	.230

We then consider the accuracy of MWE identification when multiple rules are combined. In particular, we fix the highest-ranking rule TRAIN$_{\text{CONTIG}}$, and we consider combinations involving the other rules. Table 6.2 presents the new results. The addition of the rule TRAIN$_{\text{GAPPY}}$ does manage to slightly improve the recall of TRAIN$_{\text{CONTIG}}$, but at the expense of a considerable decrease in the precision. The addition of the rule PATTERN$_{\text{PROPN}}$ improves both precision and recall, reflecting the fact that named entities are sparse, and most occurrences were not seen in training data. Similarly, the addition of the rule PATTERN$_{\text{NOUN}}$ does improve recall, suggesting that many occurrences of noun+noun compounds were not seen in the training data. However, many of these predictions are spurious (i.e. they refer to productive combinations of nouns, such as *dinner plate*, which were not annotated), and thus the precision of these 2 combined rules is sub-par. The same behavior can be seen in the last line, where we consider all rules but TRAIN$_{\text{GAPPY}}$. The combination achieves a considerably higher recall, but at the expense of a reduction in the precision.

Table 6.2: Combined baselines for rule-based MWE identifier (English dataset).

Rules	Precision	Recall	F_1
TRAIN$_{\text{CONTIG}}$ + TRAIN$_{\text{GAPPY}}$	**.831**	.232	.363
TRAIN$_{\text{CONTIG}}$ + PATTERN$_{\text{PROPN}}$	.783	.484	.599
TRAIN$_{\text{CONTIG}}$ + PATTERN$_{\text{NOUN}}$	.491	.413	.449
TRAIN$_{\text{CONTIG}}$ + PATTERN$_{\text{PROPN}}$ + PATTERN$_{\text{NOUN}}$	.561	**.665**	**.609**

We similarly consider the baseline results obtained by the rule-based MWE identifier for the French corpus. Table 6.3 presents the individual score obtained by each rule. Differently from the English corpus, many of the MWEs in the French development set had a counterpart in the training set, which contributed to a TRAIN$_{\text{CONTIG}}$ recall of more than 60% of the occurrences. The pattern-based rule PATTERN$_{\text{NOUN}}$ does find a modest amount of new MWEs, but at the cost of a very low precision. Therefore, when both rules are combined, the result is quite a bit lower than the one obtained for TRAIN$_{\text{CONTIG}}$ alone.

Table 6.3: Baseline results for rule-based MWE identifier (French dataset).

Rules	Precision	Recall	F_1
TRAIN$_{\text{CONTIG}}$	**.862**	.684	**.763**
PATTERN$_{\text{NOUN}}$	.081	.107	.092
TRAIN$_{\text{CONTIG}}$ + PATTERN$_{\text{NOUN}}$	.381	**.792**	.515

6.3.2 Rule-based identification: compositionality scores

The rule-based method can also be applied with an external lexicon of MWEs. We consider the two lexicons described in Section 6.2.2: COMP$_{\text{W2V}}$ and COMP$_{\text{PPMI}}$. These lexicons associate MWE candidates with an automatically-calculated compositionality score. We consider multiple variants of the MWE identification model by applying different thresholds on what score constitutes an idiomatic MWE. Lower thresholds should improve precision (as they only allow annotation of highly non-compositional cases), while reducing the recall due to their restrictiveness.

Table 6.4 presents the results obtained for the rule-based system with different thresholds of compositionality scores in COMP$_{\text{W2V}}$. As expected, lower thresholds are associated with a lower recall in both languages. The precision, on the other hand, presents an unexpected behavior: the lowest precision (indicated through † on the table) is not associated with the more permissive threshold of $\text{CS}_\beta \leq 1.000$. In fact, in the case of the French corpus, the precision falls monotonically as we consider stricter thresholds. In the case of the English corpus, the lowest precision is associated with a middle-range threshold of $\text{CS}_\beta \leq 0.200$, but note that other more restrictive thresholds (indicated through $^?$ on the table) are much less reliable, as the number of MWEs predicted by the system is very low ($\text{CS}_\beta \leq 0.100$ has 37 predictions, $\text{CS}_\beta \leq 0.050$ has 10 predictions, and

$\text{CS}_\beta \leq 0.000$ has 4 predictions). In the case of French, the lowest threshold of $\text{CS}_\beta \leq 0.000$ produces 85 predictions.

Table 6.4: Results for rule-based MWE identifier using COMP$_{\text{W2V}}$.

Threshold	English corpus			French corpus		
	Precision	Recall	F$_1$	Precision	Recall	F$_1$
$\text{CS}_\beta \leq 1.000$	.152	**.185**	**.167**	**.245**	**.482**	**.325**
$\text{CS}_\beta \leq 0.500$	.149	.181	.164	**.245**	**.482**	**.325**
$\text{CS}_\beta \leq 0.200$	.098†	.055	.070	.176	.239	.203
$\text{CS}_\beta \leq 0.100$	.216$^{?}$	.031	.055	.132	.111	.120
$\text{CS}_\beta \leq 0.050$	.200$^{?}$	.008	.015	.128	.029	.048
$\text{CS}_\beta \leq 0.000$	**.250**$^{?}$	.004	.008	.012†	.001	.002

We then consider whether a similar behavior can be observed when using compositionality scores from COMP$_{\text{PPMI}}$. Table 6.5 presents the results obtained for the rule-based system under these scores. As seen in Section 5.6.3, *PPMI–thresh* scores tend to be lower than in other DSMs, which explains why the scores obtained for high thresholds are almost identical. As in the case of COMP$_{\text{W2V}}$ above, lower thresholds are associated with a lower recall in both languages. Moreover, the lowest values of precision are once again associated with lower thresholds, suggesting that this is a consistent property of both datasets. These precision scores are more reliable than the ones from COMP$_{\text{W2V}}$ (for English, $\text{CS}_\beta \leq 0.000$ has 77 predictions, $\text{CS}_\beta \leq 0.005$ has 177 predictions and $\text{CS}_\beta \leq 0.010$ has 261 predictions; while for French, $\text{CS}_\beta \leq 0.000$ has 513 predictions, $\text{CS}_\beta \leq 0.005$ has 1648 predictions and $\text{CS}_\beta \leq 0.010$ has 2155 predictions).[7]

Table 6.5: Results for rule-based MWE identifier using COMP$_{\text{PPMI}}$.

Threshold	English corpus			French corpus		
	Precision	Recall	F$_1$	Precision	Recall	F$_1$
$\text{CS}_\beta \leq 1.000$	**.118**	**.240**	**.158**	**.136**	**.522**	**.215**
$\text{CS}_\beta \leq 0.500$	**.118**	**.240**	**.158**	**.136**	**.522**	**.215**
$\text{CS}_\beta \leq 0.200$	**.118**	**.240**	**.158**	.135	.519	**.215**
$\text{CS}_\beta \leq 0.100$	**.118**	.236	**.158**	.134	.514	.213
$\text{CS}_\beta \leq 0.050$	.100	.185	.130	.124	.461	.196
$\text{CS}_\beta \leq 0.020$	.094	.138	.112	.090	.284	.137
$\text{CS}_\beta \leq 0.010$	.077†	.079	.078	.060	.153	.087
$\text{CS}_\beta \leq 0.005$	.090	.063	.074	.053†	.104	.070
$\text{CS}_\beta \leq 0.000$	.104	.031	.048	.066	.040	.050

[7] Note that, due to the nature of PPMI vs PMI, $\text{CS}_\beta \leq 0$ is equivalent to $\text{CS}_\beta = 0$ for COMP$_{\text{PPMI}}$.

A possible explanation to the behavior observed above would be that these datasets also annotate collocations along with idiomatic MWEs. We investigate this hypothesis through a manual annotation of all 309 MWE candidates identified by $\textsc{comp}_{\text{w2v}}$. We classify each MWE occurrence in one of these 4 categories:

- Productive expression: when the words are combined in a fully productive manner (e.g. *nice car*). In a productive expression, any of the elements can be replaced by a similar word without any loss of meaning or increased markedness (e.g. *nice car* → *nice airplane*, *cool car*).

- Collocation: when the choice of words is conventionalized, but is still compositional (e.g. *test results*). In these cases, changes in word order and replacement by similar words or synonyms is possible, but has a distinctive markedness (e.g. the expression *results from the tests* still refers to a similar concept, but is not the preferred way of referring to *test results*).

- Idiomatic expression: when the whole expression is idiomatic, with at least one of the words not contributing to a literal sense. This includes crystallized metaphors (e.g. *extra mile*) and proper names (e.g. *New Jersey*).

- Other: for all other cases. This includes errors of POS tag (e.g. *wind blows* classified as `NOUN+NOUN`), adjacent sequences of words that do not form a phrase (e.g. *home from work*), and cases in which the intended meaning was considered hard to judge even in the context of the original sentence (e.g. an occurrence of *good sport*).

For the first three categories, we consider the fraction of their occurrences for each threshold of predicted compositionality score (Table 6.6). In particular, we consider the lines with $\text{CS}_\beta \leq 0.5$ and $\text{CS}_\beta \leq 0.2$. In the case of productive MWEs, for these two thresholds, we can see that they present a similar rate of occurrence among the different levels of compositionality score (around 55%). The rate of occurrence of idiomatic expressions has a slight variation, with a higher rate for $\text{CS}_\beta \leq 0.2$, as expected. Note however the case of collocations: they present a difference of more than 7% between the two thresholds.[8] This suggests that the difference between annotations is related to the fact that the dataset includes many collocations, which is precisely what we filter out when we use the compositionality scores.

[8] Note that the scores for $\text{CS}_\beta \leq 0.1$ and for lower thresholds are less trustworthy due to the smaller amount of predicted MWEs.

Table 6.6: Classification of MWE candidates identified by COMP$_{\text{W2V}}$.

Threshold	Productive	Collocation	Idiomatic
$\text{CS}_\beta \leq 1.000$	170/309 = 55.0%	61/309 = 19.7%	28/309 = 9.1%
$\text{CS}_\beta \leq 0.500$	170/308 = 55.2%	60/308 = 19.5%	28/308 = 9.1%
$\text{CS}_\beta \leq 0.200$	78/143 = 54.5%	17/143 = 11.9%	17/143 = 11.9%
$\text{CS}_\beta \leq 0.100$	17/37 = 45.9%	6/37 = 16.2%	8/37 = 21.6%
$\text{CS}_\beta \leq 0.050$	1/10 = 10.0%	3/10 = 30.0%	4/10 = 40.0%
$\text{CS}_\beta \leq 0.000$	0/4 = 0.0%	1/4 = 25.0%	3/4 = 75.0%

6.3.3 Probabilistic identification

The previous section has considered the use of predicted compositionality scores as part of a rule-based system of MWE identification. We considered different thresholds on these scores, and we showed that higher thresholds produced better results. An analysis of the data suggested that higher scores may also be associated with higher rates of annotation (as they may indicate the presence of collocations). Rather than pursuing the rule-based approach under different ranges of threshold, we consider a different approach: using a probabilistic classifier which considers these scores as features for the prediction of MWE occurrences.

We construct a CRF classifier based on different kinds of features. The features, described in Section 6.2.4 can be purely contextual (CTX), or involve statistical association measures (AM), or come from the previously defined lexicon of compositionality scores (COMP$_{\text{W2V}}$ and COMP$_{\text{PPMI}}$).

Table 6.7 presents the results obtained for the evaluation against the English reference corpus. Overall, there is a notable improvement in both precision and recall when additional features are considered beyond CTX. Concerning the recall, it can be seen that the use of association measures is redundant with the scores from COMP$_{\text{W2V}}$ (as the recall in the second and third lines are identical, and the one in the fourth line is only slightly higher). The scores from COMP$_{\text{PPMI}}$, on the other hand, contribute to a higher improvement in the recall than the association measures. Moreover, this feature completely subsumes AM, as can be seen from the fact that the recall in the last two lines is the same. Regarding the precision, COMP scores are considerably higher than CTX alone, and to a certain extent CTX + AM as well. The highest F_1 score of .710 is better than the baseline .609 from Section 6.3.1 by +10.1 percentage points.

Table 6.7: Results for probability-based MWE identifier (English dataset).

Features	Precision	Recall	F_1
CTX	.758	.567	.649
CTX + AM	.789	.602	.683
CTX + COMP_{W2V}	.797	.602	.686
CTX + COMP_{W2V} + AM	.798	.606	.689
CTX + $\text{COMP}_{\text{PPMI}}$	.811	**.626**	.707
CTX + $\text{COMP}_{\text{PPMI}}$ + AM	**.820**	**.626**	**.710**

Table 6.8 presents the results obtained for the evaluation against the French reference corpus. As in the case of the English data above, the probabilistic method improves both precision and recall when we consider features beyond CTX. However, the improvement for the French corpus is much less pronounced. For recall, the greatest improvement happens with the addition of association measures (with +2.2 percentage points), with COMP providing a smaller effect. In the case of the precision, only the addition of both AM and $\text{COMP}_{\text{PPMI}}$ provided an improvement, and it was considerably smaller than the one seen for English. Differently from what was observed for English, we see that the highest F_1 score obtained for the probabilistic method for French (.736) is unable to surpass the rule-based baseline of .763 obtained through $\text{TRAIN}_{\text{CONTIG}}$ in Section 6.3.1. Further work should investigate this behavior through an analysis of the French corpus annotations.

Table 6.8: Results for probability-based MWE identifier (French dataset).

Features	Precision	Recall	F_1
CTX	.817	.636	.715
CTX + AM	.818	.658	.730
CTX + COMP_{W2V}	.802	.649	.717
CTX + COMP_{W2V} + AM	.812	**.664**	.731
CTX + $\text{COMP}_{\text{PPMI}}$	.817	.650	.724
CTX + $\text{COMP}_{\text{PPMI}}$ + AM	**.826**	**.664**	**.736**

6.4 Summary

This chapter evaluated the use of automatically predicted compositionality scores as features in the task of MWE identification in two corpora. We started with a rule-based baseline, where a contiguous identification of lemmas seen in the training corpus was found to obtain high precision (higher than .8) for both languages. In the case of the

English data, the recall from this method alone was weak (.23), while the recall obtained for French was considerably higher (.68), making this a particularly hard baseline to beat in the latter language. Indeed, while overall English results could be improved with the addition of a pattern for matching proper nouns (and the precision score could further be improved through noun–noun patterns), no F_1 improvement was found for the French corpus. The highest-scoring set of rules in this baseline obtained an $F_1 = .609$ for English and $F_1 = .763$ for French.

We then considered the application of compositionality scores directly as part of the rule-based method of MWE identification. We collected a lexicon of potential MWEs (based on NC syntactic patterns) and calculated their compositionality scores. Different thresholds were then applied on the compositionality scores, with the least-compositional MWEs being automatically annotated in the corpus according to the rule-based method. The results we obtained were consistent across the two languages and the two lexicons of compositionality: an overall low precision of identification, which surprisingly drops more harshly when stricter thresholds of idiomaticity are considered. We presume that this effect is caused by a high rate of annotation of collocations, which tend to have higher compositionality scores.

Another method of identifying MWEs would be through a probabilistic approach considering multiple features. We evaluate the performance of a CRF trained on different sets of features, grouped as: lexical features, association measures, and compositionality scores. The results were highly corpus-dependent: while both association measures and compositionality scores contributed to higher values of F_1 for English, the improvement in French results was considerably weaker. Moreover, while the highest CRF scores for English (.710) convincingly beat the baseline above (by +10.1 percentage points), the highest scores for French (.736) are actually *lower* than the baseline of purely identifying all MWEs seen in training data (by −2.7 points). Further analysis would be required to understand this discrepancy.

Concerning the hypothesis $h_{\text{pred-comp} \rightarrow \text{ident-accur}}$ that compositionality scores can have a positive effect on the accuracy of MWE identification, we have obtained mixed results. In the case of rule-based methods, we found no improvement in prediction with more restrictive scores of compositionality (i.e. with a lower threshold). In fact, results suggest that the English corpus contains a high amount of annotated collocations, which would explain why a more strict threshold does not improve on the results. However, when we considered a CRF, the F_1 score for MWE identification for the English and French corpus did present an increase (in particular for the former language). In order to evaluate this hypothesis in a more favorable setting, future work should consider corpora that have been annotated particularly with idiomatic MWEs in mind. For example, the corpus for the PARSEME shared task contains an MWE category called *verbal idiom*, which is guaranteed to refer to expressions that humans have judged as idiosyncratic (SAVARY et al., 2017b), and could be a more close fit for the evaluation of this hypothesis.

One question that can be raised from the results in this chapter is whether association measures can be helpful in the identification of MWEs. While we did not find any correlation between human judgments of compositionality and a measure of conventionalization (see $\mathrm{h}_{\mathrm{idiom}\,\approx\,\mathrm{distr.convent}}$ on page 63), note that the distinction being done on the task of compositionality prediction is between compositional and idiomatic MWEs, while the annotations on these corpora might tend toward a distinction between fully productive expressions and any kind of conventionalized expressions (i.e. any kind of MWEs, in the broadest sense). In this case, we hypothesize that the improvements in MWE identification caused by association measures is related to their ability of capturing conventionalization. Future analysis is still needed to verify whether this interpretation is correct. In particular, if only the idiomatic MWEs in the corpora are taken into account (i.e. compositional cases are filtered out), we do not expect association measures to contribute with an improvement of MWE detection scores.

The MWE identification techniques presented in this chapter were implemented and are currently available as part of the mwetoolkit. A description of the implementation as well as further results can have been published as Cordeiro, Ramisch and Villavicencio (2016c), Scholivet, Ramisch and Cordeiro (2017).

7 CONCLUSIONS

This thesis has proposed a framework of multiword expression compositionality prediction, and has investigated the impact of several variables in the accuracy of the predictions. The predictive model is based on the manipulation of distributional semantic models, i.e. vectorial representations of the meaning of words and MWEs. We have presented three new datasets of human-rated compositionality scores, in three different languages, and evaluate the developed framework using these resources. Finally, we also consider the use of predicted compositionality scores as features in the task of MWE identification.

Both the construction of the datasets and the subsequent evaluations of the predictive model are associated with a set of hypotheses. Table 7.1 summarizes these hypotheses and provides a reference to the page in which they have been evaluated.

In the following section, we present the main contributions from this thesis, including an overview of our findings for the evaluated hypotheses. We then present some perspectives of future work.

7.1 Contributions

The contributions of this thesis can be summarized as follows:

- Three new human-rated datasets of compositionality scores.
- An analysis of the new datasets with regards to score distribution and correlation with human variables.
- A new framework of compositionality prediction, which relies on a systematization of DSMs and parameters.
- A large-scale multilingual evaluation of the compositionality prediction framework under a variety of settings.
- An extrinsic evaluation of predicted compositionality scores in the task of MWE identification.

Table 7.1: Hypotheses evaluated in this thesis.

Hypothesis	Sub-hypothesis	Evaluation
$h_{\text{idiom} \approx \text{distr}}$	$h_{\text{idiom} \approx \text{distr.convent}}$	Pages 63, 112
	$h_{\text{idiom} \approx \text{distr.freq}}$	Pages 63, 109
$h_{\text{accur} \leftarrow \text{MWE}}$	$h_{\text{accur} \leftarrow \text{MWE.diffic}}$	Page 114
	$h_{\text{accur} \leftarrow \text{MWE.idiom}}$	Page 114
	$h_{\text{accur} \leftarrow \text{MWE.convent}}$	Page 112
	$h_{\text{accur} \leftarrow \text{MWE.freq}}$	Page 110
$h_{\text{accur} \leftarrow \text{DSM}}$	$h_{\text{accur} \leftarrow \text{DSM.window}}$	Page 85
	$h_{\text{accur} \leftarrow \text{DSM.dims}}$	Page 87
$h_{\text{accur} \leftarrow \text{corpus}}$	$h_{\text{accur} \leftarrow \text{corpus.wordform}}$	Page 88
	$h_{\text{accur} \leftarrow \text{corpus.size}}$	Page 90
	$h_{\text{accur} \leftarrow \text{corpus.parallel}}$	Page 92
h_{strat}	$h_{\text{strat.partial-info}}$	Page 95
	$h_{\text{strat.maxsim}}$	Page 96
	$h_{\text{strat.geom}}$	Page 99
$h_{\text{pred-comp} \approx \text{comp}}$	$h_{\text{pred-comp} \approx \text{comp}}$	Chapter 5
$h_{\text{pred-comp} \rightarrow \text{ident-accur}}$	$h_{\text{pred-comp} \rightarrow \text{ident-accur}}$	Chapter 6

Many of the results presented in this thesis have also been presented in peer-reviewed publications. We refer back to Section 1.4 (page 19) for the complete list of publications. As for the contributions that have been presented in this thesis, we described them in more detail below.

Chapter 3 presented the construction of three datasets of human-rated MWE compositionality scores. The datasets encompass three languages (English, French and Portuguese), and is the first dataset of MWE compositionality for two of these languages. This resource is freely available, and can be used for evaluating and training techniques that involve some type of semantic processing, such as lexical substitution and text simplification.

We also analyzed the constructed datasets, whose scores were found to follow a uniform distribution. Moreover, the three datasets were found to have comparable levels of difficulty of annotation. We have evaluated the hypothesis that MWE idiomaticity was correlated with distributional characteristics ($h_{\text{idiom} \approx \text{distr}}$). In particular, we considered the correlation with an estimator of conventionalization ($h_{\text{idiom} \approx \text{distr.convent}}$), which was shown not to be statistically significant. We also considered the correlation with the frequency ($h_{\text{idiom} \approx \text{distr.freq}}$), which turned out to be the opposite of what one would expect: higher-frequency MWEs are actually more likely to be compositional, with lower-frequency ones being more likely to be idiomatic. Some of these results were also presented in publications (CORDEIRO; RAMISCH; VILLAVICENCIO, 2016a; WILKENS et al., 2017; RAMISCH et al., 2016).

Chapter 4 presents a systematization of different DSMs and their parameters along a common set of axes, which can be used to compare multiple distributional representations in a multilingual setting. This chapter also proposes a framework of compositionality prediction that can take into account all of these configurations. The framework has also been described in a publication (CORDEIRO; RAMISCH; VILLAVICENCIO, 2016b).

The framework above was then used in a large-scale intrinsic evaluation of multiple combinations of DSM and corpus parameters (Chapter 5). Here too we investigate the correlation between MWE idiomaticity and distributional characteristics ($\mathrm{h}_{\text{idiom} \approx \text{distr}}$). As in the case of human judgments above, a correlation with corpus frequency was also observed for model predictions. Moreover, while human judgments did not correlate with a measure of conventionalization, we did find a mild correlation in some system scores.

One of the goals of the large-scale evaluation was to determine the factors that influence the accuracy of model predictions. One of the hypothesis we considered was that the accuracy should be influenced by MWE-specific characteristics ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE}}$). We found that both idiomaticity ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE.idiom}}$) and difficulty in human judgments of compositionality ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE.diffic}}$) are associated with lower-quality predictions. On the other hand, MWE frequency ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE.freq}}$) and conventionalization ($\mathrm{h}_{\text{accur} \leftarrow \text{MWE.convent}}$) did not show clear signs of correlation with model accuracy.

We also evaluate different variations of DSMs. Our hypothesis is that DSM-specific configuration should play a crucial role in the accuracy of the results ($\mathrm{h}_{\text{accur} \leftarrow \text{DSM}}$). While we do find a high variety in the accuracy across different DSMs, the results for the two DSM parameters we considered were somewhat underwhelming. We found that a higher number of dimensions would consistently contribute to a mild improvement in the accuracy ($\mathrm{h}_{\text{accur} \leftarrow \text{DSM.dims}}$), but no cross-lingual and unified recommendations could be attained regarding the variation in context-window sizes ($\mathrm{h}_{\text{accur} \leftarrow \text{DSM.window}}$).

Along with the impact from DSM-specific parameters, we also hypothesized an influence of corpus-specific parameters in the accuracy of results ($\mathrm{h}_{\text{accur} \leftarrow \text{corpus}}$). Indeed, the results confirm that stopword removal and lemmatization are both important steps of corpus preprocessing for this task, especially in the case of languages that are morphologically richer than English. Moreover, the use of POS tags does not contribute to a higher quality in the representation of word vectors for this task, possibly due to the fact that it increases the sparsity in co-occurrence counts ($\mathrm{h}_{\text{accur} \leftarrow \text{corpus.wordform}}$). An analysis of different corpus sizes also showed that these may have a direct impact in the accuracy of results ($\mathrm{h}_{\text{accur} \leftarrow \text{corpus.size}}$). Moreover, a proposed technique of parallel predictions was shown to perform equivalently to whole-corpus predictions, while allowing for the better utilization of computational resources ($\mathrm{h}_{\text{accur} \leftarrow \text{corpus.parallel}}$).

Concerning the predictive model itself, we have considered six different strategies for deriving the compositionality scores. Our hypothesis is that different strategies would provide a different view into the data, with some strategies being more accurate than others ($\mathrm{h}_{\text{strat}}$). We evaluated two additive strategies that are commonly used in the literature, but that had never been compared, and we concluded that their results are mostly equivalent to each other. Two other strategies considered only one of the words in the NCs (head or modifier). As expected, their accuracy suffered due to the limited information ($\mathrm{h}_{\text{strat.partial-info}}$). We then evaluated two proposed strategies. We confirmed the hypotheses that the *maxsim* strategy is better suited for compositional MWEs ($\mathrm{h}_{\text{strat.maxsim}}$), while the *geom* strategy optimizes towards idiomatic cases ($\mathrm{h}_{\text{strat.geom}}$). However, in both cases, results were quite similar to the standard additive strategies, suggesting that the impact of strategy choice is not as strong as previously thought.

All of the experiments presented in Chapter 5 revolve around a common hypothesis: model predictions are correlated with human-rated MWE compositionality ($\mathrm{h}_{\text{pred-comp} \approx \text{comp}}$). Indeed, the variety of results obtained in this thesis all suggest that this hypothesis is true. While the correlation obtained for predictions using the worst DSM and corpus configurations may be considered weak, we have shown that the appropriate configurations are able to consistently produce predictions of compositionality that highly correlate with human judgments for a variety of datasets across multiple languages. The identification of patterns in the large space of more than 8 thousand configurations is one of the most salient contributions of this thesis. Some of the results on the evaluation of compositionality prediction were also published in an ACL paper (CORDEIRO et al., 2016), and we are currently working on another paper to be submitted to a journal.

Finally, one of the contributions of this thesis is the application of predicted compositionality scores to the task of MWE identification (Chapter 6). The goal was to evaluate the hypothesis $\mathrm{h}_{\text{pred-comp} \rightarrow \text{ident-accur}}$, which predicts an improvement of MWE identification with the use of predicted compositionality scores as internal features. In a rule-based algorithm, compositionality scores were not found to be a good feature for the identification of annotated MWE occurrences, likely due to the presence of collocations along with idiomatic MWEs in the annotation. We then considered a probabilistic model of identification, in which the results were mixed: while compositionality scores significantly improved the results over rule-based and probabilistic baselines for the English corpus, no such improvement was found for the French corpus. Further analysis of this phenomenon is left for future work. Intermediary results for this task have been published in Cordeiro, Ramisch and Villavicencio (2016c), and sent for publication in Scholivet, Ramisch and Cordeiro (2017).

7.2 Future work

Concerning the research on *compositionality datasets*, we envisage the extension of the dataset for each of the languages to allow better inter-language comparability (e.g. `EN` *red wine* and its translations `FR` *vin rouge* and `PT` *vinho tinto*). We also consider the collection of compositionality judgments for MWEs in additional languages, ideally from different language families for a broader generalization of results.

A different direction is to augment the dataset with judgments of similarity between compounds sharing the same head. For example, we can ask people to judge the similarity of the word *case* in the expression *nut case* against the word *case* in similar expressions with high PMI, such as *criminal case*, *special case*, *exceptional case*, *upper case* and *business case*. This judgment could also be extended to all pairs of expressions, which would allow for semantic clusters (and where clusters of a single expression could be taken as evidence of idiomaticity). This approach could steer some of the research on compositionality in the direction of lexical similarity, which is commonly used for the evaluation of DSMs in the case of single words. It would also allow further investigation of polysemy in the case of collocations sharing the same head (MOLDOVAN et al., 2004; KIM; BALDWIN, 2013). Crucially, it would allow us to peek into the DSM representation of similarly-looking compounds and to identify the ways in which the vectors of these expressions denote their difference in idiomaticity.

This thesis presents an extensive evaluation of an additive model of compositionality prediction using the constructed datasets. Similar evaluations could be done for other predictive techniques in the state of the art, such as the work of Salehi, Cook and Baldwin (2015). The examination of these works in the context of our multilingual datasets would provide a more solid indication of their accuracy. Moreover, outstanding cross-language differences between such results and the results found in this thesis would provide further directions of investigation.

Regarding the results obtained in our *compositionality prediction* methods, the highest-scoring configurations in this thesis achieved reasonably high correlation with human predictions. Nevertheless, some of the predicted MWE scores were diametrically opposite to the average of human judgments. A cross-DSM analysis of vector representations for these MWEs could reveal whether this behavior stems from deficiencies in the underlying DSM vector representation (e.g. the fact that all DSMs considered could only represent a single meaning per word). If this is the case, modifications in the DSM could be investigated so as to prevent the occurrence of such discrepant score predictions. The construction of a dataset of compound head similarity such as the one suggested above could facilitate the discovery of these DSM weak points.

As for the compositionality prediction methods themselves, we plan on examining the use of a voting scheme for combining the output of complementary DSMs. Moreover, we also plan on combining additional sources of information for building the models, such as multilingual lexicons or translation data (SALEHI; COOK; BALDWIN, 2014), to improve even further the compositionality prediction. We would also like to propose and evaluate more sophisticated compositionality functions that take into account the unbalanced contribution of individual words to the global meaning of a compound. This could be done e.g. through a combination of the *maxsim* and *geom* strategies proposed in this thesis (either on the level of the strategy itself, or in the form of an ensemble method that combines the predictions of multiple strategies).

This thesis has employed predicted compositionality scores in an *application* of MWE identification. We considered a rule-based and a probabilistic model, both of which we evaluated under a base configuration as well as in two configurations involving compositionality scores. For the probabilistic model, technical considerations required the quantization of the predicted scores for a categorical interpretation. The specific quantization used may have greatly limited the results, and future works should consider different schemes of quantization. Alternatively, this problem could be solved through the use of neural networks, which can appropriately deal with real-valued data.

Considering the results obtained by the application of MWE identification, we see that compositionality scores significantly contribute to better accuracy in the case of the English corpus, but has a less pronounced effect on the French corpus. Crucially, the probabilistic method fares worse than the baseline which identifies only MWEs seen in the training data for the French language. Future work would be needed to investigate this difference between the results for the two languages.

Finally, we also consider other applications of compositionality scores. In particular, we would like to incorporate the collected scores into a machine translation system, as an indication of whether an expression should be translated as a single indivisible unit. We also envisage the application of predicted MWE compositionality scores in MWE-aware parsers, extending the approach used in previous work on multiword prepositions (NASR et al., 2015; CONSTANT; NIVRE, 2016; WASZCZUK; SAVARY; PARMENTIER, 2016).

For the task of MWE identification, we would like to explore context-based definitions of compositionality scores. We would also like to evaluate our framework on verbal MWEs, such as the ones annotated for the PARSEME shared task (SAVARY et al., 2017b). Verbal MWEs are an understudied topic in the literature, and present some challenges that were not present in the case of the nominal compounds we used in this thesis. In particular, verbal MWEs can have extremely rigid or flexible morphosyntactic characteristics[1], and can often present discontinuities (e.g. *take [something] into account*). Work on verbal MWEs could be pursued in the context long-term research projects.

[1] For example, compare the rigid expression *bite me!*, which does not even allow the inflection of the verb, with the expression *pay a visit*, which even allows a change in word order in the passive voice.

REFERENCES

ABEILLé, A.; CLéMENT, L.; TOUSSENEL, F. Building a treebank for french. In: ABEILLé, A. (Ed.). **Treebanks: building and using parsed corpora**. Dordrecht, The Netherlands: Kluwer academic publishers, 2003. p. 165–168.

AGRAWAL, S.; AGGARWAL, A. et al. Hybrid approach: A solution for extraction of domain independent multiword expression. **Int J. of Technology Innovations and Research**, v. 5, 2013.

ARTSTEIN, R.; POESIO, M. Inter-coder agreement for computational linguistics. **Computational Linguistics**, v. 34, n. 4, p. 555–596, 2008. ISSN 0891-2017.

BALDWIN, T.; KIM, S. N. Multiword expressions. In: **Handbook of Natural Language Processing, Second Edition.** [S.l.: s.n.], 2010. p. 267–292.

BALDWIN, T.; VILLAVICENCIO, A. Extracting the unextractable: A case study on verb-particles. In: **Proceedings of CoNLL 2002**. ACL, 2002. (COLING-02), p. 1–7. Available from Internet: <http://dx.doi.org/10.3115/1118853.1118854>. Accessed: Jan 16, 2018.

BANNARD, C.; BALDWIN, T.; LASCARIDES, A. A statistical approach to the semantics of verb-particles. In: **Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (MWE '03), p. 65–72. Available from Internet: <http://dx.doi.org/10.3115/1119282.1119291>. Accessed: Jan 16, 2018.

BANNARD, C. J. **Acquiring phrasal lexicons from corpora**. Thesis (PhD) — University of Edinburgh, 2006.

BARONI, M. et al. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. **Language resources and evaluation**, Springer, v. 43, n. 3, p. 209–226, 2009.

BARONI, M.; DINU, G.; KRUSZEWSKI, G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 238–247. Available from Internet: <http://www.aclweb.org/anthology/P14-1023>. Accessed: Jan 16, 2018.

BARONI, M.; LENCI, A. Distributional memory: A general framework for corpus-based semantics. **Computational Linguistics**, MIT Press, v. 36, n. 4, p. 673–721, 2010.

BARONI, M.; ZAMPARELLI, R. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In: **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**. Cambridge, MA: Association for Computational Linguistics, 2010. p. 1183–1193. Available from Internet: <http://www.aclweb.org/anthology/D10-1115>. Accessed: Jan 16, 2018.

BICK, E. The parsing system palavras. **Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**, University of Arhus, 2000.

BOOS, R.; PRESTES, K.; VILLAVICENCIO, A. Identification of multiword expressions in the brWaC. In: **Proceedings of LREC 2014**. ELRA, 2014. p. 728–735. ISBN 978-2-9517408-8-4. ACL Anthology Identifier: L14-1429. Available from Internet: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/518_Paper.pdf>. Accessed: Jan 16, 2018.

BOTT, S. et al. Ghost-pv: A representative gold standard of german particle verbs. **COLING 2016**, p. 125, 2016.

BOUMA, G. Normalized (pointwise) mutual information in collocation extraction. **Proceedings of GSCL**, p. 31–40, 2009.

BRIDE, A.; CRUYS, T. Van de; ASHER, N. A generalisation of lexical functions for composition in distributional semantics. In: **ACL (1)**. [S.l.: s.n.], 2015. p. 281–291.

BU, F.; ZHU, X.; LI, M. Measuring the non-compositionality of multiword expressions. In: **Proceedings of the 23rd International Conference on Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (COLING '10), p. 116–124. Available from Internet: <http://dl.acm.org/citation.cfm?id=1873781.1873795>. Accessed: Jan 16, 2018.

BULLINARIA, J. A.; LEVY, J. P. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. **Behavior Research Methods**, v. 44, n. 3, p. 890–907, 2012. ISSN 1554-3528. Available from Internet: <http://dx.doi.org/10.3758/s13428-011-0183-8>. Accessed: Jan 16, 2018.

CAMACHO-COLLADOS, J.; PILEHVAR, M. T.; NAVIGLI, R. A framework for the construction of monolingual and cross-lingual word similarity datasets. In: **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**. Beijing, China: Association for Computational Linguistics, 2015. p. 1–7. Available from Internet: <http://www.aclweb.org/anthology/P15-2001>. Accessed: Jan 16, 2018.

CAP, F. Show me your variance and i tell you who you are–deriving compound compositionality from word alignments. **MWE 2017**, p. 102, 2017.

CAP, F. et al. How to account for idiomatic german support verb constructions in statistical machine translation. In: **Proceedings of the 11th Workshop on Multiword Expressions**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 19–28. Available from Internet: <http://www.aclweb.org/anthology/W15-0903>. Accessed: Jan 16, 2018.

CARPUAT, M.; DIAB, M. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In: **Proc. of NAACL/HLT 2010**. Los Angeles, CA: [s.n.], 2010. p. 242–245.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational linguistics**, MIT Press, v. 16, n. 1, p. 22–29, 1990.

CIARAMITA, M.; JOHNSON, M. Supersense tagging of unknown nouns in wordnet. In: **Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (EMNLP '03), p. 168–175. Available from Internet: <https://doi.org/10.3115/1119355.1119377>. Accessed: Jan 16, 2018.

CONSTANT, M. et al. Multiword expression processing: A survey. **Computational Linguistics**, MIT Press, 2017.

CONSTANT, M.; NIVRE, J. A transition-based system for joint lexical and syntactic analysis. In: **ACL**. [S.l.: s.n.], 2016.

CONSTANT, M.; SIGOGNE, A. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In: **Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World**. Portland, Oregon, USA: Association for Computational Linguistics, 2011. p. 49–56. Available from Internet: <http://www.aclweb.org/anthology/W11-0809>. Accessed: Jan 16, 2018.

CORDEIRO, S. R. et al. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.: s.n.], 2016. v. 1, p. 1986–1997.

CORDEIRO, S. R.; RAMISCH, C.; VILLAVICENCIO, A. Token-based MWE identification strategies in the mwetoolkit. In: **PARSEME's 4th general meeting**. Malta: [s.n.], 2015.

CORDEIRO, S. R.; RAMISCH, C.; VILLAVICENCIO, A. Filtering and measuring the intrinsic quality of human compositionality judgments. In: **Proceedings of the 12th Workshop on Multiword Expressions**. [S.l.: s.n.], 2016. p. 32–37.

CORDEIRO, S. R.; RAMISCH, C.; VILLAVICENCIO, A. mwetoolkit+sem: Integrating word embeddings in the mwetoolkit for semantic MWE processing. In: **Proc. of LREC 2016**. Portoroz, Slovenia: [s.n.], 2016.

CORDEIRO, S. R.; RAMISCH, C.; VILLAVICENCIO, A. Ufrgs&lif at semeval-2016 task 10: Rule-based MWE identification and predominant-supersense tagging. In: **SemEval at NAACL-HLT**. [S.l.: s.n.], 2016. p. 910–917.

DUNNING, T. Accurate methods for the statistics of surprise and coincidence. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 19, n. 1, p. 61–74, mar. 1993. ISSN 0891-2017. Available from Internet: <http://dl.acm.org/citation.cfm?id=972450.972454>. Accessed: Jan 16, 2018.

ERK, K.; PADÓ, S. Exemplar-based models for word meaning in context. In: **ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, Short Papers**. The Association for Computer Linguistics, 2010. p. 92–97. Available from Internet: <http://www.aclweb.org/anthology/P10-2017>. Accessed: Jan 16, 2018.

EVERT, S.; KRENN, B. Using small random samples for the manual evaluation of statistical association measures. **Computer Speech & Language**, Elsevier, v. 19, n. 4, p. 450–466, 2005.

FARAHMAND, M.; SMITH, A.; NIVRE, J. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In: **Proceedings of the 11th Workshop on Multiword Expressions**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 29–33. Available from Internet: <http://www.aclweb.org/anthology/W15-0904>. Accessed: Jan 16, 2018.

FAZLY, A.; STEVENSON, S. Automatically constructing a lexicon of verb phrase idiomatic combinations. In: **EACL**. [S.l.: s.n.], 2006.

FELLBAUM, C. (Ed.). **WordNet: An Electronic Lexical Database (Language, Speech, and Communication)**. [S.l.]: MITPRESS, 1998. 423 p. ISBN 0-262-06197-X.

FERRET, O. Compounds and distributional thesauri. In: CALZOLARI, N. et al. (Ed.). **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 2979–2984. ISBN 978-2-9517408-8-4. ACL Anthology Identifier: L14-1590. Available from Internet: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/754_Paper.pdf>. Accessed: Jan 16, 2018.

FIRTH, J. R. A synopsis of linguistic theory, 1930-1955. Blackwell, 1957.

FRANTZI, K.; ANANIADOU, S.; MIMA, H. Automatic recognition of multi-word terms: the c-value/nc-value method. **International Journal on Digital Libraries**, Springer, v. 3, n. 2, p. 115–130, 2000.

FREGE, G. Über sinn und bedeutung. **Zeitschrift für Philosophie und philosophische Kritik**, v. 100, p. 25–50, 1892/1960. Translated, as 'On Sense and Reference', by Max Black.

FREITAG, D. et al. New experiments in distributional representations of synonymy. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Ninth Conference on Computational Natural Language Learning**. [S.l.], 2005. p. 25–32.

GANITKEVITCH, J.; DURME, B. V.; CALLISON-BURCH, C. Ppdb: The paraphrase database. In: **HLT-NAACL**. [S.l.: s.n.], 2013. p. 758–764.

GIRJU, R. et al. On the semantics of noun compounds. **Computer speech & language**, Elsevier, v. 19, n. 4, p. 479–496, 2005.

GOLDBERG, A. E. Compositionality. In: ______. **The Routledge Handbook of Semantics**. [S.l.]: Routledge, 2015. chp. 24.

GUEVARA, E. Computing semantic compositionality in distributional semantics. In: **Proceedings of the Ninth International Conference on Computational Semantics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (IWCS '11), p. 135–144. Available from Internet: <http://dl.acm.org/citation.cfm?id=2002669.2002684>. Accessed: Jan 16, 2018.

GURRUTXAGA, A.; ALEGRIA, I. n. Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In: **Proceedings**

of the 9th Workshop on Multiword Expressions. Atlanta, Georgia, USA: Association for Computational Linguistics, 2013. p. 116–125. Available from Internet: <http://www.aclweb.org/anthology/W13-1017>. Accessed: Jan 16, 2018.

HARRIS, Z. Distributional structure. **Word**, v. 10, p. 146–162, 1954.

HARTUNG, M. et al. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In: **Proceedings of the 15th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)**. [S.l.: s.n.], 2017.

HENDRICKX, I. et al. Semeval-2013 task 4: Free paraphrases of noun compounds. In: **Proceedings of *SEM 2013, Volume 2 – SemEval**. ACL, 2013. p. 138–143. Available from Internet: <http://www.aclweb.org/anthology/S13-2025>. Accessed: Jan 16, 2018.

HUANG, Z.; XU, W.; YU, K. Bidirectional lstm-crf models for sequence tagging. **arXiv preprint arXiv:1508.01991**, 2015.

HWANG, J. D. et al. Propbank annotation of multilingual light verb constructions. In: ACL. **Proc. of the LAW 2010**. [S.l.], 2010. p. 82–90.

JACKENDOFF, R. **The Architecture of the Language Faculty**. [S.l.]: MIT Press, 1997. xvi+262+ p. ISBN 0-262-60025-0.

JAGFELD, G.; PLAS, L. van der. Towards a better semantic role labelling of complex predicates. In: **Proc. of NAACL Student Research Workshop**. Denver, US: [s.n.], 2015. p. 33–39.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall, 2008. 1024 p. ISBN 0-13-187321-0.

JUSTESON, J. S.; KATZ, S. M. Technical terminology: some linguistic properties and an algorithm for identification in text. **Natural language engineering**, Cambridge Univ Press, v. 1, n. 01, p. 9–27, 1995.

KIELA, D.; CLARK, S. A systematic study of semantic vector space model parameters. In: **Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL**. [S.l.: s.n.], 2014. p. 21–30.

KIM, S. N.; BALDWIN, T. Word sense and semantic relations in noun compounds. **ACM Trans. Speech Lang. Process.**, ACM, New York, NY, USA, v. 10, n. 3, p. 9:1–9:17, jul. 2013. ISSN 1550-4875. Available from Internet: <http://doi.acm.org/10.1145/2483969.2483971>. Accessed: Jan 16, 2018.

KÖPER, M.; WALDE, S. S. im. Distinguishing literal and non-literal usage of german particle verbs. In: **HLT-NAACL**. [S.l.: s.n.], 2016. p. 353–362.

KULKARNI, N.; FINLAYSON, M. jMWE: A Java toolkit for detecting multi-word expressions. In: **Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (MWE '11), p. 122–124.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: **Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. p. 282–289. ISBN 1-55860-778-1. Available from Internet: <http://dl.acm.org/citation.cfm?id=645530.655813>. Accessed: Jan 16, 2018.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. **Discourse processes**, Taylor & Francis, v. 25, n. 2-3, p. 259–284, 1998.

LANDES, S.; LEACOCK, C.; TENGI, R. I. Building semantic concordances. **WordNet: an electronic lexical database**, MIT Press, Cambridge, MA, v. 199, n. 216, p. 199–216, 1998.

LAPESA, G.; EVERT, S. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. **Transactions of the Association for Computational Linguistics**, v. 2, p. 531–545, 2014.

LAPESA, G.; EVERT, S. Large-scale evaluation of dependency-based dsms: Are they worth the effort? **EACL 2017**, p. 394, 2017.

LAUER, M. How much is enough?: Data requirements for statistical NLP. **CoRR**, abs/cmp-lg/9509001, 1995. Available from Internet: <http://arxiv.org/abs/cmp-lg/9509001>. Accessed: Jan 16, 2018.

LEVY, O.; GOLDBERG, Y.; DAGAN, I. Improving distributional similarity with lessons learned from word embeddings. **Transactions of the Association for Computational Linguistics**, v. 3, p. 211–225, 2015. ISSN 2307-387X. Available from Internet: <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570>. Accessed: Jan 16, 2018.

LIN, D. Automatic retrieval and clustering of similar words. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 17th international conference on Computational linguistics-Volume 2**. [S.l.], 1998. p. 768–774.

LIN, D. Automatic identification of non-compositional phrases. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics**. [S.l.], 1999. p. 317–324.

MAAROUF, I. E.; OAKES, M. Statistical measures for characterising mwes. In: **IC1207 COST PARSEME 5th general meeting**. [S.l.: s.n.], 2015.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge, USA: [s.n.], 1999. 620 p. ISBN 0-262-13360-1.

MCCARTHY, D.; KELLER, B.; CARROLL, J. Detecting a continuum of compositionality in phrasal verbs. In: **Proceedings of the ACL 2003 Workshop on Multiword Expressions**. ACL, 2003. (MWE '03), p. 73–80. Available from Internet: <http://dx.doi.org/10.3115/1119282.1119292>. Accessed: Jan 16, 2018.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2013. p. 3111–3119.

MITCHELL, J.; LAPATA, M. Vector-based models of semantic composition. In: **ACL**. [S.l.: s.n.], 2008. p. 236–244.

MITCHELL, J.; LAPATA, M. Composition in distributional models of semantics. **Cognitive science**, Wiley Online Library, v. 34, n. 8, p. 1388–1429, 2010.

MITKOV, R. **The Oxford handbook of computational linguistics**. [S.l.]: Oxford University Press, 2005.

MOLDOVAN, D. et al. Models for the semantic classification of noun phrases. In: **Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. (CLS '04), p. 60–67. Available from Internet: <http://dl.acm.org/citation.cfm?id=1596431.1596440>. Accessed: Jan 16, 2018.

NAKOV, P. Paraphrasing verbs for noun compound interpretation. In: **Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)**. [S.l.: s.n.], 2008. p. 46–49.

NAKOV, P. On the interpretation of noun compounds: Syntax, semantics, and entailment. **Natural Language Engineering**, v. 19, n. 3, p. 291–330, 2013. Available from Internet: <http://dx.doi.org/10.1017/S1351324913000065>. Accessed: Jan 16, 2018.

NASR, A. et al. Joint dependency parsing and multiword expression tokenisation. In: **Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2015. p. 1116–1126.

NAVIGLI, R.; PONZETTO, S. P. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. **Artificial Intelligence**, Elsevier Science Publishers Ltd., v. 193, p. 217–250, 2012.

NIVRE, J.; HALL, J.; NILSSON, J. Maltparser: A data-driven parser-generator for dependency parsing. In: **Proceedings of LREC**. [S.l.: s.n.], 2006. v. 6, p. 2216–2219.

OKAZAKI, N. **CRFsuite: a fast implementation of Conditional Random Fields (CRFs)**. 2007. Available from Internet: <http://www.chokkan.org/software/crfsuite/>. Accessed: Jan 16, 2018.

PADÓ, S.; LAPATA, M. Dependency-based construction of semantic space models. **Computational Linguistics**, MIT Press, v. 33, n. 2, p. 161–199, 2007.

PADRÓ, M. et al. Nothing like good old frequency: Studying context filters for distributional thesauri. In: **Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014) - short papers**. Doha, Qatar: [s.n.], 2014. Available from Internet: <http://www.aclweb.org/anthology/D14-1047>. Accessed: Jan 16, 2018.

PARTEE, B. H. **Montague grammar**. [S.l.]: Elsevier, 2014.

PECINA, P. Lexical association measures and collocation extraction. **Language resources and evaluation**, Springer, v. 44, n. 1-2, p. 137–158, 2010.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Available from Internet: <http://www.aclweb.org/anthology/D14-1162>. Accessed: Jan 16, 2018.

PETROV, S.; DAS, D.; MCDONALD, R. A universal part-of-speech tagset. **arXiv preprint arXiv:1104.2086**, 2011.

PINKER, S. Language acquisition. **Language: An invitation to cognitive science**, v. 1, p. 135–82, 1995.

RAMISCH, C. **Multiword Expressions Acquisition - A Generic and Open Framework**. Springer, 2015. (Theory and Applications of Natural Language Processing). Available from Internet: <http://dx.doi.org/10.1007/978-3-319-09207-2>. Accessed: Jan 16, 2018.

RAMISCH, C. et al. How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. [S.l.: s.n.], 2016. p. 156.

RAMSHAW, L.; MARCUS, M. Text chunking using transformation-based learning. In: **Third Workshop on Very Large Corpora**. [s.n.], 1995. Available from Internet: <http://aclweb.org/anthology/W95-0107>. Accessed: Jan 16, 2018.

REDDY, S.; MCCARTHY, D.; MANANDHAR, S. An empirical study on compositionality in compound nouns. In: **Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011)**. Chiang Mai, Thailand: [s.n.], 2011. Available from Internet: <http://sivareddy.in/papers/ijcnlp2011empirical.pdf>. Accessed: Jan 16, 2018.

REN, Z. et al. Improving statistical machine translation using domain bilingual multiword expressions. In: **Proc. of the ACL 2009 Workshop on MWEs**. Singapore: [s.n.], 2009. p. 47–54.

REY, D.; NEUHÄUSER, M. Wilcoxon-signed-rank test. In: **International encyclopedia of statistical science**. [S.l.]: Springer, 2011. p. 1658–1659.

RIEDL, M.; BIEMANN, C. A single word is not enough: Ranking multiword expressions using distributional semantics. In: **EMNLP**. [S.l.: s.n.], 2015. p. 2430–2440.

ROLLER, S.; WALDE, S. Schulte im. Feature norms of german noun compounds. In: **Proceedings of the 10th Workshop on Multiword Expressions (MWE)**. ACL, 2014. p. 104–108. Available from Internet: <http://www.aclweb.org/anthology/W14-0818>. Accessed: Jan 16, 2018.

ROLLER, S.; WALDE, S. Schulte im; SCHEIBLE, S. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In:

Proceedings of the 9th Workshop on Multiword Expressions. ACL, 2013. p. 32–41. Available from Internet: <http://www.aclweb.org/anthology/W13-1005>. Accessed: Jan 16, 2018.

SAG, I. A. et al. Multiword expressions: A pain in the neck for nlp. In: **Computational Linguistics and Intelligent Text Processing**. [S.l.]: Springer, 2002. p. 1–15.

SALEHI, B.; COOK, P.; BALDWIN, T. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In: BOUMA, G.; PARMENTIER, Y. (Ed.). **Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics**. Gothenburg, Sweden: The Association for Computer Linguistics, 2014. p. 472–481. Available from Internet: <http://aclweb.org/anthology/E/E14/E14-1050.pdf>. Accessed: Jan 16, 2018.

SALEHI, B.; COOK, P.; BALDWIN, T. A word embedding approach to predicting the compositionality of multiword expressions. In: **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 977–983. Available from Internet: <http://www.aclweb.org/anthology/N15-1099>. Accessed: Jan 16, 2018.

SALEHI, B. et al. The impact of multiword expression compositionality on machine translation evaluation. In: **Proceedings of the 11th Workshop on Multiword Expressions**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 54–59. Available from Internet: <http://www.aclweb.org/anthology/W15-0909>. Accessed: Jan 16, 2018.

SALLE, A.; VILLAVICENCIO, A.; IDIART, M. Matrix factorization using window sampling and negative sampling for improved word representations. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 419–424. Available from Internet: <http://anthology.aclweb.org/P16-2068>. Accessed: Jan 16, 2018.

SAVARY, A. Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. In: MANETH, S. (Ed.). **CIAA**. Springer, 2009. (Lecture Notes in Computer Science, v. 5642), p. 237–240. ISBN 978-3-642-02978-3. Available from Internet: <http://dblp.uni-trier.de/db/conf/wia/ciaa2009.html#Savary09>. Accessed: Jan 16, 2018.

SAVARY, A. et al. Parseme multilingual corpus of verbal multiword expressions. In: **Phraseology and Multiword Expressions**. [S.l.]: Language Science Press (LangSci), 2017. Submitted for review.

SAVARY, A. et al. The parseme shared task on automatic identification of verbal multiword expressions. In: **Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)**. [S.l.: s.n.], 2017. p. 31–47.

SAVARY, A. et al. The parseme shared task on automatic identification of verbal multiword expressions. In: **Phraseology and Multiword Expressions**. [S.l.]: Language Science Press (LangSci), 2017. Submitted for review.

SAVARY, A. et al. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In: **7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)**. Poznań, Poland: [s.n.], 2015.

SCHMID, H. Treetagger — a language independent part-of-speech tagger. **Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart**, v. 43, p. 28, 1995.

SCHNEIDER, N. **Lexical Semantic Analysis in Natural Language Text**. Thesis (PhD) — Ph. D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2014.

SCHNEIDER, N. et al. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. **Transactions of the Association for Computational Linguistics**, v. 2, p. 193–206, abr. 2014. Available from Internet: <http://www.transacl.org/wp-content/uploads/2014/04/51.pdf>. Accessed: Jan 16, 2018.

SCHNEIDER, N. et al. SemEval 2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In: **Proc. of SemEval**. San Diego, California, USA: [s.n.], 2016.

SCHNEIDER, N. et al. A corpus of preposition supersenses. In: **Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL**. [S.l.: s.n.], 2016. p. 99–109.

SCHNEIDER, N. et al. Comprehensive annotation of multiword expressions in a social web corpus. 2014.

SCHNEIDER, N. et al. Comprehensive annotation of multiword expressions in a social web corpus. In: CALZOLARI, N. et al. (Ed.). **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 455–461. ISBN 978-2-9517408-8-4. ACL Anthology Identifier: L14-1433. Available from Internet: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf>. Accessed: Jan 16, 2018.

SCHOLIVET, M.; RAMISCH, C.; CORDEIRO, S. R. Sequence models and lexical resources for mwe identification in french. In: **Phraseology and Multiword Expressions**. [S.l.]: Language Science Press (LangSci), 2017. Submitted for review.

SCHONE, P.; JURAFSKY, D. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In: **Proceedings of Empirical Methods in Natural Language Processing**. Pittsburgh, PA: [s.n.], 2001. Available from Internet: <http://www.colorado.edu/ling/jurafsky/emnlp2001_mwu_iii.pdf>. Accessed: Jan 16, 2018.

SEDDAH, D. et al. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In: **Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages**. Seattle, WA, USA: Association for Computational Linguistics, 2013. p. 146–182. Available from Internet: <http://www.aclweb.org/anthology/W13-4917>. Accessed: Jan 16, 2018.

SERETAN, V. **Syntax-Based Collocation Extraction**. 1st. ed. Dordrecht, Netherlands: Springer, 2011. (Text, Speech and Language Technology, v. 44). 212 p. ISBN 978-94-007-0133-5.

SHLENS, J. A tutorial on principal component analysis. **arXiv preprint arXiv:1404.1100**, 2014.

SMADJA, F. Retrieving collocations from text: Xtract. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 19, n. 1, p. 143–177, mar. 1993. ISSN 0891-2017. Available from Internet: <http://dl.acm.org/citation.cfm?id=972450.972458>. Accessed: Jan 16, 2018.

SOCHER, R. et al. Semantic compositionality through recursive matrix-vector spaces. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**. [S.l.], 2012. p. 1201–1211.

SPORLEDER, C.; LI, L. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In: **Proc. of EACL 2009**. Athens: [s.n.], 2009. p. 754–762.

STEVENSON, S.; FAZLY, A.; NORTH, R. Statistical measures of the semi-productivity of light verb constructions. In: **Proceedings of the Workshop on Multiword Expressions: Integrating Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. (MWE '04), p. 1–8. Available from Internet: <http://dl.acm.org/citation.cfm?id=1613186.1613187>. Accessed: Jan 16, 2018.

STYMNE, S.; CANCEDDA, N.; AHRENBERG, L. Generation of compound words in statistical machine translation into compounding languages. **Computational Linguistics**, MIT Press, v. 39, n. 4, p. 1067–1108, 2013.

TSVETKOV, Y.; WINTNER, S. Identification of multiword expressions by combining multiple linguistic information sources. **Computational Linguistics**, MIT Press, v. 40, n. 2, p. 449–468, 2014.

WALDE, S. Schulte im et al. Ghost-nn: A representative gold standard of german noun-noun compounds. In: **LREC**. [S.l.: s.n.], 2016.

WALDE, S. Schulte im; MüLLER, S.; ROLLER, S. Exploring vector space models to predict the compositionality of german noun-noun compounds. In: **Proceedings of *SEM 2013, Volume 1**. ACL, 2013. p. 255–265. Available from Internet: <http://www.aclweb.org/anthology/S13-1038>. Accessed: Jan 16, 2018.

WASZCZUK, J.; SAVARY, A.; PARMENTIER, Y. Promoting multiword expressions in a* tag parsing. In: **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 429–439. Available from Internet: <http://aclweb.org/anthology/C16-1042>. Accessed: Jan 16, 2018.

WILKENS, R. et al. Lexsubnc: A dataset of lexical substitution for nominal compounds. In: **Proceedings of the Twelfth International Conference on Computational Semantics**. [S.l.]: Association for Computational Linguistics, 2017. (IWCS'17).

YAZDANI, M.; FARAHMAND, M.; HENDERSON, J. Learning semantic composition to detect non-compositionality of multiword expressions. In: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 1733–1742. Available from Internet: <http://aclweb.org/anthology/D15-1201>. Accessed: Jan 16, 2018.

ZHOU, Z.-H. **Ensemble methods: foundations and algorithms**. [S.l.]: CRC press, 2012.

ZILIO, L. et al. Joining forces for multiword expression identification. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2016. p. 233–238.

INDEX

APPENDIX

A. List of English Compounds

We present below the 90 nominal compounds in *EN-comp*$_{90}$, along with their human-rated compositionality scores. We refer to Reddy, McCarthy and Manandhar (2011) for the other 90 compounds.

Compounds	$c_{\mathbf{WC}}$
ancient history	1.95
armchair critic	1.33
baby buggy	3.94
bad hat	0.62
benign tumour	4.69
big fish	0.85
birth rate	4.60
black cherry	3.11
bow tie	4.25
brain teaser	2.65
busy bee	0.88
carpet bombing	1.24
cellular phone	3.78
close call	1.59
closed book	0.68
computer program	4.50
con artist	2.10
cooking stove	4.68
cotton candy	1.79
critical review	4.06
dead end	1.32
dirty money	2.21
dirty word	2.48
disc jockey	1.25
divine service	3.11
dry land	3.95
dry wall	3.33
dust storm	3.85
eager beaver	0.36
economic aid	4.33
elbow grease	0.56
elbow room	0.61
entrance hall	4.17
eternal rest	3.25
fish story	1.68
flower child	0.50
food market	3.82
foot soldier	1.95
front man	1.64
goose egg	0.48
grey matter	2.39
guinea pig	0.45
half sister	2.84
half wit	1.16
health check	4.17
high life	1.67
inner circle	1.56
inner product	3.00
insane asylum	3.95
insurance company	5.00
insurance policy	4.15
iron collar	3.88
labour union	4.76
life belt	2.84
life vest	3.44
lime tree	4.61
loan shark	1.00
loose woman	2.53
mail service	4.69
market place	3.00
mental disorder	4.89
middle school	3.84
milk tooth	1.43
mother tongue	0.59
narrow escape	1.75
net income	2.94
news agency	4.39
noble gas	1.18
nut case	0.44
old flame	0.58
old hat	0.35
old timer	0.89
phone book	4.25
pillow slip	3.70
pocket book	1.42
prison guard	4.89
prison term	4.79
private eye	0.82
record book	3.70
research lab	4.75
sex bomb	0.53
silver lining	0.35
sound judgement	3.39
sparkling water	3.14
street girl	3.16
subway system	4.63
tennis elbow	2.50
top dog	1.05
wet blanket	0.21
word painting	1.62

B. List of French Compounds

We present below the 180 nominal compounds in *FR-comp*, along with their human-rated compositionality scores.

Compounds	$c_{\mathbf{WC}}$	Gloss
activité physique	4.93	'physical activity' (lit. *activity physical*)
année scolaire	3.60	'school year' (lit. *year scholar*)
art contemporain	4.60	'contemporary art' (lit. *art contemporary*)
baie vitrée	3.64	'open glass window' (lit. *opening glassy*)
bas côté	1.31	'aisle' (lit. *low side*)
beau frère	0.67	'brother-in-law' (lit. *beautiful brother*)
beau père	1.18	'father-in-law' (lit. *beautiful father*)
belle mère	0.80	'mother-in-law' (lit. *beautiful mother*)
berger allemand	1.29	'German shepherd' (lit. *shepherd German*)
bon sens	3.57	'common sense' (lit. *good sense*)
bon vent	0.87	'good luck' (lit. *good/fair wind*)
bon vivant	2.57	'bon vivant' (lit. *good "liver"*)
bonne humeur	4.53	'good mood' (lit. *good humor*)
bonne poire	0.42	'sucker, soft touch' (lit. *good pear*)
bonne pratique	4.47	'good practice' (lit. *good practice*)
bouc émissaire	0.23	'scapegoat' (lit. *goat emissary*)
bras cassé	0.57	'lame duck' (lit. *arm broken*)
bras droit	0.40	'right arm' (lit. *arm right*)
brebis galeuse	0.55	'black sheep' (lit. *sheep scabby*)
carte blanche	0.20	'carte blanche' (lit. *card white*)
carte bleue	1.94	'bank card' (lit. *card blue*)
carte grise	3.08	'vehicle registration' (lit. *card grey*)
carte vitale	1.70	'healthcare card' (lit. *card vital*)
carton plein	0.78	'clean sweep' (lit. *cardboard full*)
casque bleu	1.85	'UN peacekeeper' (lit. *helmet blue*)
centre commercial	3.93	'shopping center' (lit. *center commercial*)
cercle vicieux	2.15	'vicious circle' (lit. *circle vicious*)
cerf volant	0.64	'kite' (lit. *deer flying*)
chambre froide	4.27	'cold chamber' (lit. *chamber cold*)
changement climatique	4.79	'climate change' (lit. *change climatic*)
chapeau bas	0.64	'bravo' (lit. *hat low*)
charge sociale	3.00	'social security contribution' (lit. *charge social*)
chauve souris	0.33	'bat' (lit. *bald mouse*)
chute libre	3.64	'free fall' (lit. *fall free*)
club privé	4.58	'private club' (lit. *club private*)
coffre fort	3.67	'safe, vault' (lit. *chest/box strong*)
communauté urbaine	4.57	'urban community' (lit. *community urban*)
conseil municipal	4.00	'city council' (lit. *council municipal*)
coup dur	2.40	'hard blow' (lit. *blow hard*)
coup franc	1.71	'free kick (soccer)' (lit. *blow free/frank*)
courrier électronique	4.57	'e-mail' (lit. *mail electronic*)
court circuit	1.69	'short circuit' (lit. *short circuit*)
court métrage	2.36	'short film' (lit. *short length/footage*)
crème fraîche	3.73	'French sour cream' (lit. *cream fresh*)
crème glacée	4.75	'ice cream' (lit. *cream icy*)
dernier cri	0.67	'latest, trendy' (lit. *last cry*)
dernier mot	3.09	'final say' (lit. *last word*)
directeur général	3.87	'chief executive officer' (lit. *director general*)
disque dur	2.83	'hard drive' (lit. *disk hard*)
douche froide	1.18	'damper' (lit. *cold shower*)
droit fondamental	4.27	'fundamental right' (lit. *right fundamental*)
développement économique	4.46	'economic development' (lit. *development economic*)
eau chaude	5.00	'hot water' (lit. *water hot*)
eau douce	2.33	'fresh water' (lit. *water sweet*)
eau minérale	4.00	'mineral water' (lit. *water mineral*)

Compounds	$c_{\mathbf{WC}}$	Gloss
eau potable	5.00	‘drinking water’ (lit. *water potable*)
eau vive	3.44	‘jellyfish’ (lit. *water living*)
eau forte	0.90	‘etching’ (lit. *water strong*)
eaux usées	4.54	‘sewage’ (lit. *waters used*)
effet spécial	3.67	‘special effect’ (lit. *effect special*)
expérience professionnelle	4.86	‘professional experience’ (lit. *experience professional*)
fait divers	3.69	‘news story’ (lit. *fact diverse*)
famille nombreuse	4.90	‘large family’ (lit. *family numerous*)
faux ami	1.25	‘false friend’ (lit. *false friend*)
faux cul	0.31	‘hypocrite’ (lit. *false arse*)
faux pas	1.82	‘blunder’ (lit. *false step*)
faux semblant	3.57	‘false pretence’ (lit. *false appearance*)
feu rouge	2.60	‘red traffic light’ (lit. *fire red*)
feu vert	0.71	‘green light, permission’ (lit. *fire green*)
fil conducteur	1.25	‘underlying theme’ (lit. *thread conducting*)
fleur bleue	0.45	‘sentimental’ (lit. *flower blue*)
foie gras	4.54	‘foie gras’ (lit. *liver fatty*)
fou rire	2.33	‘giggle’ (lit. *crazy laughter*)
grand air	1.33	‘outdoors’ (lit. *big air*)
grand jour	1.07	‘broad daylight’ (lit. *big day*)
grand saut	2.17	‘move forward’ (lit. *big leap*)
grand écran	3.14	‘silver screen’ (lit. *big screen*)
grande entreprise	4.54	‘big company’ (lit. *big company*)
grande surface	3.14	‘department store’ (lit. *big surface*)
grippe aviaire	3.58	‘avian flu’ (lit. *flu avian*)
gros mot	1.40	‘swearword’ (lit. *large word*)
gros plan	1.87	‘close-up’ (lit. *large plan*)
guerre civile	3.43	‘civil war’ (lit. *war civil*)
haut parleur	1.83	‘loudspeaker’ (lit. *loud/high speaker*)
haute mer	2.54	‘high seas’ (lit. *high sea*)
haute montagne	4.13	‘high mountains’ (lit. *high mountain*)
heure supplémentaire	4.00	‘overtime hour’ (lit. *hour extra*)
huile essentielle	2.25	‘essential oil’ (lit. *oil essential*)
idée reçue	2.90	‘popular belief’ (lit. *idea received*)
insertion professionnelle	4.27	‘professional insertion’ (lit. *insertion professional*)
intérêt général	4.36	‘general interest’ (lit. *interest general*)
jeune fille	4.64	‘young girl, maiden’ (lit. *young girl*)
journal officiel	4.50	‘official gazette’ (lit. *newspaper official*)
langue française	4.85	‘French language’ (lit. *language French*)
marée noire	3.00	‘oil spill’ (lit. *tide black*)
match nul	2.46	‘draw, stalemate’ (lit. *match null*)
matière grasse	5.00	‘fat’ (lit. *matter greasy*)
matière grise	2.15	‘grey matter’ (lit. *material grey*)
matière première	2.90	‘raw material’ (lit. *material primary*)
mauvaise foi	2.38	‘bad faith’ (lit. *bad faith*)
mauvaise langue	2.21	‘gossip’ (lit. *bad tongue*)
montagnes russes	1.08	‘roller coaster’ (lit. *mountains Russian*)
monument historique	4.79	‘historical monument’ (lit. *monument historical*)
mort né	3.23	‘stillborn’ (lit. *dead born*)
nouveau monde	2.73	‘New World, Americas’ (lit. *new world*)
nuit blanche	1.07	‘sleepless night’ (lit. *night white*)
numéro vert	1.50	‘toll-free number’ (lit. *number green*)
ordure ménagère	4.20	‘household waste’ (lit. *garbage household*)
organisation syndicale	4.90	‘trade union’ (lit. *organisation of-trade-union*)
pages jaunes	3.00	‘yellow pages’ (lit. *pages yellow*)
parachute doré	0.50	‘golden parachute’ (lit. *parachute golden*)
parc naturel	4.33	‘nature park’ (lit. *park natural*)
parti politique	4.88	‘political party’ (lit. *party political*)
parti pris	2.69	‘bias’ (lit. *party taken*)
partie fine	0.80	‘orgy’ (lit. *party fine/delicate*)
petit ami	0.86	‘boyfriend’ (lit. *small friend*)
petit beurre	1.64	‘butter biscuit’ (lit. *small butter*)

Compounds	$c_{\mathbf{WC}}$	Gloss
petit déjeuner	2.27	‘breakfast’ (lit. *small lunch*)
petit joueur	1.00	‘amateur’ (lit. *small player*)
petit pois	4.14	‘pea’ (lit. *small pea*)
petit salé	1.15	‘salted pork’ (lit. *small salty*)
petit écran	2.50	‘television’ (lit. *small screen*)
petit enfant	2.79	‘grandchild’ (lit. *small child*)
petit four	0.92	‘type of dessert’ (lit. *small oven*)
petit nègre	0.50	‘pidgin French’ (lit. *little black-man*)
petite annonce	2.69	‘classified ad’ (lit. *small announcement*)
petite nature	0.47	‘squeamish’ (lit. *small nature*)
pied noir	0.13	‘French expats from Algeria’ (lit. *foot black*)
pièce montée	2.47	‘tiered cake’ (lit. *piece assembled*)
pleine lune	3.54	‘full moon’ (lit. *full moon*)
poids lourd	2.08	‘truck’ (lit. *weight heavy*)
point faible	2.46	‘weak point’ (lit. *point weak*)
point mort	1.00	‘standstill’ (lit. *point dead*)
pot pourri	0.40	‘medley’ (lit. *pot/jar rotten*)
poule mouillée	0.00	‘coward’ (lit. *chicken wet*)
poupée russe	3.75	‘Russian nesting doll’ (lit. *doll Russian*)
premier ministre	3.67	‘first minister’ (lit. *first minister*)
premier plan	2.82	‘foreground’ (lit. *first plan*)
première dame	1.92	‘first lady’ (lit. *first lady*)
prince charmant	2.00	‘Prince Charming’ (lit. *prince charming*)
prévision météorologique	4.70	‘weather forecast’ (lit. *forecast meteorological*)
recherche scientifique	4.92	‘scientific research’ (lit. *research scientific*)
ressources humaines	3.91	‘human resources’ (lit. *resources human*)
rond point	3.18	‘roundabout’ (lit. *round point*)
roulette russe	0.87	‘Russian roulette’ (lit. *roulette Russian*)
réchauffement climatique	4.40	‘global warming’ (lit. *warming climatic*)
région parisienne	4.43	‘Paris region’ (lit. *region Parisian*)
réseau social	4.09	‘social network’ (lit. *network social*)
sang froid	0.47	‘self-control’ (lit. *blood cold*)
second degré	1.40	‘tongue-in-cheek’ (lit. *second degree*)
second rôle	3.64	‘supporting role’ (lit. *second role*)
septième ciel	0.21	‘cloud nine’ (lit. *seventh heaven*)
service public	4.71	‘public service’ (lit. *service public*)
site officiel	4.85	‘official website’ (lit. *website official*)
soirée privée	4.53	‘private party’ (lit. *party private*)
sucre roux	4.31	‘brown sugar’ (lit. *sugar ginger-colored*)
sécurité routière	4.55	‘road safety’ (lit. *safety of-road*)
sécurité sociale	3.67	‘social security’ (lit. *security social*)
table basse	4.79	‘coffee table’ (lit. *table low*)
table ronde	1.46	‘round table’ (lit. *table round*)
tapis rouge	3.31	‘red carpet’ (lit. *carpet red*)
temps fort	1.87	‘key moment, highlight’ (lit. *time strong*)
temps mort	2.07	‘wasted time, idleness’ (lit. *time dead*)
temps partiel	3.62	‘part-time (work)’ (lit. *time partial*)
temps plein	3.08	‘full-time (work)’ (lit. *time full*)
temps réel	3.00	‘real time’ (lit. *time real*)
travaux publics	4.09	‘public works’ (lit. *works public*)
trou noir	2.58	‘black hole’ (lit. *hole black*)
trou normand	0.78	‘palate cleanser’ (lit. *hole Norman*)
téléphone arabe	0.23	‘Chinese whispers’ (lit. *telephone Arabic*)
téléphone portable	5.00	‘cellphone’ (lit. *telephone portable*)
valeur sûre	3.64	‘safe bet’ (lit. *value safe/sure*)
vie associative	4.00	‘community life’ (lit. *life associative*)
vie quotidienne	4.31	‘everyday life’ (lit. *life daily*)
vieille fille	2.42	‘spinster’ (lit. *old girl/maid*)
vin blanc	3.80	‘white wine’ (lit. *wine white*)
vin rouge	4.69	‘red wine’ (lit. *wine red*)
yeux rouges	4.36	‘red eyes’ (lit. *eyes red*)
école primaire	3.92	‘primary school’ (lit. *school primary*)
étoile filante	3.20	‘shooting star’ (lit. *star slipping*)

C. List of Portuguese Compounds

We present below the 180 nominal compounds in *PT-comp*, along with their human-rated compositionality scores.

Compounds	$c_{\mathbf{WC}}$	Gloss
abalo sísmico	4.42	‘earthquake’ (lit. *shock seismic*)
acampamento militar	4.82	‘military camp’ (lit. *camp military*)
agente secreto	4.58	‘secret agent’ (lit. *agent secret*)
alarme falso	3.24	‘false alarm’ (lit. *alarm false*)
algodão doce	1.28	‘cotton candy’ (lit. *cotton sweet*)
alta temporada	2.04	‘high season’ (lit. *high season*)
alta costura	1.52	‘haute couture’ (lit. *high sewing*)
alto mar	1.35	‘high seas’ (lit. *high sea*)
alto falante	0.88	‘loudspeaker’ (lit. *loud/high speaker*)
amigo oculto	2.89	‘secret Santa’ (lit. *friend hidden*)
amigo secreto	3.11	‘secret Santa’ (lit. *friend secret*)
amor próprio	3.91	‘self-esteem’ (lit. *love own*)
ano novo	4.29	‘new year’ (lit. *year new*)
ar condicionado	2.44	‘air conditioning’ (lit. *air conditioned*)
ar livre	1.95	‘open air’ (lit. *air free*)
arma branca	0.65	‘cold weapon’ (lit. *weapon white*)
ato falho	3.50	‘Freudian slip’ (lit. *act faulty*)
banho turco	2.19	‘Turkish bath’ (lit. *bath Turkish*)
batata doce	4.24	‘sweet potato’ (lit. *potato sweet*)
bebida alcoólica	5.00	‘alcoholic drink’ (lit. *drink alcoholic*)
bode expiatório	0.47	‘scapegoat’ (lit. *goat expiatory*)
braço direito	0.57	‘right arm’ (lit. *arm right*)
buraco negro	2.88	‘black hole’ (lit. *hole black/dark*)
café colonial	2.70	‘afternoon tea’ (lit. *breakfast colonial*)
caixa forte	3.19	‘safe, vault’ (lit. *box strong*)
caixa preta	0.94	‘black box’ (lit. *box black*)
caixeiro viajante	3.43	‘traveling salesman’ (lit. *clerk traveling*)
carne branca	2.85	‘white meat’ (lit. *meat white*)
carne vermelha	3.66	‘red meat’ (lit. *meat red*)
carro forte	2.62	‘armored car’ (lit. *car strong*)
carta aberta	3.64	‘open letter’ (lit. *letter open*)
centro comercial	3.68	‘shopping mall’ (lit. *center commercial*)
centro espírita	3.43	‘Spiritualist center’ (lit. *center spiritualist*)
cerca viva	3.58	‘hedge’ (lit. *fence living*)
cheiro verde	0.67	‘parsley’ (lit. *smell green*)
circuito integrado	4.52	‘integrated circuit’ (lit. *circuit integrated*)
classe executiva	2.67	‘business class’ (lit. *class executive*)
coluna social	2.45	‘gossip column’ (lit. *column social*)
colégio militar	4.88	‘military high-school’ (lit. *high-school military*)
comida caseira	4.11	‘homemade food’ (lit. *food homemade*)
companhia aérea	3.11	‘airline’ (lit. *company aerial*)
conta corrente	2.71	‘checking account’ (lit. *account current*)
coração partido	1.06	‘broken heart’ (lit. *heart broken*)
corda bamba	1.31	‘tightrope, bad situation’ (lit. *rope wobbly*)
cordas vocais	2.32	‘vocal chords’ (lit. *chords vocal*)
curto circuito	1.96	‘short circuit’ (lit. *short circuit*)
câmara fria	4.65	‘cold chamber’ (lit. *chamber cold*)
céu aberto	1.68	‘outdoors, open air’ (lit. *sky open*)
círculo vicioso	2.17	‘vicious circle’ (lit. *circle vicious*)
círculo virtuoso	2.39	‘virtuous circle’ (lit. *circle virtuous*)
deputado federal	4.92	‘federal deputy’ (lit. *deputy federal*)
desfile militar	4.93	‘military parade’ (lit. *parade military*)
direitos humanos	3.86	‘human rights’ (lit. *rights human*)
disco rígido	2.76	‘hard drive’ (lit. *disk rigid*)

Compounds	$c_{\mathbf{WC}}$	Gloss
disco voador	2.94	‘flying saucer’ (lit. *disk flying*)
efeitos especiais	3.37	‘special effects’ (lit. *effects special*)
elefante branco	0.16	‘white elephant’ (lit. *elephant white*)
escada rolante	3.85	‘escalator’ (lit. *stair rolling*)
estrela cadente	2.52	‘shooting star’ (lit. *star falling*)
exame clínico	4.75	‘clinical examination’ (lit. *examination clinical*)
exames laboratoriais	4.90	‘laboratory tests’ (lit. *examinations laboratory*)
farinha integral	4.72	‘wholemeal flour’ (lit. *flour integral*)
febre amarela	1.43	‘yellow fever’ (lit. *fever yellow*)
ficha limpa	2.97	‘clean criminal records’ (lit. *file clean*)
fila indiana	1.17	‘single file’ (lit. *queue Indian*)
fio condutor	1.58	‘underlying theme’ (lit. *thread conductor*)
força bruta	3.33	‘brute force’ (lit. *force brute*)
gatos pingados	0.00	‘a few people’ (lit. *cats dropped*)
gelo seco	2.33	‘dry ice’ (lit. *ice dry*)
golpe baixo	2.03	‘low blow’ (lit. *punch low*)
governo federal	4.97	‘federal government’ (lit. *government federal*)
gripe aviária	3.11	‘avian flu’ (lit. *flu avian*)
gripe suína	2.48	‘swine flu’ (lit. *flu swine*)
guarda florestal	4.16	‘forest ranger’ (lit. *guard forest*)
jogo duro	1.13	‘rough play’ (lit. *game hard*)
juízo final	3.60	‘doomsday’ (lit. *judgement final*)
leite integral	4.67	‘whole milk’ (lit. *milk integral*)
lista negra	1.60	‘black list’ (lit. *list black*)
livre-docente	2.63	‘professor’ (lit. *free lecturer*)
livro aberto	0.79	‘open book’ (lit. *book open*)
longa data	1.63	‘longtime’ (lit. *date long*)
longa-metragem	0.96	‘feature film’ (lit. *long length/footage*)
lua cheia	3.52	‘full moon’ (lit. *moon full*)
lua nova	1.40	‘new moon’ (lit. *moon new*)
lugar comum	1.52	‘cliché’ (lit. *place common*)
magia negra	1.72	‘black magic’ (lit. *magic black*)
mar aberto	2.87	‘open sea’ (lit. *sea open*)
maré alta	4.03	‘high tide’ (lit. *tide high*)
maré baixa	4.18	‘low tide’ (lit. *tide low*)
massa cinzenta	1.69	‘grey matter’ (lit. *mass grey*)
mau contato	2.84	‘faulty contact’ (lit. *bad contact*)
mau humor	4.29	‘bad mood’ (lit. *bad humour*)
mau olhado	1.97	‘evil eye’ (lit. *bad glance*)
mercado negro	1.06	‘black market’ (lit. *black market*)
mesa redonda	1.10	‘round table’ (lit. *table round*)
montanha russa	0.31	‘roller coaster’ (lit. *mountain Russian*)
má fé	1.62	‘bad faith’ (lit. *bad faith*)
máquina virtual	3.76	‘virtual machine’ (lit. *machine virtual*)
mão fechada	1.06	‘stingy’ (lit. *hand closed*)
navio negreiro	3.52	‘slave ship’ (lit. *ship black-slave*)
novo mundo	2.29	‘new world’ (lit. *new world*)
novo rico	3.62	‘new rich, new money’ (lit. *new rich*)
nó cego	0.74	‘difficult situation’ (lit. *knot blind*)
núcleo atômico	4.93	‘atomic nucleus’ (lit. *nucleus atomic*)
olho gordo	0.28	‘evil eye’ (lit. *eye fat*)
olho mágico	0.27	‘peephole’ (lit. *eye magic*)
olho nu	2.15	‘naked eye’ (lit. *eye naked*)
ovelha negra	0.45	‘black sheep’ (lit. *sheep black*)
papel higiênico	4.27	‘toilet paper’ (lit. *paper hygienic*)
paraíso fiscal	1.47	‘tax haven’ (lit. *paradise fiscal*)
pastor alemão	0.90	‘German shepherd’ (lit. *shepherd German*)
pau mandado	0.30	‘subservient, stooge’ (lit. *stick ordered*)
pavio curto	0.80	‘short-tempered’ (lit. *fuse short*)
pente fino	0.53	‘careful research’ (lit. *comb thin*)
peso morto	0.90	‘dead weight’ (lit. *weight dead*)
planta baixa	0.74	‘floor plan’ (lit. *plant short*)
ponto cego	1.92	‘blind spot’ (lit. *point blind*)

Compounds	$c_{\mathbf{WC}}$	Gloss
ponto forte	1.51	‘strong point’ (lit. *point strong*)
ponto fraco	2.27	‘weak point’ (lit. *point weak*)
poção mágica	3.29	‘magic potion’ (lit. *potion magic*)
prato feito	3.14	‘blue-plate special’ (lit. *plate ready-made*)
primeira infância	3.70	‘early childhood’ (lit. *first infancy*)
primeira-mão	0.71	‘first hand’ (lit. *first hand*)
primeira necessidade	3.97	‘first necessity’ (lit. *first necessity*)
primeira-dama	1.52	‘first lady’ (lit. *first dame*)
primeiro-ministro	2.87	‘first minister’ (lit. *first minister*)
primeiro plano	2.00	‘forefront’ (lit. *first plan*)
processo seletivo	4.78	‘selection process’ (lit. *process selective*)
pronto socorro	2.76	‘first-aid posts’ (lit. *ready aid*)
príncipe encantado	1.72	‘prince charming’ (lit. *prince enchanted*)
puro sangue	1.55	‘pure blood’ (lit. *pure blood*)
pão-duro	0.12	‘stingy’ (lit. *bread hard*)
pé quente	0.09	‘lucky’ (lit. *foot hot*)
pé-direito	0.10	‘ceiling height’ (lit. *foot right*)
pé frio	0.23	‘unlucky’ (lit. *foot cold*)
pólo aquático	2.87	‘water polo’ (lit. *aquatic pole/polo*)
quadro negro	2.94	‘blackboard’ (lit. *board black*)
queda livre	3.48	‘free fall’ (lit. *fall free*)
quinta categoria	1.00	‘second-rate’ (lit. *fifth category*)
rede social	3.27	‘social network’ (lit. *network social*)
regime político	4.00	‘political system’ (lit. *regime political*)
relógio analógico	4.92	‘analog clock’ (lit. *clock analog*)
relógio biológico	2.12	‘biological clock’ (lit. *clock biological*)
reta final	1.12	‘final stretch’ (lit. *straight line final*)
roda gigante	4.20	‘Ferris wheel’ (lit. *wheel giant*)
roleta russa	0.29	‘Russian roulette’ (lit. *roulette Russian*)
saia justa	0.37	‘tight spot’ (lit. *skirt tight*)
sala cirúrgica	4.47	‘operating room’ (lit. *room surgical*)
salão paroquial	4.52	‘parish hall’ (lit. *hall parish*)
sangue azul	0.15	‘blue-blooded’ (lit. *blood blue*)
sangue frio	0.52	‘cold-blooded’ (lit. *blood cold*)
sangue quente	0.87	‘hot-blooded’ (lit. *blood hot*)
secretária eletrônica	2.52	‘answering machine’ (lit. *secretary electronic*)
segundas intenções	2.11	‘ulterior motives’ (lit. *second intentions*)
segundo plano	1.55	‘aside, in the background’ (lit. *second plan*)
sentença judicial	4.67	‘court ruling’ (lit. *sentence judicial*)
sexto sentido	1.40	‘sixth sense’ (lit. *sixth sense*)
sinal verde	1.39	‘green lights’ (lit. *signal green*)
sistema político	4.36	‘political system’ (lit. *system political*)
sétima arte	2.19	‘seventh art’ (lit. *seventh art*)
tapete vermelho	3.76	‘red carpet’ (lit. *carpet red*)
tartaruga marinha	5.00	‘sea turtle’ (lit. *turtle marine*)
tela plana	4.96	‘flat screen TV’ (lit. *screen flat*)
tempo real	2.81	‘real time’ (lit. *time real*)
terceira idade	1.70	‘elder’ (lit. *third age*)
terceira pessoa	2.00	‘third person’ (lit. *third person*)
tiro livre	1.58	‘free kick (soccer)’ (lit. *shot free*)
trabalho braçal	3.55	‘manual labor’ (lit. *work arm*)
trabalho escravo	4.24	‘slave work’ (lit. *work slave*)
vaca louca	1.23	‘mad cow’ (lit. *cow crazy/mad*)
vinho branco	3.40	‘white wine’ (lit. *wine white*)
vinho tinto	4.08	‘red wine’ (lit. *wine dark-red*)
vista grossa	0.50	‘turn a blind eye’ (lit. *vision thick*)
viva voz	1.70	‘aloud’ (lit. *live voice*)
voto secreto	4.82	‘secret ballot’ (lit. *vote secret*)
vôo doméstico	3.41	‘domestic flight’ (lit. *flight domestic*)
vôo internacional	4.96	‘international flight’ (lit. *flight international*)
água doce	1.45	‘fresh water’ (lit. *water sweet*)
água mineral	4.21	‘mineral water’ (lit. *water mineral*)
ônibus executivo	2.63	‘minibus’ (lit. *bus executive*)