Unsupervised Compositionality Prediction of Nominal Compounds

Silvio Cordeiro
Federal University of Rio Grande do Sul
and Aix Marseille University, CNRS, LIS
silvioricardoc@gmail.com

Aline Villavicencio
University of Essex and
Federal University of Rio Grande do Sul
alinev@gmail.com

Marco Idiart
Federal University of Rio Grande do Sul
marco.idiart@gmail.com

Carlos Ramisch
Aix Marseille University, CNRS, LIS
carlos.ramisch@lis-lab.fr

Nominal compounds such as red wine and nut case display a continuum of compositionality, with varying contributions from the components of the compound to its semantics. This article proposes a framework for compound compositionality prediction using distributional semantic models, evaluating to what extent they capture idiomaticity compared to human judgments. For evaluation, we introduce data sets containing human judgments in three languages: English, French, and Portuguese. The results obtained reveal a high agreement between the models and human predictions, suggesting that they are able to incorporate information about idiomaticity. We also present an in-depth evaluation of various factors that can affect prediction, such as model and corpus parameters and compositionality operations. General crosslingual analyses reveal the impact of morphological variation and corpus size in the ability of the model to predict compositionality, and of a uniform combination of the components for best results.

1. Introduction

It is a universally acknowledged assumption that the meaning of phrases, expressions, or sentences can be determined by the meanings of their parts and by the rules used

Submission received: 4 December 2017; revised version received: 22 June 2018; accepted for publication: 8 August 2018.

doi:10.1162/COLI_a_00341

to combine them. Part of the appeal of this **principle of compositionality**[1] is that it implies that a meaning can be assigned even to a new sentence involving an unseen combination of familiar words (Goldberg 2015). Indeed, for natural language processing (NLP), this is an attractive way of linearly deriving the meaning of larger units from their components, performing the semantic interpretation of any text.

For representing the meaning of individual words and their combinations in computational systems, **distributional semantic models (DSMs)** have been widely used. 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 DSMs, words are usually represented as vectors that, to some extent, capture cooccurrence patterns in corpora (Lin 1998; Landauer, Foltz, and Laham 1998; Mikolov et al. 2013; Baroni, Dinu, and Kruszewski 2014). Evaluation of DSMs has focused on obtaining accurate semantic representations for words, and state-of-the-art models are already capable of obtaining a high level of agreement with human judgments for predicting synonymy or similarity between words (Freitag et al. 2005; Camacho-Collados, Pilehvar, and Navigli 2015; Lapesa and Evert 2017) and for modeling syntactic and semantic analogies between word pairs (Mikolov, Yih, and Zweig 2013). These representations for individual words can also be combined to create representations for larger units such as phrases, sentences, and even whole documents, using simple additive and multiplicative vector operations (Mitchell and Lapata 2010; Reddy, McCarthy, and Manandhar 2011; Mikolov et al. 2013; Salehi, Cook, and Baldwin 2015), syntax-based lexical functions (Socher et al. 2012), or matrix and tensor operations (Baroni and Lenci 2010; Bride, Van de Cruys, and Asher 2015). However, it is not clear to what extent this approach is adequate in the case of idiomatic **multiword expressions (MWEs)**. MWEs fall into a wide spectrum of compositionality; that is, some MWEs are more compositional (e.g., *olive oil*) while others are more idiomatic (Sag et al. 2002; Baldwin and Kim 2010). In the latter case, the meaning of the MWE may not be straightforwardly related to the meanings of its parts, creating a challenge for the principle of compositionality (e.g., *snake oil* as a product of questionable benefit, not necessarily an *oil* and certainly not extracted from *snakes*).

In this article, we discuss approaches for automatically detecting to what extent the meaning of an MWE can be directly computed from the meanings of its component words, represented using DSMs. We evaluate how accurately DSMs can model the semantics of MWEs with various levels of compositionality compared to human judgments. Since MWEs encompass a large amount of related but distinct phenomena, we focus exclusively on a subcategory of MWEs: **nominal compounds**. They represent an ideal case study for this work, thanks to their relatively homogeneous syntax (as opposed to other categories of MWEs such as verbal idioms) and their pervasiveness in language. We assume that models able to predict the compositionality of nominal compounds could be generalized to other MWE categories by addressing their variability in future work. Furthermore, to determine to what extent these approaches are also adequate cross-lingually, we evaluate them in three languages: English, French, and Portuguese.

Given that MWEs are frequent in languages (Sag et al. 2002), identifying idiomaticity and producing accurate semantic representations for compositional and idiomatic cases is of relevance to NLP tasks and applications that involve some form of semantic processing, including semantic parsing (Hwang et al. 2010; Jagfeld and van der Plas 2015), word sense disambiguation (Finlayson and Kulkarni 2011; Schneider et al. 2016),

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

and machine translation (Ren et al. 2009; Carpuat and Diab 2010; Cap et al. 2015; Salehi et al. 2015). Moreover, the evaluation of DSMs on tasks involving MWEs, such as compositionality prediction, has the potential to drive their development towards new directions.

The main hypothesis of our work is that, if the meaning of a compositional nominal compound can be derived from a combination of its parts, this translates in DSMs as similar vectors for a compositional nominal compound and for the combination of the vectors of its parts using some vector operation, that we refer to as **composition function**. Conversely we can use the lack of similarity between the nominal compound vector representation and a combination of its parts to detect idiomaticity. Furthermore, we hypothesize that accuracy in predicting compositionality depends both on the characteristics of the DSMs used to represent expressions and their components and on the composition function adopted. Therefore, we have built 684 DSMs and performed an extensive evaluation, involving over 9,072 analyses, investigating various types of DSMs, their configurations, the corpora used to train them, and the composition function used to build vectors for expressions.[2]

This article is structured as follows. Section 2 presents related work on distributional semantics, compositionality prediction, and nominal compounds. Section 3 presents the data sets created for our evaluation. Section 4 describes the compositionality prediction framework, along with the composition functions which we evaluate. Section 5 specifies the experimental setup (corpora, DSMs, parameters, and evaluation measures). Section 6 presents the overall results of the evaluated models. Sections 7 and 8 evaluate the impact of DSM and corpus parameters, and of composition functions on compositionality prediction. Section 9 discusses system predictions through an error analysis. Section 10 summarizes our conclusions. Appendix A contains a glossary, Appendix B presents extra sanity-check experiments, Appendix C contains the questionnaire used for data collection, and Appendices D, E, and F list the compounds in the data sets.

2. Related Work

The literature on distributional semantics is extensive (Lin 1998; Turney and Pantel 2010; Baroni and Lenci 2010; Mohammad and Hirst 2012), so we provide only a brief introduction here, underlining their most relevant characteristics to our framework (Section 2.1). Then, we define compositionality prediction and discuss existing approaches, focusing on distributional techniques for multiword expressions (Section 2.2). Our framework is evaluated on nominal compounds, and we discuss their relevant properties (Section 2.3) along with existing data sets for evaluating compositionality prediction (Section 2.4).

[2] This article significantly extends and updates previous publications:

1. We consolidate the description of the data sets introduced in Ramisch et al. (2016) and Ramisch, Cordeiro, and Villavicencio (2016) by adding details about data collection, filtering, and results of a thorough analysis studying the correlation between compositionality and related variables.
2. We extend the compositionality prediction framework described in Cordeiro, Ramisch, and Villavicencio (2016) by adding and evaluating new composition functions and DSMs.
3. We extend the evaluation reported in Cordeiro et al. (2016) not only by adding Portuguese, but also by evaluating additional parameters: corpus size, composition functions, and new DSMs.

2.1 Distributional Semantic Models

Distributional semantic models (DSMs) use context information to represent the meaning of lexical units as vectors. These vectors are built assuming the **distributional hypothesis**, whose central idea is that the meaning of a word can be learned based on the contexts where it appears—or, as popularized by Firth (1957), "you shall know a word by the company it keeps."

Formally, a DSM attempts to encode the meaning of each target word w_i of a vocabulary V as a vector of real numbers $\mathbf{v}(w_i)$ in $\mathbb{R}^{|V|}$. Each component of $\mathbf{v}(w_i)$ is a function of the co-occurrence between w_i and the other words in the vocabulary (its contexts w_c). This function can be simply a co-occurrence count $c(w_i, w_c)$, or some measure of the association between w_i and each w_c, such as pointwise mutual information (PMI, Church and Hanks [1990], Lin [1999]) or positive PMI (PPMI, Baroni, Dinu, and Kruszewski [2014]; Levy, Goldberg, and Dagan [2015]).

In DSMs, co-occurrence can be defined as two words co-occurring in the same document, sentence, or sentence fragment in a corpus. Intrasentential models are often based on a sliding window; that is, a context word w_c co-occurs within a certain window of W words around the target w_i. Alternatively, co-occurrence can also be based on syntactic relations obtained from parsed corpora, where a context word w_c appears within specific syntactic relations with w_i (Lin 1998; Padó and Lapata 2007; Lapesa and Evert 2017).

The set of all vectors $\mathbf{v}(w_i)$, $\forall w_i \in V$ can be represented as a sparse co-occurrence matrix $V \times V \rightarrow \mathbb{R}$. Given that most word pairs in this matrix co-occur rarely (if ever), a threshold on the number of co-occurrences is often applied to discard irrelevant pairs. Additionally, co-occurrence vectors can be transformed to have a significantly smaller number of dimensions, converting vectors in $\mathbb{R}^{|V|}$ into vectors in $\mathbb{R}^d$, with $d \ll |V|$.[3] Two solutions are commonly employed in the literature. The first one consists in using context thresholds, where all target–context pairs that do not belong to the top-d most relevant pairs are discarded (Salehi, Cook, and Baldwin 2014; Padró et al. 2014b). The second solution consists in applying a dimensionality reduction technique such as singular value decomposition on the co-occurrence matrix where only the d largest singular values are retained (Deerwester et al. 1990). Similar techniques focus on the factorization of the logarithm of the co-occurrence matrix (Pennington, Socher, and Manning 2014) and on alternative factorizations of the PPMI matrix (Salle, Villavicencio, and Idiart 2016).

Alternatively, DSMs can be constructed by training a neural network to predict target–context relationships. For instance, a network can be trained to predict a target word w_i among all possible words in V given as input a window of surrounding context words. This is known as the continuous bag-of-words model. Conversely, the network can try to predict context words for a target word given as input, and this is known as the skip-gram model (Mikolov et al. 2013). In both cases, the network training procedure allows encoding in the hidden layer semantic information about words as a side effect of trying to solve the prediction task. The weight parameters that connect the unity representing w_i with the d-dimensional hidden layer are taken as its vector representation $\mathbf{v}(w_i)$.

There are a number of factors that may influence the ability of a DSM to accurately learn a semantic representation. These include characteristics of the training corpus such

[3] After dimensionality reduction, nowadays word vectors are often called *word embeddings*.

as size (Mikolov, Yih, and Zweig 2013) as well as frequency thresholds and filters (Ferret 2013; Padró et al. 2014b), genre (Lapesa and Evert 2014), preprocessing (Padó and Lapata 2003, 2007), and type of context (window vs. syntactic dependencies) (Agirre et al. 2009; Lapesa and Evert 2017). Characteristics of the model include the choice of association and similarity measures (Curran and Moens 2002), dimensionality reduction strategies (Van de Cruys et al. 2012), and the use of subsampling and negative sampling techniques (Mikolov, Yih, and Zweig 2013). However, the particular impact of these factors on the quality of the resulting DSM may be heterogeneous and depends on the task and model (Lapesa and Evert 2014). Because there is no consensus about a single optimal model that works for all tasks, we compare a variety of models (Section 5) to determine which are best suited for our compositionality prediction framework.

2.2 Compositionality Prediction

Before adopting the principle of compositionality to determine the meaning of a larger unit, such as a phrase or multiword expression (MWE), it is important to determine whether it is idiomatic or not.[4] This problem, known as **compositionality prediction**, can be solved using methods that measure directly the extent to which an expression is constructed from a combination of its parts, or indirectly via language-dependent properties of MWEs linked to idiomaticity like the degree of determiner variability and morphological flexibility (Fazly, Cook, and Stevenson 2009; Tsvetkov and Wintner 2012; Salehi, Cook, and Baldwin 2015; Köper and Schulte im Walde 2016). In this article, we focus on direct prediction methods in order to evaluate the target languages under similar conditions. Nonetheless, this does not exclude the future integration of information used by indirect prediction methods, as a complement to the methods discussed here.

For direct prediction methods, three ingredients are necessary. First, we need vector representations of single-word meanings, such as those built using DSMs (Section 2.1). Second, we need a mathematical model of how the compositional meaning of a phrase is calculated from the meanings of its parts. Third, we need the compositionality measure itself, which estimates the similarity between the compositionally constructed meaning of a phrase and its observed meaning, derived from corpora. There are a number of alternatives for each of the ingredients, and throughout this article we call a specific choice of the three ingredients a **compositionality prediction configuration**.

Regarding the second ingredient, that is, the mathematical model of compositional meaning, the most natural choice is the **additive model** (Mitchell and Lapata 2008). In the additive model, the compositional meaning of a phrase $w_1 w_2 \ldots w_n$ is calculated as a linear combination of the word vectors of its components: $\sum_i \beta_i \mathbf{v}(w_i)$, where $\mathbf{v}(w_i)$ is a d-dimensional vector for each word w_i, and the β_i coefficients assign different weights to the representation of each word (Reddy, McCarthy, and Manandhar 2011; Schulte im Walde, Müller, and Roller 2013; Salehi, Cook, and Baldwin 2015). These weights can capture the asymmetric contribution of each of the components to the semantics of the whole phrase (Bannard, Baldwin, and Lascarides 2003; Reddy, McCarthy, and Manandhar 2011). For example, in *flea market*, it is the head (*market*) that has a clear contribution to the overall meaning, whereas in *couch potato* it is the modifier (*couch*).

The additive model can be generalized to use a matrix of multiplicative coefficients, which can be estimated through linear regression (Guevara 2011). This model can be

[4] The task of determining whether a phrase is compositional is closely related to MWE discovery (Constant et al. 2017), which aims to automatically extract MWE lists from corpora.

further modified to learn polynomial projections of higher degree, with quadratic projections yielding particularly promising results (Yazdani, Farahmand, and Henderson 2015). These models come with the caveat of being supervised, requiring some amount of pre-annotated data in the target language. Because of these requirements, our study focuses on unsupervised compositionality prediction methods only, based exclusively on automatically POS-tagged and lemmatized monolingual corpora.

Alternatives to the additive model include the multiplicative model and its variants (Mitchell and Lapata 2008). However, results suggest that this representation is inferior to the one obtained through the additive model (Reddy, McCarthy, and Manandhar 2011; Salehi, Cook, and Baldwin 2015). Recent work on predicting intracompound semantics also supports that additive models tend to yield better results than multiplicative models (Hartung et al. 2017).

The third ingredient is the measure of similarity between the compositionally constructed vector and its actual corpus-based representation. Cosine similarity is the most commonly used measure for compositionality prediction in the literature (Schone and Jurafsky 2001; Reddy, McCarthy, and Manandhar 2011; Schulte im Walde, Müller, and Roller 2013; Salehi, Cook, and Baldwin 2015). Alternatively, one can calculate the overlap between the distributional neighbors of the whole phrase and those of the component words (McCarthy, Keller, and Carroll 2003), or the number of single-word distributional neighbors of the whole phrase (Riedl and Biemann 2015).

2.3 Nominal Compounds

Instead of covering compositionality prediction for MWEs in general, we focus on a particular category of phenomena represented by nominal compounds. We define a **nominal compound** as a syntactically well-formed and conventionalized noun phrase containing two or more content words, whose head is a noun.[5] They are conventionalized (or institutionalized) in the sense that their particular realization is statistically idiosyncratic, and their constituents cannot be replaced by synonyms (Sag et al. 2002; Baldwin and Kim 2010; Farahmand, Smith, and Nivre 2015). Their semantic interpretation may be straightforwardly compositional, with contributions from both elements (e.g., *climate change*), partly compositional, with contribution mainly from one of the elements (e.g., *grandfather clock*), or idiomatic (e.g., *cloud nine*) (Nakov 2013).

The syntactic realization of nominal compounds varies across languages. In English, they are often expressed as a sequence of two nouns, with the second noun as the syntactic head, modified by the first noun. This is the most frequently annotated POS-tag pattern in the MWE-annotated DiMSUM English corpus (Schneider et al. 2016). In French and Portuguese, they often assume the form of adjective–noun or noun–adjective pairs, where the adjective modifies the noun. Examples of such constructions include the adjective–noun compound FR *petite annonce* (lit. *small announcement* 'classified ad') and the noun–adjective compound PT *buraco negro* (lit. *hole black* 'black hole').[6] Additionally, compounds may also involve prepositions linking the modifier with the head, as in the case of FR *cochon d'Inde* (lit. *pig of India* 'guinea pig') and PT *dente de leite* (lit. *tooth of milk* 'milk tooth'). Because prepositions are highly polysemous and their representation in DSMs is tricky, we do not include compounds containing prepositions

[5] The terms *noun compound* and *compound noun* are usually reserved for nominal compounds formed by sequences of nouns only, typical of Germanic languages but not frequent in Romance languages.

[6] In this article, examples are preceded by their language codes: EN for English, FR for French, and PT for Brazilian Portuguese. In the absence of a language code, English is implied.

in this article. Hence, we focus on 2-word nominal compounds of the form noun$_1$–noun$_2$ (in English), and noun–adjective and adjective–noun (in the three languages).

Regarding the meaning of nominal compounds, the implicit relation between the components of compositional compounds can be described in terms of free paraphrases involving verbs, such as *flu virus* as *virus that causes/creates flu* (Nakov 2008),[7] or prepositions, such as *olive oil* as *oil from olives* (Lauer 1995). These implicit relations can often be seen explicitly in the equivalent expressions in other languages (e.g., FR *huile d'olive* and PT *azeite de oliva* for EN *olive oil*).

Alternatively, the meaning of compositional nominal compounds can be described using a closed inventory of relations which make the role of the modifier explicit with respect to the head noun, including syntactic tags such as *subject* and *object*, and semantic tags such as *instrument* and *location* (Girju et al. 2005). The degree of compositionality of a nominal compound can also be represented using numerical scores (Section 2.4) to indicate to what extent the component words allow predicting the meaning of the whole (Reddy, McCarthy, and Manandhar 2011; Roller, Schulte im Walde, and Scheible 2013; Salehi et al. 2015). The latter is the representation that we adopted in this article.

2.4 Numerical Compositionality Data sets

The evaluation of compositionality prediction models can be performed extrinsically or intrinsically. In extrinsic evaluation, compositionality information can be used to decide how a compound should be treated in NLP systems such as machine translation or text simplification. For instance, for machine translation, idiomatic compounds need to be treated as atomic phrases, as current methods of morphological compound processing cannot be applied to them (Stymne, Cancedda, and Ahrenberg 2013; Cap et al. 2015).

Although potentially interesting, extrinsic evaluation is not straightforward, as results may be influenced both by the compositionality prediction model and by the strategy for integration of compositionality information into the NLP system. Therefore, most related work focuses on an intrinsic evaluation, where the compositionality scores produced by a model are compared to a gold standard, usually a data set where nominal compound semantics have been annotated manually. Intrinsic evaluation thus requires the existence of data sets where each nominal compound has one (or several) numerical scores associated with it, indicating its compositionality. Annotations can be provided by expert linguist annotators or by crowdsourcing, often requiring that several annotators judge the same compound to reduce the impact of subjectivity on the scores. Relevant compositionality data sets of this type are listed below, some of which were used in our experiments.

- Reddy, McCarthy, and Manandhar (2011) collected judgments for a set of 90 English noun–noun (e.g., *zebra crossing*) and adjective–noun (e.g., *sacred cow*) 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. The data set was built through crowdsourcing, and the final scores are the average of 30 judgments per compound.This data set will be referred to as *Reddy* in our experiments.

[7] Nakov (2008) also proposes a method for automatically extracting paraphrases from the web to classify nominal compounds. This was extended in a SemEval 2013 task, where participants had to rank free paraphrases according to the semantic relations in the compounds (Hendrickx et al. 2013).

- Farahmand, Smith, and Nivre (2015) collected judgments for 1,042 English noun–noun compounds. Each compound has binary judgments regarding non-compositionality and conventionalization given by four expert annotators (both native and non-native speakers). A hard threshold is applied so that compounds are considered as noncompositional if at least two annotators say so (Yazdani, Farahmand, and Henderson 2015), and the total compositionality score is given by the sum of the four binary judgments. This data set will be referred to as *Farahmand* in our experiments.
- Kruszewski and Baroni (2014) built the *Norwegian Blue Parrot* data set, containing judgments for modifier-head phrases in English. The judgments consider whether the phrase is (1) an instance of the concept denoted by the head (e.g., *dead parrot* and *parrot*) and (2) a member of the more general concept that includes the head (e.g., *dead parrot* and *pet*), along with typicality ratings, with 5,849 judgments in total.
- Roller, Schulte im 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 and Schulte im Walde 2014).
- Schulte im Walde et al. (2016) collected judgments for a set of 868 German noun–noun compounds, including human judgments of compositionality on a scale of 1 to 6. Compounds are judged by multiple annotators, and the final compositionality score is the average across annotators. The data set is also annotated for in-corpus frequency, productivity, and ambiguity, and a subset of 180 compounds has been selected for balancing these variables. The annotations were performed by the authors, linguists, and through crowdsourcing. For the balanced subset of 180 compounds, compositionality annotations were performed by experts only, excluding the authors.

For a multilingual evaluation, in this work, we construct two data sets, one for French and one for Portuguese compounds, and extend the *Reddy* data set for English using the same protocol as Reddy, McCarthy, and Manandhar (2011).

3. Creation of a Multilingual Compositionality Data set

In Section 3.1, we describe the construction of data sets of 180 compounds for French (*FR-comp*) and Portuguese (*PT-comp*). For English, the complete data set contains 280 compounds, of which 190 are new and 90 come from the *Reddy* data set. We use 180 of these (*EN-comp*) for cross-lingual comparisons (90 from the original *Reddy* data set combined with 90 new ones from *EN-comp*$_{90}$), and 100 new compounds as held-out data (*EN-comp*$_{Ext}$), to evaluate the robustness of the results obtained (Section 6.3). These data sets containing compositionality scores for 2–word nominal compounds are used to evaluate our framework (Section 4), and we discuss their characteristics in Section 3.2.[8]

[8] For English, only *EN-comp*$_{90}$ and *EN-comp*$_{Ext}$ (90 and 100 new compounds, respectively) are considered. *Reddy* (included in *EN-comp*) is analyzed in Reddy, McCarthy, and Manandhar (2011).

3.1 Data Collection

For each of the target languages, we collected, via crowdsourcing, a set of numerical scores corresponding to the level of compositionality of the target nominal compounds. We asked non-expert participants to judge each compound considering three sentences where the compound occurred. After reading the sentences, participants assess the degree to which the meaning of the compound is related to the meanings of its parts. This follows from the assumption that a fully compositional compound will have an interpretation whose meaning stems from both words (e.g., *lime tree* as a *tree* of *limes*), while a fully idiomatic compound will have a meaning that is unrelated to its components (e.g., *nut case* as an eccentric person).

Our work follows the protocol proposed by Reddy, McCarthy, and Manandhar (2011), where compositionality is explained in terms of the literality of the individual parts. This type of indirect annotation does not require expert linguistic knowledge, and still provides reliable data, as we show later. For each language, data collection involved four steps: compound selection, sentence selection, questionnaire design, and data aggregation.

Compound Selection. For each data set, we manually selected nominal compounds from dictionaries, corpus searches, and by linguistic introspection, maintaining an equal proportion of compounds that are compositional, partly compositional, and idiomatic.[9] We considered them to be compositional if their semantics are related to both components (e.g., *benign tumor*), partly compositional if their semantics are related to only one of the components (e.g., *grandfather clock*), and idiomatic if they are not directly related to either (e.g., *old flame*). This preclassification was used only to select a balanced set of compounds and was not shown to the participants nor used at any later stage. For all languages, all compounds are required to have a head that is unambiguously a noun, and additionally for French and Portuguese, all compounds have an adjective as modifier.

Sentence Selection. Compounds may be polysemous (e.g., FR *bras droit* may mean *most reliable helper* or literally *right arm*). To avoid any potential sense uncertainty, each compound was presented to the participants with the same sense in three sentences. These sentences were manually selected from the WaC corpora: ukWaC (Baroni et al. 2009), frWaC, and brWaC (Boos, Prestes, and Villavicencio 2014), presented in detail in Section 5.

Questionnaire Design. For each compound, after reading three sentences, participants are asked to:

- provide synonyms for the compound in these sentences. The synonyms are used as additional validation of the quality of the judgments: if unrelated words are provided, the answers are discarded.
- assess the contribution of the *head* noun to the meaning of the compound (e.g., is a *busy bee* always literally a *bee*?)

[9] We have not attempted to select compounds that are translations of each other, as a compound in a given language may be realized differently in the other languages.

- assess the contribution of the *modifier* noun or adjective to the meaning of the compound (e.g., is a *busy bee* always literally *busy*?)
- assess the degree to which the *compound* can be seen as a combination of its parts (e.g., is a *busy bee* always literally a *bee* that is *busy*?)

Participants answer the last three items using a Likert scale from 0 (idiomatic/non-literal) to 5 (compositional/literal), following Reddy, McCarthy, and Manandhar (2011). To qualify for the task, participants had to submit demographic information confirming that they are native speakers, and to undergo training in the form of four example questions with annotated answers in an external form (see Appendix C for details).

Data Aggregation. For English and French, we collected answers using Amazon Mechanical Turk (AMT), manually removing answers that were not from native speakers or where the synonyms provided were unrelated to the target compound sense. Because AMT has few Brazilian Portuguese native speakers, we developed an in-house web interface for the questionnaire, which was sent out to Portuguese-speaking NLP mailing lists.

For a given compound and question we calculate aggregated scores as the arithmetic averages of all answers across participants. We will refer to these averaged scores as the **human compositionality score** (hc)s. We average the answers to the three questions independently, generating three scores: $\mathrm{hc}_{\mathbf{H}}$ for the head noun, $\mathrm{hc}_{\mathbf{M}}$ for the modifier, and $\mathrm{hc}_{\mathbf{HM}}$ for the whole compound. In our framework, we try to predict $\mathrm{hc}_{\mathbf{HM}}$ automatically (Section 5). To assess the variability of the answers (Section 3.2.1), we also calculate the standard deviation across participants for each question ($\sigma_{\mathbf{H}}$, $\sigma_{\mathbf{M}}$, and $\sigma_{\mathbf{HM}}$).

The list of compounds, their translations, glosses, and compositionality scores are given in Appendices D (*EN-comp*$_{90}$ and *EN-comp*$_{Ext}$), E (*FR-comp*), and F (*PT-comp*).[10]

3.2 Data set Analysis

In this section, we present different measures of agreement among participants (Section 3.2.1) and examine possible correlations between compositionality scores, familiarity, and conventionalization (Section 3.2.2) in the data sets created for this article.

3.2.1 Measuring Data set Quality. To assess the quality of the collected human compositionality scores, we use standard deviation and inter-annotator agreement scores.

Standard Deviation ($\overline{\sigma}$ *and* $P_{\sigma>1.5}$). The standard deviation (σ) of the participants' answers can be used as an indication of their agreement: for each compound and for each of the three questions, small σ values suggest greater agreement. In addition, if the instructions are clear, σ can also be seen as an indication of the level of difficulty of the task. In other words, all other things being equal, compounds with larger σ can be considered intrinsically harder to analyze by the participants. For each data set, we consider two aggregated metrics based on σ:

- $\overline{\sigma}$ — The average of σ in the data set.
- $P_{\sigma>1.5}$ — The proportion of compounds whose σ is higher than 1.5.

[10] Freely available at: `http://pageperso.lis-lab.fr/~carlos.ramisch/?page=downloads/compounds`

Table 1
Average number of answers per compound $\overline{n}$, average standard deviation $\overline{\sigma}$, proportion of high standard deviation $P_{\sigma>1.5}$, for the compound (**HM**), head (**H**), and modifier (**M**).

Data set	$\overline{n}$	$\overline{\sigma_{\mathbf{HM}}}$	$\overline{\sigma_{\mathbf{H}}}$	$\overline{\sigma_{\mathbf{M}}}$	$P_{\sigma_{\mathbf{HM}}>1.5}$	$P_{\sigma_{\mathbf{H}}>1.5}$	$P_{\sigma_{\mathbf{M}}>1.5}$
FR-comp	14.9	1.15	1.08	1.21	22.78%	24.44%	30.56%
PT-comp	31.8	1.22	1.09	1.20	14.44%	17.22%	19.44%
EN-comp$_{90}$	18.8	1.17	1.05	1.18	18.89%	16.67%	27.78%
EN-comp$_{Ext}$	22.6	1.21	1.27	1.16	17.00%	29.00%	18.00%
Reddy	28.4	0.99	0.94	0.89	5.56%	11.11%	8.89%

Table 1 presents the result of these metrics when applied to our in-house data sets, as well as to the original *Reddy* data set. The column $\overline{n}$ indicates the average number of answers per compound, while the other six columns present the values of $\overline{\sigma}$ and $P_{\sigma>1.5}$ for compound (**HM**), head-only (**H**), and modifier-only (**M**) scores.

These values are below what would be expected for random decisions ($\overline{\sigma}_{\text{rand}} \simeq 1.71$, for the Likert scale). Although our data sets exhibit higher variability than *Reddy*, this may be partly due to the application of filters done by Reddy, McCarthy, and Manandhar (2011) to remove outliers.[11] These values could also be due to the collection of fewer answers per compound for some of the data sets. However, there is no clear tendency in the variation of the standard deviation of the answers and the number of participants n. The values of $\overline{\sigma}$ are quite homogeneous, ranging from 1.05 for *EN-comp*$_{90}$ (head) to 1.27 for *EN-comp*$_{Ext}$ (head). The low agreement for modifiers may be related to a greater variability in semantic relations between modifiers and compounds: these include material (e.g., *brass ring*), attribute (e.g., *black cherry*), and time (e.g., *night owl*).

Figure 1(a) shows standard deviation ($\sigma_{\mathbf{HM}}$, $\sigma_{\mathbf{H}}$, and $\sigma_{\mathbf{M}}$) for each compound of *FR-comp* as a function of its average compound score $\text{hc}_{\mathbf{HM}}$.[12] For all three languages, greater agreement was found for compounds at the extremes of the compositionality scale (fully compositional or fully idiomatic) for all scores. These findings can be partly explained by end-of-scale effects, that result in greater variability for the intermediate scores in the Likert scale (from 1 to 4) that correspond to the partly compositional cases. Hence, we expect that it will be easier to predict the compositionality of idiomatic/compositional compounds than of partly compositional ones.

Inter-Annotator Agreement (α). To measure inter-annotator agreement of multiple participants, taking into account the distance between the ordinal ratings of the Likert scale, we adopt the α score (Artstein and Poesio 2008). The α score is more appropriate for ordinal data than traditional agreement scores for categorical data, such as Cohen's and Fleiss' κ (Cohen 1960; Fleiss and Cohen 1973). However, due to the use of crowdsourcing, most participants rated only a small number of compounds with very limited chance of overlap among them: the average number of answers per participant is 13.6 for *EN-comp*$_{90}$, 10.2 for *EN-comp*$_{Ext}$, 33.7 for *FR-comp*, and 53.5 for *PT-comp*. Because the

[11] Participants with negative correlation with the mean, and answers farther than ±1.5 from the mean.

[12] Only *FR-comp* is shown as the other data sets display similar patterns.

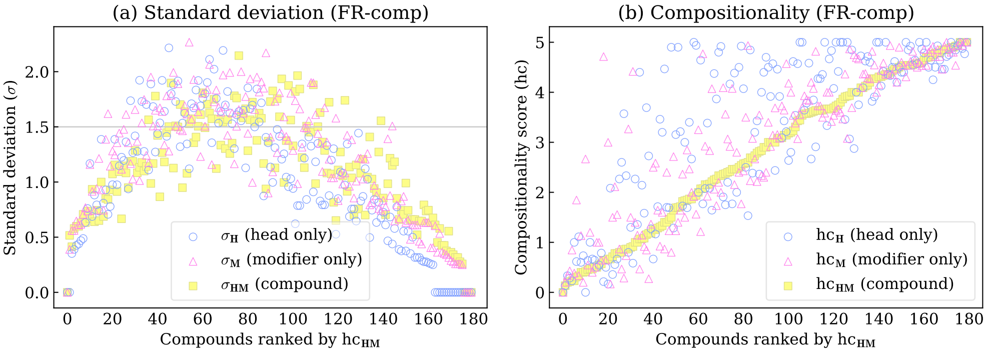


Figure 1
Left: Standard deviations ($\sigma_{\mathbf{H}}$, $\sigma_{\mathbf{M}}$, and $\sigma_{\mathbf{HM}}$) as a function of $hc_{\mathbf{HM}}$ in *FR-comp*. Right: Average compositionality ($hc_{\mathbf{H}}$, $hc_{\mathbf{M}}$, and $hc_{\mathbf{HM}}$) as a function of $hc_{\mathbf{HM}}$ in *FR-comp*.

α score assumes that each participant rates all the items, we focus on the answers provided by three of the participants, who rated the whole set of 180 compounds in *PT-comp*.

Using a linear distance schema between the answers,[13] we obtain an agreement of $\alpha = .58$ for head-only, $\alpha = .44$ for modifier-only, and $\alpha = .44$ for the whole compound. To further assess the difficulty of this task, we also calculate α for a single expert annotator, judging the same set of compounds after an interval of one month. The scores were $\alpha = .69$ for the head and $\alpha = .59$ for both the compound and for the modifier. The Spearman correlation between these two annotations performed by the same expert is $\rho = 0.77$ for $hc_{\mathbf{HM}}$. This can be seen as a qualitative upper bound for automatic compositionality prediction on *PT-comp*.

3.2.2 Compositionality, Familiarity, and Conventionalization. Figure 1(b) shows the average scores ($hc_{\mathbf{HM}}$, $hc_{\mathbf{H}}$, and $hc_{\mathbf{M}}$) for the compounds ranked according to the average compound score $hc_{\mathbf{HM}}$. Although this figure is for *FR-comp*, similar patterns were found for the other data sets. For all three languages, the human compositionality scores provide additional confirmation that the data sets are balanced, with the compound scores ($hc_{\mathbf{HM}}$) being distributed linearly along the scale. Furthermore, we have calculated the average $hc_{\mathbf{HM}}$ values separately for the compounds in each of the three compositionality classes used for compound selection: idiomatic, partly compositional and compositional (Section 3.1). These averages are, respectively, 1.0, 2.4, and 4.0 for *EN-comp*$_{90}$; 1.1, 2.4, and 4.2 for *EN-comp*$_{Ext}$; 1.3, 2.7, and 4.3 for *FR-comp*; and 1.3, 2.5, and 3.9 for *PT-comp*, indicating that our attempt to select a balanced number of compounds from each class is visible in the collected $hc_{\mathbf{HM}}$ scores.

Additionally, the human scores also suggest an asymmetric impact of the non-literal parts over the compound: whenever participants judged an element of the compound as non-literal, the whole compound was also rated as idiomatic. Thus, most head and modifier scores ($hc_{\mathbf{H}}$ and $hc_{\mathbf{M}}$) are close to or above the diagonal line in Figure 1(b). In other words, a component of the compound is seldom rated as less literal than the compositionality of the whole compound $hc_{\mathbf{HM}}$, although the opposite is more common.

[13] A disagreement between answers *a* and *b* is weighted $|a - b|$.

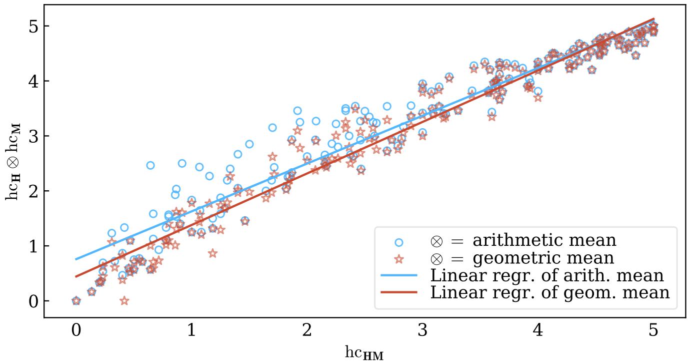


Figure 2
Relation between $\mathrm{hc}_\mathbf{H} \otimes \mathrm{hc}_\mathbf{M}$ and $\mathrm{hc}_\mathbf{HM}$ in *FR-comp*, using arithmetic and geometric means.

Table 2
Spearman ρ correlation between compositionality, frequency, and PMI for the three data sets.

Data set	frequency	PMI
FR-comp	0.598 ($p < 10^{-18}$)	0.164 ($p > 0.01$)
PT-comp	0.109 ($p > 0.1$)	0.076 ($p > 0.1$)
EN-comp$_{90}$	0.305 ($p < 10^{-2}$)	−0.024 ($p > 0.1$)
EN-comp$_{Ext}$	0.384 ($p < 10^{-5}$)	0.138 ($p > 0.1$)

To evaluate if it is possible to predict $\mathrm{hc}_\mathbf{HM}$ from the $\mathrm{hc}_\mathbf{H}$ and $\mathrm{hc}_\mathbf{M}$, we calculate the arithmetic and geometric means between $\mathrm{hc}_\mathbf{H}$ and $\mathrm{hc}_\mathbf{M}$ for each compound. Figure 2 shows the linear regression of both measures for *FR-comp*. The goodness of fit is $r^2_{\text{arith}} = .93$ for the arithmetic mean, and $r^2_{\text{geom}} = .96$ for the geometric mean, confirming that they are good predictors of $\mathrm{hc}_\mathbf{HM}$.[14] Thus, we assume that $\mathrm{hc}_\mathbf{HM}$ summarizes $\mathrm{hc}_\mathbf{H}$ and $\mathrm{hc}_\mathbf{M}$, and focus on predicting $\mathrm{hc}_\mathbf{HM}$ instead of $\mathrm{hc}_\mathbf{H}$ and $\mathrm{hc}_\mathbf{M}$ separately. These findings also inspired the pc_{arith} and pc_{geom} compositionality prediction functions (Section 4).

To examine whether there is an effect of the familiarity of a compound on hc scores, in particular if more idiomatic compounds need to be more familiar, we also calculated the correlation between the compositionality score for a compound $\mathrm{hc}_\mathbf{HM}$ and its frequency in a corpus, as a proxy for familiarity. In this case we used the WaC corpora and calculated the frequencies based on the lemmas. The results, in Table 2, show a statistically significant positive Spearman correlation of ρ = 0.305 for *EN-comp*$_{90}$, ρ = 0.384 for *EN-comp*$_{Ext}$, and ρ = 0.598 for *FR-comp*, indicating that, contrary to our expectations, compounds that are more frequent tend to be assigned higher compositionality scores. However, frequency alone is not enough to predict compositionality, and further investigation is needed to determine if compositionality and frequency are also correlated with other factors.

[14] r^2_{arith} and r^2_{geom} are .91 and .96 in *PT-comp*, .90 and .96 in *EN-comp*$_{90}$, and .92 and .95 in *EN-comp*$_{Ext}$.

We also analyzed the correlation between compositionality and conventionalization to determine if more idiomatic compounds correspond to more conventionalized ones. We use PMI (Church and Hanks 1990) as a measure of conventionalization, as it indicates the strength of association between the components (Farahmand, Smith, and Nivre 2015). We found no statistically significant correlation between compositionality and PMI.

4. Compositionality Prediction Framework

We propose a **compositionality prediction framework**[15] including the following elements: a **DSM**, created from corpora using existing state-of-the-art models that generate **corpus-derived vectors**[16] for compounds w_1w_2 and for their components w_1 and w_2; a **composition function**; and a set of **predicted compositionality scores** (pc). The framework, shown in Figure 3, is evaluated by measuring the correlation between the scores predicted by the models (pc) and the human compositionality scores (hc) for the list of compounds in our data sets (Section 3). The predicted compositionality scores are obtained from the cosine similarity between the corpus-derived vector of the compound, $\mathbf{v}(w_1w_2)$, and the **compositionally constructed vector**, $\mathbf{v}_\beta(w_1, w_2)$:

$$\text{pc}_\beta(w_1w_2) = \cos(\,\mathbf{v}(w_1w_2),\ \mathbf{v}_\beta(w_1, w_2)\,).$$

For $\mathbf{v}_\beta(w_1, w_2)$, we use the additive model (Mitchell and Lapata 2008), in which the composition function is a weighted linear combination:

$$\mathbf{v}_\beta(w_1w_2) = \beta\frac{\mathbf{v}(w_{head})}{||\mathbf{v}(w_{head})||} + (1-\beta)\frac{\mathbf{v}(w_{mod})}{||\mathbf{v}(w_{mod})||},$$

where w_{head} (or w_{mod}) indicates the *head* (or *modifier*) of the compound w_1w_2, $||\cdot||$ is the Euclidean norm, and $\beta \in [0, 1]$ is a parameter that controls the relative importance of the head to the compound's compositionally constructed vector. The normalization of both vectors allows taking only their directions into account, regardless of their norms, which are usually proportional to their frequency and irrelevant to meaning.

We define six compositionality scores based on pc_β. Three of them $\text{pc}_{head}(w_1w_2)$, $\text{pc}_{mod}(w_1w_2)$, and $\text{pc}_{uniform}(w_1w_2)$, correspond to different assumptions about how we model compositionality: if dependent on the head ($\beta = 1$, for e.g., *crocodile tears*), on the modifier ($\beta = 0$, for e.g., *busy bee*), or in equal measure on the head and modifier ($\beta = 1/2$, for e.g., *graduate student*). The fourth score is based on the assumption that compositionality may be distributed differently between head and modifier for different compounds. We implement this idea by setting individually for each compound the

[15] Implemented as `feat_compositionality.py` in the mwetoolkit: `http://mwetoolkit.sf.net`.

[16] Except when explicitly indicated, the term *vector* refers to *corpus-derived vectors* output by DSMs.

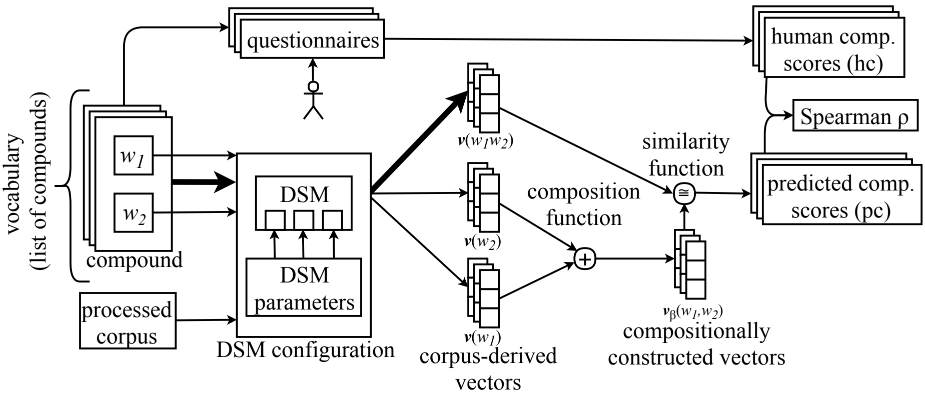


Figure 3
Schema of a compositionality prediction configuration based on a composition function. Thick arrows indicate corpus-based vectors of two-word compounds treated as a single token. The schema also covers the evaluation of the compositionality prediction configuration (top right).

value for β that yields maximal similarity in the predicted compositionality score, that is:[17]

$$\mathrm{pc}_{maxsim}(w_1w_2) = \max_{0\leq\beta\leq 1} \mathrm{pc}_\beta(w_1w_2)$$

Two other scores are not based on the additive model and do not require a composition function. Instead, they are based on the intuitive notion that compositionality is related to the average similarity between the compound and its components:

$$\mathrm{pc}_{avg}(w_1w_2) = \mathrm{avg}(pc_{head}(w_1w_2), pc_{mod}(w_1w_2))$$

We test two possibilities: the arithmetic mean $\mathrm{pc}_{arith}(w_1w_2)$ considers that compositionality is linearly related to the similarity of each component of the compound, whereas the geometric mean $\mathrm{pc}_{geom}(w_1w_2)$ reflects the tendency found in human annotations to assign compound scores $\mathrm{hc}_{\mathbf{HM}}$ closer to the lowest score between that for the head $\mathrm{hc}_{\mathbf{H}}$ and for the modifier $\mathrm{hc}_{\mathbf{M}}$ (Section 3.2).

5. Experimental Setup

This section describes the common setup used for evaluating compositionality prediction, such as corpora (Section 5.1), DSMs (Section 5.2), and evaluation metrics (Section 5.3).

[17] In practice, for the special case of two words, we do not need to perform parameter search for β, which has a closed form obtained by solving the equation $\frac{\partial}{\partial\beta}\mathrm{pc}_\beta(w_1w_2) = 0$:
$\beta = \frac{\cos(w_1w_2,w_1) - \cos(w_1w_2,w_2) \times \cos(w_1,w_2)}{(\cos(w_1w_2,w_1) + \cos(w_1w_2,w_2)) \times (1-\cos(w_1,w_2))}$.

5.1 Corpora

In this work we used the lemmatized and POS-tagged versions of the WaC corpora not only for building DSMs, but also as sources of information about the target compounds for the analyses performed (e.g., in Sections 3.2.2, 9.1, and 9.2):

- for English, the ukWaC (Baroni et al. 2009), with 2.25 billion tokens, parsed with MaltParser (Nivre, Hall, and Nilsson 2006);
- for French, the frWaC with 1.61 billion tokens preprocessed with TreeTagger (Schmid 1995); and
- for Brazilian Portuguese, a combination of brWaC (Boos, Prestes, and Villavicencio 2014), Corpus Brasileiro,[18] and all Wikipedia entries,[19] with a total of 1.91 billion tokens, all parsed with PALAVRAS (Bick 2000).

For all compounds contained in our data sets, we transformed their occurrences into single tokens by joining their component words with an underscore (e.g., `EN` *monkey business* → *monkey_business* and `FR` *belle-mère* → *belle_mère*).[20,21] To handle POS-tagging and lemmatization irregularities, we retagged the compounds' components using the gold POS and lemma in our data sets (e.g., for `EN` *sitting duck, sit/verb_duck/noun→sitting/adjective_duck/noun*). We also simplified all POS tags using coarse-grained labels (e.g., *verb* instead of *vvz*). 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[22] is applied to normalize words across English spelling variants (e.g., *color* → *colour*).

To evaluate the influence of preprocessing on compositionality prediction (Section 7.3), we generated four versions of each corpus, with different levels of linguistic information. We expect lemmatization to reduce data sparseness by merging morphologically inflected variants of the same lemma:

1. *surface*$^{+}$: the original raw corpus with no preprocessing, containing surface forms.
2. *surface*: stopword removal, generating a corpus of surface forms of content words.
3. *lemma*$_{PoS}$: stopword removal, lemmatization,[23] and POS-tagging; generating a corpus of content words distinguished by POS tags, represented as lemma/POS-tag.
4. *lemma*: stopword removal and lemmatization; generating a corpus containing only lemmas of content words.

[18] `http://corpusbrasileiro.pucsp.br/cb/Inicial.html`
[19] Wikipedia articles downloaded on June 2016.
[20] Hyphenated compounds are also re-tokenized with an underscore separator.
[21] Therefore, in Section 5.2, the terms target/context *words* may actually refer to *compounds.*
[22] `https://hunspell.github.io`
[23] In the lemmatized corpora, the lemmas of proper names are replaced by placeholders.

5.2 DSMs

In this section, we describe the state-of-the-art DSMs used for compositionality prediction.

Positive Pointwise Mutual Information (PPMI). In the models based on the PPMI matrix, the representation of a target word is a vector containing the PPMI association scores between the target and its contexts (Bullinaria and Levy 2012). 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, and Dagan 2015).[24] We consider three models that differ in how the contexts are selected:

- In *PPMI–thresh*, the vectors are $|V|$-dimensional but only the top d contexts with highest PPMI scores for each target word are kept, while the others are set to zero (Padró et al. 2014a).[25]
- In *PPMI–TopK*, the vectors are d-dimensional, and each of the d dimensions corresponds to a context word taken from a fixed list of k contexts, identical for all target words. We chose k as the $1,000$ most frequent words in the corpus after removing the top 50 most frequent words (Salehi, Cook, and Baldwin 2015).
- In *PPMI–SVD*, singular value decomposition is used to factorize the PPMI matrix and reduce its dimensionality from $|V|$ to d.[26] We set the value of the context distribution smoothing factor to 0.75, and the negative sampling factor to 5 (Levy, Goldberg, and Dagan 2015). We use the default minimum word count threshold of 5.

Word2vec (w2v). Word2vec[27] relies on a neural network to predict target/context pairs (Mikolov et al. 2013). We use its two variants: continuous bag-of-words (*w2v–cbow*) and skip-gram (*w2v–sg*). We adopt the default configurations recommended in the documentation, except for: no hierarchical softmax, 25 negative samples, frequent-word down-sampling rate of 10^{-6}, execution of 15 training iterations, and minimum word count threshold of 5.

Global Vectors (glove). GloVe[28] implements a factorization of the logarithm of the positional co-occurrence count matrix (Pennington, Socher, and Manning 2014). We adopt the default configurations from the documentation, except for: internal cutoff parameter $x_{max} = 75$ and processing of the corpus in 15 iterations. For the corpora versions *lemma* and *lemma*$_{PoS}$ (Section 5.1), we use the minimum word count threshold of 5. For *surface* and *surface*$^{+}$, due to the larger vocabulary sizes, we use thresholds of 15 and 20.[29]

[24] In previous work adjectives and adverbs were also included as contexts, but the results obtained with only verbs and nouns were better (Padró et al. 2014a).

[25] Vectors still have $|V|$ dimensions but we use d as a shortcut to represent the fact that we only retain the most relevant target-context pairs for each target word.

[26] `https://bitbucket.org/omerlevy/hyperwords`

[27] `https://code.google.com/archive/p/word2vec/`

[28] `https://nlp.stanford.edu/projects/glove/`

[29] Thresholds were selected so as to not use more than 128 GB of RAM.

Table 3
Summary of DSMs, their parameters, and evaluated parameter values. The combination of these DSMs and their parameter values leads to 228 DSM configurations evaluated per language ($1 \times 1 \times 4 \times 3 = 12$ for *PPMI–TopK*, plus $6 \times 3 \times 4 \times 3 = 216$ for the other models).

DSM	Dimension	WordForm	WindowSize
PPMI–TopK	$d = 1000$	*surface*$^+$, *surface*, *lemma*, *lemma*$_{PoS}$	$W = 1{+}1$, $W = 4{+}4$, $W = 8{+}8$
PPMI–thresh	$d = 250$, $d = 500$, $d = 750$		
PPMI–SVD			
w2v–cbow			
w2v–sg			
glove			
lexvec			

Lexical Vectors (lexvec). The LexVec model[30] factorizes the PPMI matrix in a way that penalizes errors on frequent words (Salle, Villavicencio, and Idiart 2016). We adopt the default configurations in the documentation, except for: 25 negative samples, subsampling rate of 10^{-6}, and processing of the corpus in 15 iterations. Due to the vocabulary sizes, we use a word count threshold of 10 for *lemma* and *lemma*$_{PoS}$, and 100 for *surface* and *surface*$^+$.[31]

5.2.1 DSM Parameters. In addition to model-specific parameters, the DSMs described above have some shared **DSM parameters**. We construct multiple **DSM configurations** by varying the values of these parameters. These combinations produce a total of 228 DSMs per language (see Table 3). In particular, we evaluate the influence of the following parameters on compositionality prediction:

- WindowSize: Number of context words to the left/right of the target word when searching for target-context co-occurrence pairs. The assumption is that larger windows are better for capturing semantic relations (Jurafsky and Martin 2009) and may be more suitable for compositionality prediction. We use window sizes of 1+1, 4+4, and 8+8.[32]
- Dimension: Number of dimensions of each vector. The underlying hypothesis is that, the higher the number of dimensions, the more accurate the representation of the context is going to be. We evaluate our framework with vectors of 250, 500, and 750 dimensions.
- WordForm: One of the four word-form and stopword removal variants used to represent a corpus, in Section 5.1: *surface*$^+$, *surface*, *lemma*, and *lemma*$_{PoS}$. They represent different levels of specificity in the informational content of the tokens, and may have a language-dependent impact on the performance of compositionality prediction.

[30] https://github.com/alexandres/lexvec
[31] This is in line with the authors' threshold suggestions (Salle, Villavicencio, and Idiart 2016).
[32] Common window sizes are between 1+1 and 10+10, but a few works adopt larger sizes like 16+16 or 20+20 (Kiela and Clark 2014; Lapesa and Evert 2014).

5.3 Evaluation Metrics

To evaluate a compositionality prediction configuration, we calculate Spearman's ρ rank correlation between the predicted compositionality scores (pc)s and the human compositionality scores (hc)s for the compounds that appear in the evaluation data set. We mostly use the rank correlation instead of linear correlation (Pearson) because we are interested in the framework's ability to order compounds from least to most compositional, regardless of the actual predicted values.

For English, besides the evaluation data sets presented in Section 3, we also use *Reddy* and *Farahmand* (see Section 2.4) to enable comparison with related work. For *Farahmand*, since it contains binary judgments[33] instead of graded compositionality scores, results are reported using the best F_1 (BF_1) score, which is the highest F_1 score found using the top n compounds classified as noncompositional, when n is varied (Yazdani, Farahmand, and Henderson 2015). For *Reddy*, we sometimes present Pearson scores to enable comparison with related work.

Because of the large number of compositionality prediction configurations evaluated, we only report the best performance for each configuration over all possible DSM parameter values. The generalization of these analyses is then ensured using cross-validation and held-out data. To determine whether the difference between two prediction results are statistically different, we use nonparametric Wilcoxon's sign-rank test.

6. Overall Results

In this section, we present the overall results obtained on the *Reddy*, *Farahmand*, *EN-comp*, *FR-comp*, and *PT-comp* data sets, comparing all possible configurations (Section 6.1). To determine their robustness we also report evaluation for all languages using cross-validation (Section 6.2) and for English using the held-out data set *EN-comp*$_{Ext}$ (Section 6.3). All results reported in this section use the $\text{pc}_{uniform}$ function.

6.1 Distributional Semantic Models

Table 4 shows the highest overall values obtained for each DSM (columns) on each data set (rows). For English (*Reddy*, *EN-comp*, and *Farahmand*), the highest results for the compounds found in the corpus were obtained with *w2v* and *PPMI–thresh*, shown as the first value in each pair in Table 4. Not all compounds in the English data sets are present in our corpus. Therefore, we also report results adopting a fallback strategy (the second value). Because its impact depends on the data set, and the relative performance of the models is similar with or without it, for the remainder of the article we discuss only the results without fallback.[34]

The best *w2v–cbow* and *w2v–sg* configurations are not significantly different from each other, but both are different from *PPMI–thresh* ($p < 0.05$). In a direct comparison

[33] A compound is considered as noncompositional if at least 2 out of 4 annotators annotate it as noncompositional.

[34] This refers to 5 out of 180 in *EN-comp* and 129 out of 1,042 in *Farahmand*. For these, the fallback strategy assigns the average compositionality score (Salehi, Cook, and Baldwin 2015). Although fallback produces slightly better results for *EN-comp*, it does the opposite for *Farahmand*, which contains a larger proportion of missing compounds (2.8% vs. 12.4%).

Table 4
Highest results for each DSM, using BF_1 for *Farahmand* data set, Pearson r for *Reddy* (r), and Spearman ρ for all the other data sets. For English, in each pair of values, the first is for the compounds found in the corpus, and the second uses fallback for missing compounds.

Data set	*PPMI–SVD*	*PPMI–TopK*	*PPMI–thresh*	*glove*	*lexvec*	*w2v–cbow*	*w2v–sg*
Farahmand	.487/.424	.435/.376	.472/.404	.400/.358	.449/.431	**.512/.471**	.507/.468
Reddy (r)	.738/.726	.732/.717	.762/.768	.783/.787	.787/.787	.803/.798	**.814/.814**
Reddy (ρ)	.743/.743	.706/.716	.791/.803	.754/.759	.774/.773	.796/.796	**.812/.812**
EN-comp	.655/.666	.624/.632	.688/.704	.638/.651	.646/.658	.716/.730	**.726/.741**
FR-comp	.584	.550	**.702**	.680	.677	.652	.653
PT-comp	.530	.519	**.602**	.555	.570	.588	.586

with related work, our best result for the *Reddy* data set (Spearman $\rho = .812$, Pearson $r = .814$) improves upon the best correlation reported by Reddy, McCarthy, and Manandhar (2011) ($\rho = .714$), and by Salehi, Cook, and Baldwin (2015) ($r = .796$). For *Farahmand*, these results are comparable to those reported by Yazdani, Farahmand, and Henderson (2015) ($BF_1 = .487$), but our work adopts an unsupervised approach for compositionality prediction. For both *FR-comp* and *PT-comp*, the *w2v* models are outperformed by *PPMI–thresh*, whose predictions are significantly different from the predictions of other models ($p < 0.05$).

In short, these results suggest language-dependent trends for DSMs, by which *w2v* models perform better for the English data sets, and *PPMI–thresh* for French and Portuguese. While this may be due to the level of morphological inflection in these languages, it may also be due to differences in corpus size or to particular DSM parameters used in each case. In Section 7, we analyze the impact of individual DSM and corpus parameters to better understand this language dependency.

6.2 Cross-Validation

Table 4 reports the best configurations for the *EN-comp*, *FR-comp*, and *PT-comp* data sets. However, to determine whether the Spearman scores obtained are robust and generalizable, in this section we report evaluation using cross-validation. For each data set, we partition the 180 compounds into 5 folds of 36 compounds ($\mathrm{f}_1, \mathrm{f}_2, \ldots, \mathrm{f}_5$). Then, for each fold f_i, we exhaustively look for the best configuration (values of WindowSize, Dimension, and WordForm) for the union of the other folds ($\cup_{j \neq i} \mathrm{f}_j$), and predict the 36 compositionality scores for f_i using this configuration. The predicted scores for the 5 folds are then grouped into a single set of predictions, which is evaluated against the 180 human judgments.

The partition of compounds into folds is performed automatically, based on random shuffling.[35] To avoid relying on a single arbitrary fold partition, we run cross-validation 10 times, with different fold partitions each time. This process generates 10 Spearman correlations, for which we calculate the average value and a 95% confidence interval.

[35] We have also considered separating folds so as to be balanced regarding their compositionality scores. The results were similar to the ones reported here.

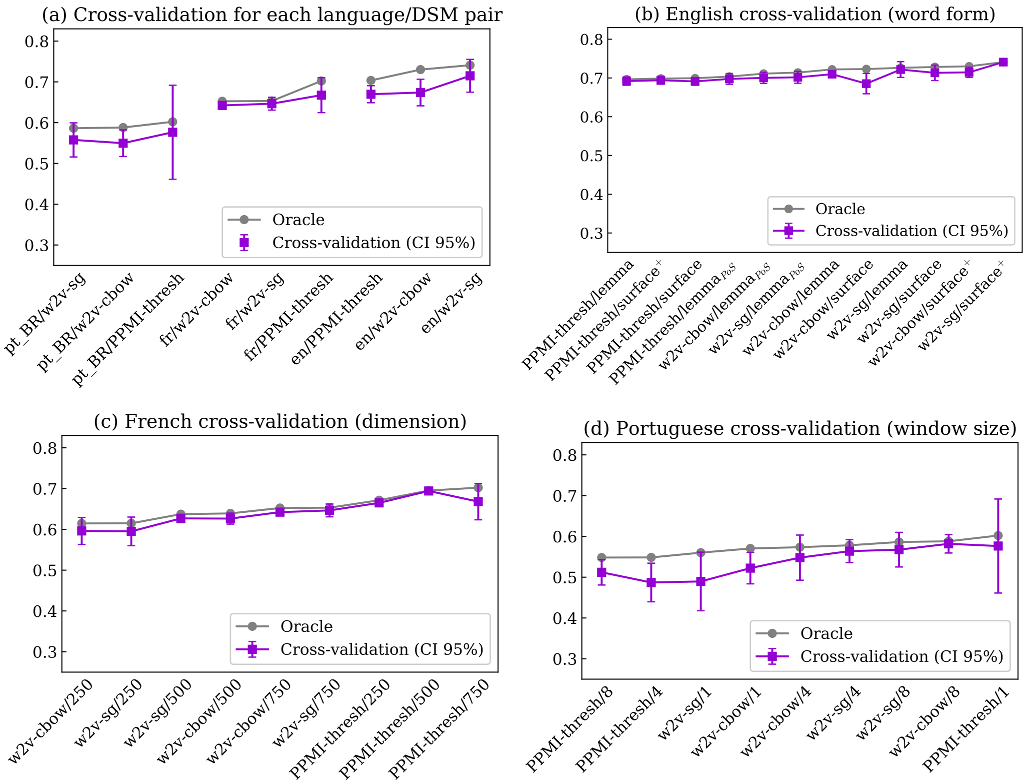


Figure 4
Results with highest Spearman for oracle and cross-validation, the latter with a confidence interval of 95%; (a) top left: overall Spearman correlations per DSM and language, (b) top right: different WORDFORM values and DSMs for English, (c) bottom left: different DIMENSION values and DSMs for French, and (d) bottom right: different WINDOWSIZE values and DSMs for Portuguese.

We have calculated cross-validation scores for a wide range of configurations, focusing on the following DSMs: *PPMI–thresh*, *w2v–cbow*, and *w2v–sg*. Figure 4 presents the average Spearman correlations of cross-validation experiments compared with the best results reported in the previous section, referred to as *oracle*. In the top left panel the x-axis indicates the DSMs for each language using the best oracle configuration, Figure 4(a). In the other panels, it indicates the best oracle configuration for a specific DSM and a fixed parameter for a given language. We present only a sample of the results for fixed parameters, as they are stable across languages. Results are presented in ascending order of oracle Spearman correlation. For each oracle datapoint, the associated average Spearman from cross-validation is presented along with the 95% confidence interval.

The Spearman correlations obtained through cross-validation are comparable to the ones obtained by the oracle. Moreover, the results are quite stable: increasingly better configurations of oracle tend to be correlated with increasingly better cross-validation scores. Indeed, the Pearson r correlation between the 9 oracle points and the 9 cross-validation points in the top-left panel is 0.969, attesting to the correlation between cross-validation and oracle scores.

For *PT-comp*, the confidence intervals are quite wide, meaning that prediction quality is sensitive to the choice of compounds used to estimate the best configurations. Probably a larger data set would be required to stabilize cross-validation results. Nonetheless, the other two data sets seem representative enough, so that the small confidence intervals show that, even if we fix the value of a given parameter (e.g., $d = 750$), the results using cross-validation are stable and very similar to the oracle.

The confidence intervals overlapping with oracle data points also indicate that most cross-validation results are not statistically different from the oracle. This suggests that the highest-Spearman oracle configurations could be trusted as reasonable approximations of the best configurations for other data sets collected for the same language constructed using similar guidelines.

6.3 Evaluation on Held-Out Data

As an additional test of the robustness of the results obtained, we calculated the performance of the best models obtained for one of the data sets (*EN-comp*), on a separate held-out data set ($EN\text{-}comp_{Ext}$). The latter contains 100 compounds balanced for compositionality, not included in *EN-comp* (that is, not used in any of the preceding experiments). The results obtained on $EN\text{-}comp_{Ext}$ are shown in Table 5. They are comparable and mostly better than those for the oracle and for cross-validation. As the items are different in the two data sets, a direct comparison of the results is not possible, but the equivalent performances confirm the robustness of the models and configurations for compositionality prediction. Moreover, these results are obtained in an unsupervised manner, as the compositionality scores are not used to train any of the models. The scores are used only for comparative purposes for determining the impact of various factors in the ability of these DSMs to predict compositionality.

7. Influence of DSM Parameters

In this section, we analyze the influence of DSM parameters on compositionality prediction. We consider different window sizes (Section 7.1), numbers of vector dimensions (Section 7.2), types of corpus preprocessing (Section 7.3), and corpus sizes. For each parameter, we analyze all possible values of other parameters. In other words, we report the best results obtained by fixing a value and considering all possible configurations of other parameters. Results reported in this section use the $\text{pc}_{uniform}$ function.

Table 5
Configurations with best performances on *EN-comp* and on $EN\text{-}comp_{Ext}$. Best performances are measured on *EN-comp* and the corresponding configurations are applied to $EN\text{-}comp_{Ext}$.

DSM	WordForm	WindowSize	Dimension	ρ *EN-comp*	ρ $EN\text{-}comp_{Ext}$
PPMI–SVD	*surface*	1+1	250	0.655	0.692
PPMI–TopK	$lemma_{PoS}$	8+8	1,000	0.624	0.680
PPMI–thresh	$lemma_{PoS}$	8+8	750	0.688	0.675
glove	$lemma_{PoS}$	8+8	500	0.637	0.670
lexvec	$lemma_{PoS}$	8+8	250	0.646	0.685
w2v–cbow	$surface^{+}$	1+1	750	0.716	0.731
w2v–sg	$surface^{+}$	1+1	750	0.726	0.733

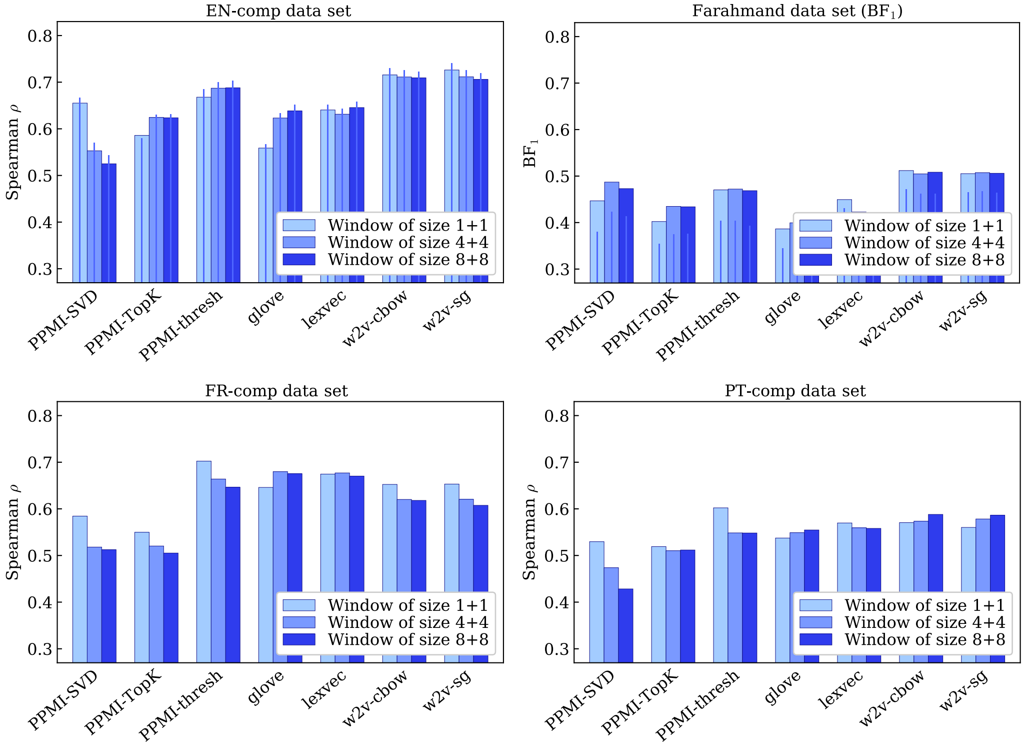


Figure 5
Best results for each DSM and WindowSize (1+1, 4+4, and 8+8), using BF_1 for *Farahmand*, and Spearman ρ for other data sets. Thin bars indicate the use of fallback in English. Differences between the two highest Spearman correlations for each model are statistically significant ($p < 0.05$), except for *PPMI–SVD*, according to Wilcoxon's sign-rank test.

7.1 Window Size

DSMs build the representation of every word based on the frequency of other words that appear in its context. Our hypothesis is that larger window sizes result in higher scores, as the additional data allows a better representation of word-level semantics. However, as some of these models adopt different weight decays for larger windows,[36] variation in their behavior related to window size is to be expected.

Contrary to our expectations, for the best models in each language, large windows did not lead to better compositionality prediction. Figure 5 shows the best results obtained for each window size.[37] For English, *w2v* is the best model, and its performance does not seem to depend much on the size of the window, but with a small trend for smaller sizes to be better. For French and Portuguese, *PPMI–thresh* is only the best model for the minimal window size, and there is a large gap in performance for *PPMI–thresh* as window size increases, such that for larger windows it is outperformed by other models.

[36] For *PPMI–SVD* with WindowSize=8+8, a context word at distance D from its target word is weighted $\frac{8-D}{8}$. For *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 potential noise introduced by distant contexts.

[37] Henceforth, we omit results for *EN-comp*$_{90}$ and *Reddy*, as they are included in *EN-comp*.

To assess which of these differences are statistically significant, we have performed Wilcoxon's sign-rank test on the two highest Spearman values for each DSM in each language. All differences are statistically significant ($p < 0.05$), with the exception of *PPMI–SVD*.

The appropriate choice of window size has been shown to be task-specific (Lapesa and Evert 2017), and the results above suggest that, for compositionality prediction, it depends also on the DSM used. Overall, the trend is for smaller windows to lead to better compositionality prediction.

7.2 Dimension

When creating corpus-derived vectors with a DSM, the question is whether additional dimensions can be informative in compositionality prediction. Our hypothesis is that the larger the number of dimensions, the more precise the representations, and the more accurate the compositionality prediction.

The results shown in Figure 6 for each of the comparable data sets confirm this trend in the case of the best DSMs: *w2v* and *PPMI–thresh*. Moreover, the effect of changing the vector dimensions for the best models seems to be consistent across these languages. The results for *PPMI–SVD*, *lexvec*, and *glove* are more varied, but they are never among

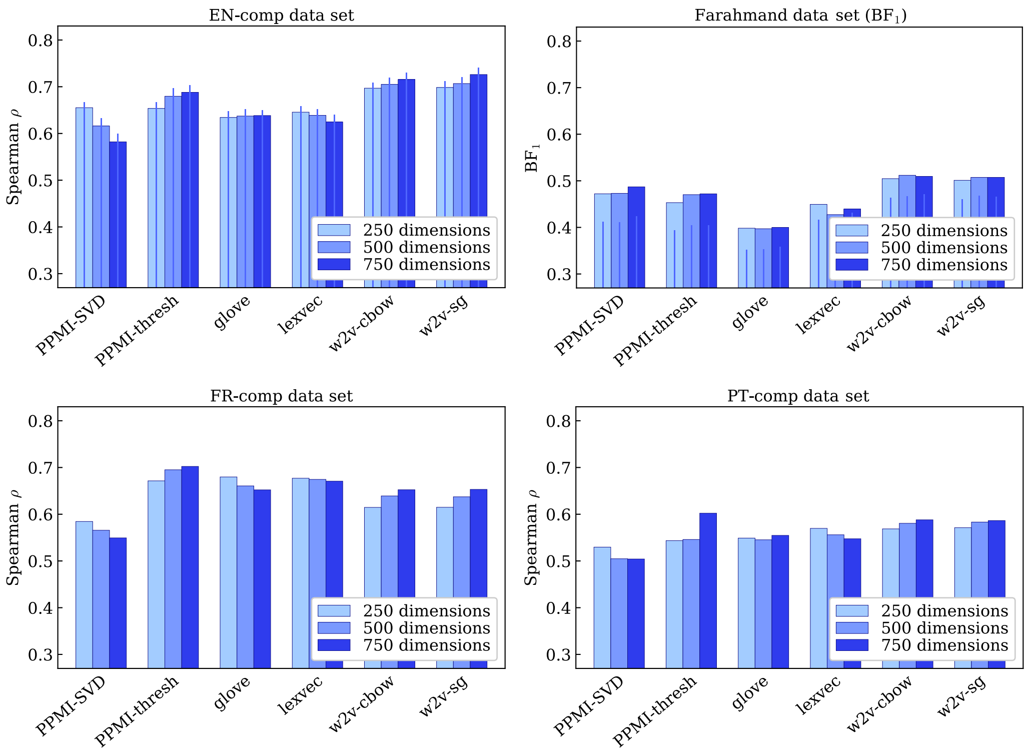


Figure 6
Best results for each DSM and DIMENSION, using BF$_1$ for *Farahmand* data set, and Spearman ρ for all the other data sets. For English, the thin bars indicate results using fallback. Differences between two highest Spearman correlations for each model are statistically significant ($p < 0.05$), except for *PPMI–SVD* for *FR-comp*, according to Wilcoxon's sign-rank test.

the best models for compositionality prediction in any of the languages.[38] All differences between the two highest Spearman correlations are statistically significant ($p < 0.05$), with the exception of *PPMI–SVD* for *FR-comp*, according to Wilcoxon's sign-rank test.

7.3 Type of Preprocessing

In related work, DSMs are constructed from corpora with various levels of preprocessing (Bullinaria and Levy 2012; Mikolov et al. 2013; Pennington, Socher, and Manning 2014; Kiela and Clark 2014; Levy, Goldberg, and Dagan 2015; Salle, Villavicencio, and Idiart 2016). In this work, we compare four levels: WORDFORM= *surface*$^+$, *surface*, *lemma*$_{PoS}$ and *lemma*, described in Section 5.1, corresponding to decreasing amounts of information. Testing different varieties of corpus preprocessing allows us to explore the trade-off between informational content and the statistical significance related to data sparsity for compositionality prediction.

Figure 7 presents the impact of different types of corpus preprocessing on the quality of compositionality prediction. In *EN-comp*, all differences between the two highest Spearman values for each DSM were significant, according to Wilcoxon's sign-rank test, except for *PPMI–thresh*, whereas in *FR-comp* and *PT-comp* they were significant only for *PPMI–TopK* and *lexvec*. However, note that the top two results are often both obtained on representations based on lemmas. If we compare the highest lemma-based result with the highest surface-based result for the same DSM, we find a statistically significant difference in every single case ($p < 0.05$).

When considering the results themselves, although the results for English are heterogeneous, for French and Portuguese, the lemma-based representations consistently allow a better prediction of compositionality scores. This may be explained by the fact that these two languages are morphologically richer than English, and lemma-based representations reduce the sparsity in the data, allowing more information to be gathered from the same amount of data. Moreover, adding POS information (*lemma*$_{PoS}$ vs. *lemma*) does not seem to bring consistent improvements that are statistically significant. This suggests that words that share the same lemma are semantically close enough that any gains from disambiguation are masked by the sparsity of a higher vocabulary size. Finally, the impact of stopword removal is also inconclusive (*surface* vs. *surface*$^+$), considering the best models for each language.

7.4 Corpus Size

If we assume that the bigger the corpus, the better the DSM, this could explain why the results for English are better than those for French and Portuguese, although it does not explain why Portuguese is behind French.[39] In this section, we examine the impact of corpus size on prediction quality by incrementally increasing the amount of data used to generate the DSMs while monitoring the Spearman correlation (ρ) with the human annotations. We use only the best DSMs for these languages, *PPMI–thresh* and *w2v–sg*, with the configurations that produced highest Spearman scores for each full corpus.

As expected, the results in Figure 8 show a smooth, roughly monotonic increase of the ρ values with corpus size, for *PPMI–thresh* and *w2v–sg* for each language and

[38] For *PPMI–SVD* and *lexvec*, this behavior might be related to the fact that both methods perform a factorization of the PPMI matrix.

[39] As the characteristics of *Farahmand* are different from the other data sets, in this analysis we only use the other more comparable data sets.

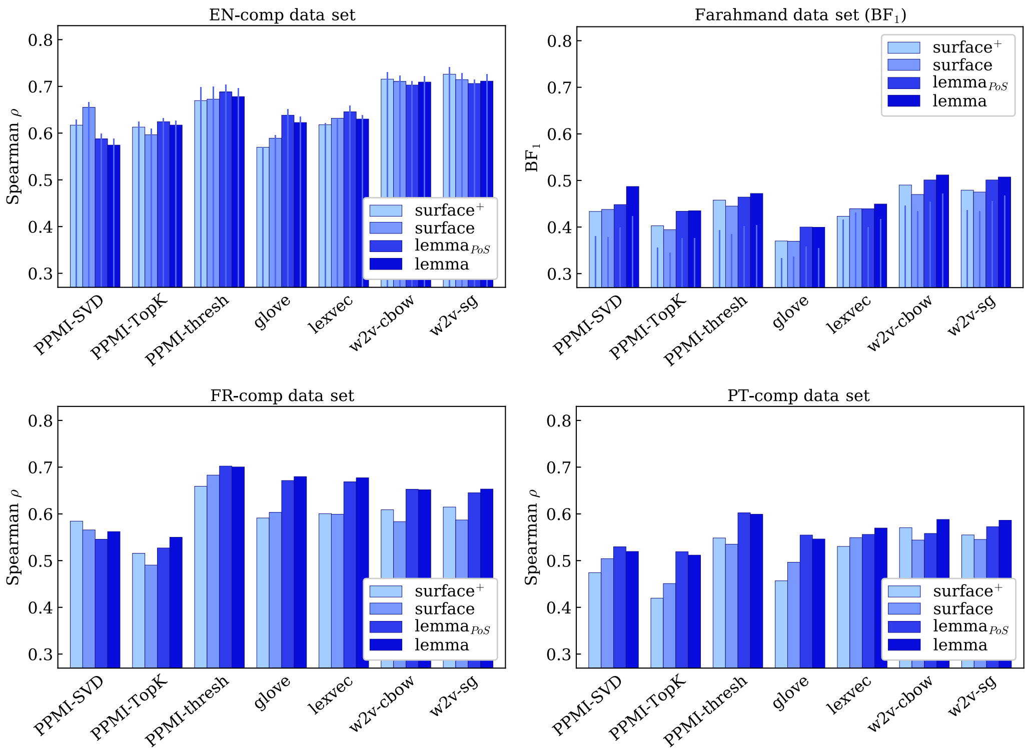


Figure 7
Best results for each DSM and WordForm, using BF$_1$ for *Farahmand* data set, and Spearman ρ for all the other data sets. For English, the thin bars indicate results using fallback. In *EN-comp* all differences between the two highest Spearman values for each DSM were significant, according to Wilcoxon's sign-rank test, except for *PPMI–thresh*, while in *FR-comp* and *PT-comp* they were only significant for *PPMI–TopK* and *lexvec*.

data set.[40] In all cases there is a clear saturation behavior, so that we can safely say that after one billion tokens, the quality of the predictions reaches a plateau and additional corpus fragments do not bring improvements. This suggests that differences in compositionality prediction performance for these languages cannot be totally explained by differences in corpus sizes.

8. Influence of Compositionality Prediction Function

Up to this point, the predicted compositionality scores for the compounds were calculated using a uniform function that assumes that each component contributes 50% to

[40] For *PPMI–thresh*, eight different samplings of corpus fragments were performed (for a total of 800 DSMs per language), with each y-axis data point presenting the average and standard deviation of the ρ obtained from those samplings. For *w2v–sg*, since it is much more time-consuming, a single sampling was used, and thus only one execution was performed for each datapoint (for a total of 100 DSMs per language).

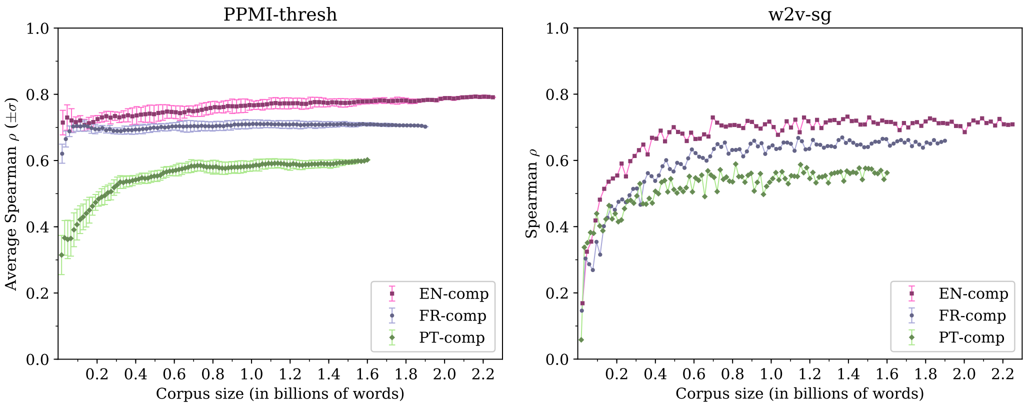


Figure 8
Spearman's ρ for increasing corpus sizes for *PPMI–thresh* (left) and *w2v–sg* (right) for *EN-comp* in red, *FR-comp* in blue, and *PT-comp* in green. Corpus sizes are in the x-axis in billion words. Curves for *PPMI–thresh* show average and standard deviation (error bars) across 8 samplings of the corpus.

the meaning of the compound ($\text{pc}_{uniform}$). However, this might not accurately capture a faithful representation of compounds whose meaning is more semantically related to one of the components (e.g., *crocodile tears*, which is semantically closer to the head *tears*; and *night owl*, which is semantically closer to the modifier *night*). As this may have an impact on the success of compositionality prediction, in this section we evaluate how different compositionality prediction functions model these compounds. In particular, we proposed pc_{maxsim}, (Section 4) for dynamically determining weights that assign maximal similarity between the compound and each of its components. We have also proposed pc_{geom}, which favors idiomatic readings through the geometric mean of the similarities between a compound and its components. Our hypotheses are that pc_{maxsim} will be better correlated with human scores for compositional and partly compositional compounds, while pc_{geom} can better capture the semantics of idiomatic ones (Section 8.1).

First, to verify whether other prediction functions improve results obtained for the best $\text{pc}_{uniform}$ configurations reported up to now, we have evaluated every strategy on all DSM configurations. Table 6 shows that the functions that combine both components (columns $\text{pc}_{uniform}$ to pc_{arith}) generate better compositionality predictions than functions that ignore one of the individual components (columns pc_{head} and pc_{mod}). There is some variation among the combined scores, with the best score indicated in bold. Every best score is statistically different from all other scores in its row ($p < 0.05$). The results for pc_{arith} and $\text{pc}_{uniform}$ are very similar, reflecting their similar formulations.[41]

Here we focus on the issue of adjusting β in the compositionally constructed vector; that is, we consider the use of pc_{maxsim} instead of $\text{pc}_{uniform}$. This score seems to be beneficial in the case of English (*EN-comp*), but not in the case of French or Portuguese.

[41] The Pearson correlations (averaged across 7 DSMs) between pc_{arith} and $\text{pc}_{uniform}$ are $r = .972$ for *EN-comp*, $r = .991$ for *FR-comp*, and $r = .969$ for *PT-comp*, confirming their similar results.

Table 6
Spearman ρ for the proposed compositionality prediction scores, using the best DSM configuration for each score.

Data set	$\mathrm{pc}_{uniform}$	pc_{maxsim}	pc_{geom}	pc_{arith}	pc_{head}	pc_{mod}
EN-comp	.726	**.730**	.677	.718	.555	.677
FR-comp	.702	.693	.699	**.703**	.617	.645
PT-comp	**.602**	.590	.580	.598	.558	.486

Table 7 presents the best pc_{maxsim} model for each data set, along with the average weights assigned to head and modifier for every compound in the data set. Before analyzing the results in Table 7, we have to verify whether the data sets are balanced for the influence of each component to the meaning of the whole, or if there is any bias towards heads/modifiers. The influence of the head, estimated as the average of $\mathrm{hc}_{\mathbf{H}}/(\mathrm{hc}_{\mathbf{H}} + \mathrm{hc}_{\mathbf{M}})$ over all compounds of a data set, is 0.50 for *EN-comp*, 0.52 for *FR-comp*, and 0.52 for *PT-comp*. This indicates that the data sets are balanced in terms of the influence of each component, and neither head nor modifier predominates as more compositional or idiomatic than the other.

As for the average β weights in pc_{maxsim}, while the weights that maximize compositionality are fairly similar for *EN-comp*, they strongly favor the head for both *FR-comp* and *PT-comp*. This may be explained by the fact that, for the latter, the modifiers are all adjectives, while *EN-comp* has mostly nouns as modifiers. Surprisingly, this seemingly more realistic weighting of the compound components for French and Portuguese does not reflect in better compositionality scores, and does not correspond to the average influence of modifiers in these data sets, estimated as 0.48 on average. One possible explanation could be that, in these cases, the adjectives may be contributing to some specific more idiomatic meaning that is not found in isolated occurrences of the adjective itself, such as FR *beau* (lit. *beautiful*), which is used in the translation of most *in-law* family members, such as FR *beau-frère* (lit. *beautiful-brother* 'brother-in-law'). In the next section, we investigate which compounds are affected the most by these different scores.

Table 7
DSM and Separman ρ of pc_{maxsim}, as well as the average weights for the head ($\overline{\beta}$) and for the modifier ($\overline{1-\beta}$) on each data set.

Data set	DSM	ρ_{maxsim}	$\overline{\beta}$ (head)	$\overline{1-\beta}$ (mod.)
EN-comp	*w2v–sg*	.730	.55	.45
FR-comp	*PPMI–thresh*	.693	.68	.32
PT-comp	*w2v–sg*	.590	.68	.32

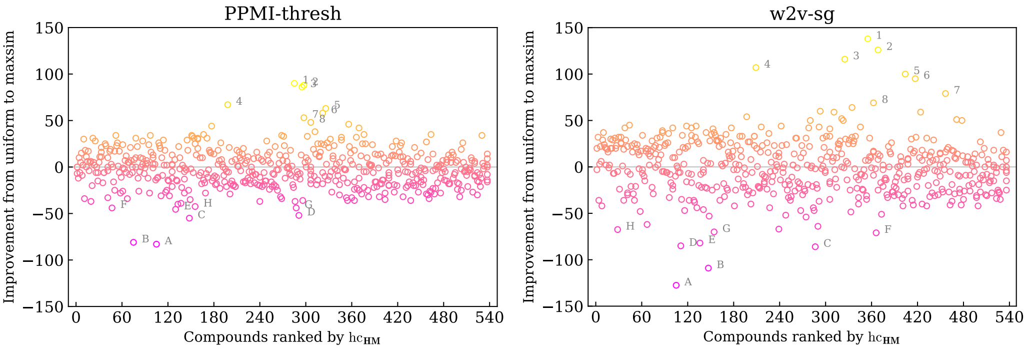


Figure 9
Distribution of improv$_{maxsim}$ (y-axis) as a function of rk$_{human}$ (x-axis). Outliers are indicated by numbers 1–8 (positive improvement) and letters A–H (negative improvement).

8.1 Rank Improvement Analysis

To better evaluate the effect of adjusting β for the individual compounds with respect to the pc$_{uniform}$ score, we define the rank improvement as:

$$\text{improv}_f(w_1w_2) = |\text{rk}_{uniform}(w_1w_2) - \text{rk}_{human}(w_1w_2)| - |\text{rk}_f(w_1w_2) - \text{rk}_{human}(w_1w_2)|,$$

where rk indicates the rank of the compound w_1w_2 in the data set when ordered according to pc$_{uniform}$, human annotations hc$_{\mathbf{HM}}$, or the compositionality prediction function f. For instance, when $f = maxsim$, positive improv$_{maxsim}$ values indicate that pc$_{maxsim}$ yields a better approximation of the ranks assigned by hc$_{\mathbf{HM}}$ than pc$_{uniform}$, whereas negative values indicate that pc$_{uniform}$ provides a better ranking.

We perform a cross-lingual analysis, grouping the hc$_{\mathbf{HM}}$ scores of the *EN-comp*, *FR-comp*, and *PT-comp* into a unique data set (henceforth *ALL-comp*), containing 540 compounds. Figure 9 presents the values of rank improvement for the best *PPMI–thresh* and *w2v–sg* configurations, ranked according to hc$_{\mathbf{HM}}$ (rk$_{human}$): compounds that are better predicted by pc$_{maxsim}$ have positive rank movements (above the 0 line).[42] The density of movement on either side of the 0 (no movement) line appears to be similar for both models with pc$_{maxsim}$ performing as well as pc$_{uniform}$.

Figure 9 also marks the outlier compounds with the highest improvements (numbers from 1 to 8) and those with the lowest improvements (letters from A to H), and Table 8 shows their improvement scores. In the case of these outliers, the adjustment seems to be more beneficial to compositional compounds than to idiomatic cases. This is confirmed by a linear regression of the movement of the 8+8 outliers as a function of the compositionality scores hc$_{\mathbf{HM}}$, where we obtain a positive coefficient of $r = 0.73$ and $r = 0.72$ for *PPMI–thresh* and *w2v–sg*, respectively. There are more outlier compounds for Portuguese and French (particularly the former), suggesting that pc$_{maxsim}$ has a stronger impact on those languages than on English. Moreover, some compounds had a similar improvement under both DSMs, with, for example, high improvement

[42] We focus on one representative of PPMI-based DSMs and one representative of word-embedding models. Similar results were observed for the best configurations of other DSMs.

Table 8
Outlier compounds with extreme positive/negative improv$_{maxsim}$ values. Example identifiers correspond to numbers/letters shown in Figure 9.

improv$_{maxsim}$ for *PPMI–thresh*

ID	improv	hc$_{\text{HM}}$	Compound 'translation' (*gloss*)
1	+90	2.82	FR *premier plan* 'foreground' (lit. *first plan*)
2	+88	2.90	FR *matiére premiére* 'raw material' (lit. *matter primary*)
3	+86	2.89	PT *amigo oculto* 'secret Santa' (lit. *friend hidden*)
4	+67	1.92	FR *premiére dame* 'first lady' (lit. *first lady*)
5	+63	3.19	PT *caixa forte* 'safe, vault' (lit. *box strong*)
6	+58	3.14	PT *prato feito* 'blue-plate special' (lit. *plate ready-made*)
7	+53	2.90	FR *idée reçue* 'popular belief' (lit. *idea received*)
8	+48	3.00	FR *marée noire* 'oil spill' (lit. *tide black*)
H	−42	1.52	PT *alta costura* 'haute couture' (lit. *high sewing*)
G	−44	2.84	EN *half sister*
F	−44	0.54	EN *melting pot*
E	−46	1.29	FR *berger allemand* 'German shepherd' (lit. *shepherd German*)
D	−52	2.87	PT *mar aberto* 'open sea' (lit. *sea open*)
C	−55	1.43	PT *febre amarela* 'yellow fever' (lit. *fever yellow*)
B	−81	0.79	PT *livro aberto* 'open book' (lit. *book open*)
A	−83	1.06	PT *coração partido* 'broken heart' (lit. *heart broken*)

improv$_{maxsim}$ for *w2v–sg*

ID	improv	hc$_{\text{HM}}$	Compound 'translation' (*gloss*)
1	+138	3.58	PT *cerca viva* 'hedge' (lit. *fence living*)
2	+126	3.67	FR *coffre fort* 'safe, vault' (lit. *chest/box strong*)
3	+116	3.19	PT *caixa forte* 'safe, vault' (lit. *chest/box strong*)
4	+107	2.03	PT *golpe baixo* 'low blow' (lit. *punch low*)
5	+100	3.97	PT *primeira necessidade* 'first necessity' (lit. *first necessity*)
6	+95	4.11	EN *role model*
7	+79	4.47	FR *bonne pratique* 'good practice' (lit. *good practice*)
8	+69	3.64	PT *carta aberta* 'open letter' (lit. *letter open*)
H	−68	0.40	FR *bras droit* 'most important helper/assistant' (lit. *arm right*)
G	−70	1.52	PT *alta costura* 'haute couture' (lit. *high sewing*)
F	−71	3.66	PT *carne vermelha* 'red meat' (lit. *meat red*)
E	−82	1.35	PT *alto mar* 'high seas' (lit. *high sea*)
D	−85	1.10	PT *mesa redonda* 'round table' (lit. *table round*)
C	−86	2.84	EN *half sister*
B	−109	1.43	PT *febre amarela* 'yellow fever' (lit. *fever yellow*)
A	−128	1.06	PT *coração partido* 'broken heart' (lit. *heart broken*)

for PT *caixa forte* literally *box strong* 'safe' and low improvement for PT *coração partido* 'broken heart'. In addition, pc$_{maxsim}$ also affected some equivalent compounds in different languages, as in the case of PT *caixa forte* and FR *coffre fort*. Overall, pc$_{maxsim}$ does not present a considerable impact on the predictions, obtaining an average improvement of improv$_{maxsim}$ = +0.41 across all compounds in *ALL-comp*.

Figure 10 shows the same analysis for $f = geom$, showing the improvement score of pc$_{geom}$ over pc$_{uniform}$. We hypothesized that pc$_{geom}$ should more accurately represent

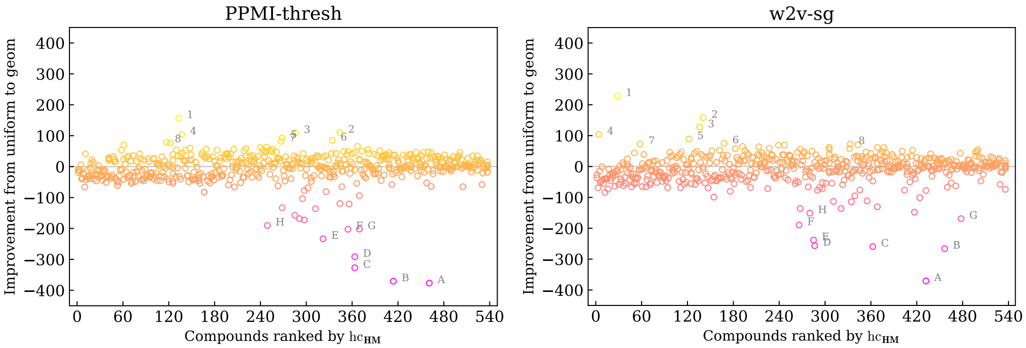


Figure 10
Distribution of improv_{geom} (y-axis) as a function of rk_{human} (x-axis). Outliers are indicated by numbers 1–8 (positive improvement) and letters A–H (negative improvement).

idiomatic compounds. From the previous sections, we know that pc_{geom} has lower performance than $\text{pc}_{uniform}$ when used to estimate the compositionality of the entire data sets (cf. Table 6). This is confirmed by an average score of $\overline{\text{improv}_{geom}} = -7.87$. As in Figure 9, Figure 10 shows a random distribution of improvements. However, the outliers have the opposite pattern, indicating that large reclassifications due to pc_{geom} tend to favor idiomatic instead of compositional compounds. The linear regression of the movement of the outliers as a function of the compositionality scores results in $r = -0.73$ and $r = -0.82$ for *PPMI–thresh* and *w2v–sg*, respectively. These confirm our hypothesis for the behavior of pc_{geom}.

Table 9 lists the outlier compounds indicated in Figure 10 along with their improvement values. Here again, the majority of the outliers belong to *PT-comp*. Some of the compounds that were found as outliers in pc_{maxsim} re-appear as outliers for pc_{geom} with inverted polarity in the improvement score, such as the ranks predicted by *PPMI–thresh* for PT *prato feito* literally *plate made* 'blue-plate special' ($\text{improv}_{maxsim} = +58$, $\text{improv}_{geom} = -234$) and by *w2v–sg* for FR *bras droit* literally *arm right* 'assistant' ($\text{improv}_{maxsim} = -68$, $\text{improv}_{geom} = +228$). This suggests that, as future work, we should consider combining both approaches into a single prediction that decides which score to use for each compound as a function of $\text{pc}_{uniform}$.

9. Characterization of the Predicted Compositionality

In the previous sections, we examined the performance of the compositionality prediction framework in terms of the correlation between automatic predictions and human judgments across languages. We now investigate the relation between predicted scores and other variables that may have an impact on results, such as familiarity (Section 9.1) and conventionalization (Section 9.2). We also compare the predicted compositionality scores with trends previously found in human scores (Section 9.3). The experiments focus on the *ALL-comp* data set, which groups the predicted scores from the best configurations on *EN-comp*, *FR-comp*, and *PT-comp* (cf. Table 4).

Table 9
Outlier compounds with extreme positive/negative improv$_{geom}$ values. Example identifiers correspond to numbers/letters shown on Figure 10.

improv$_{geom}$ for *PPMI–thresh*

ID	improv	hc$_{\text{HM}}$	Compound 'translation' (*gloss*)
1	+157	1.31	EN *snail mail*
2	+110	3.43	FR *guerre civile* 'civil war' (lit. *war civil*)
3	+109	2.83	FR *disque dur* 'hard drive' (lit. *disk hard*)
4	+104	1.35	PT *alto mar* 'high seas' (lit. *high sea*)
5	+93	2.63	PT *ônibus executivo* 'minibus' (lit. *bus executive*)
6	+85	3.32	EN *search engine*
7	+82	2.62	PT *carro forte* 'armored car' (lit. *car strong*)
8	+79	1.18	EN *noble gas*
H	−190	2.44	PT *ar condicionado* 'air conditioning' (lit. *air conditioned*)
G	−202	3.67	FR *coffre fort* 'safe, vault' (lit. *chest/box strong*)
F	−202	3.57	FR *bon sens* 'common sense' (lit. *good sense*)
E	−234	3.14	PT *prato feito* 'blue-plate special' (lit. *plate ready-made*)
D	−292	3.64	FR *baie vitrée* 'open glass window' (lit. *opening glassy*)
C	−327	3.64	PT *carta aberta* 'open letter' (lit. *letter open*)
B	−370	4.08	PT *vinho tinto* 'red wine' (lit. *wine dark-red*)
A	−376	1.69	PT *circuito integrado* 'short circuit' (lit. *short circuit*)

improv$_{geom}$ for *w2v–sg*

ID	improv	hc$_{\text{HM}}$	Compound 'translation' (*gloss*)
1	+228	0.40	FR *bras droit* 'most important helper/assistant' (lit. *arm right*)
2	+158	1.40	PT *lua nova* 'new moon' (lit. *moon new*)
3	+127	1.35	PT *alto mar* 'high seas' (lit. *high sea*)
4	+104	0.10	PT *pé direito* 'ceiling height' (lit. *foot right*)
5	+89	1.24	EN *carpet bombing*
6	+75	1.60	PT *lista negra* 'black list' (lit. *list black*)
7	+73	0.65	PT *arma branca* 'cold weapon' (lit. *weapon white*)
8	+72	3.32	EN *search engine*
H	−151	2.76	PT *disco rígido* 'hard drive' (lit. *disk rigid*)
G	−169	4.63	EN *subway system*
F	−190	2.62	PT *carro forte* 'armored car' (lit. *car strong*)
E	−238	2.83	FR *disque dur* 'hard drive' (lit. *disk hard*)
D	−256	2.84	EN *half sister*
C	−260	3.64	PT *carta aberta* 'open letter' (lit. *letter open*)
B	−266	4.47	FR *bonne pratique* 'good practice' (lit. *good practice*)
A	−370	4.25	EN *end user*

9.1 Predicted Compositionality and Familiarity

Results from Section 3.2.2 show that the familiarity of compounds measured as frequency in large corpora is associated with the compositionality scores assigned by humans. We would like to know whether this correlation also holds true to system predictions: Are the most frequent compounds being predicted as more compositional? As expected, the rank correlation between frequency and pc$_{uniform}$ shows medium to

Table 10
Spearman ρ correlations between different variables. We consider the set of predicted scores (pc), the set of human–prediction differences (diff), the compound frequencies (freq), and the compound PMI. The predicted scores are the ones from the best configurations of each sub–data set in *ALL-comp*. Correlations are indicated only when significant ($p < 0.05$).

DSM	ρ(pc,freq)	ρ(diff,freq)	ρ(pc, PMI)	ρ(diff, PMI)
PPMI–SVD	0.36	0.17	0.28	*
PPMI–thresh	0.46	0.22	0.13	*
glove	0.68	−0.19	0.26	*
lexvec	0.54	*	0.26	−0.12
PPMI–TopK	0.28	0.15	*	*
w2v–cbow	0.51	*	0.17	*
w2v–sg	0.50	*	0.17	*

strong correlation (see Table 10, column ρ[pc,freq]), though the level of correlation is somewhat DSM-dependent, are in line with the correlation observed between frequency and human scores, and with the high correlation between predicted and human scores.

Another hypothesis we test is whether frequent compounds are easier to model. A first intuition would be that this hypothesis is true, as a higher number of occurrences is associated with a larger amount of data, from which more representative vectors can be built. To test this hypothesis, we define a compound's *difficulty* as the difference between the predicted score and the normalized human score, $\text{diff} = |\text{pc} - (\text{hc}_{\textbf{HM}}/5)|$, where high values indicate a compound whose compositionality is harder to predict.[43]

We found a weak (though statistically significant) correlation between frequency and difficulty for some of the DSMs (Table 10, column ρ[diff,freq]). They are mostly positive, indicating that frequency is correlated with difficulty, which is a surprising result, as it implies that the compositionality of rarer compounds was mildly *easier* to predict for these systems, disproving the hypothesis above. These results either point to an overall lack of correlation between frequency and difficulty, or indicate mild DSM-specific behavior, which should be investigated in further research.

9.2 Predicted Compositionality and Conventionalization

PMI is not only a well-known estimator of the level of conventionalization of a multiword expression (Church and Hanks 1990; Evert 2004; Farahmand, Smith, and Nivre 2015), but it is also used in some DSMs as a way to estimate the strength of association between target and context words. To assess if what our models are implicitly measuring is the association between the component words of a compound rather than compositionality, we now examine the correlation between compositionality scores and PMI.

We found only a weak but statistically significant correlation between predicted compositionality and PMI (Table 10, column ρ[pc, PMI]), which suggests that these DSMs preserve some information regarding conventionalization. However, given that no significant correlation between PMI and human compositionality scores was found

[43] We linearly normalize predicted scores to be between 0 and 1. However, given that negative scores are rare in practice, unreported correlation with non-normalized pc are similar to the ones reported.

(Section 3.2.2) and as DSM predictions are strongly correlated to human predictions, these results indicate that our models capture more than conventionalization. They may also be a feature of this particular set of compounds, as even the compositional cases are also conventional to some extent (e.g., *white/?yellow wine*). Therefore, further investigation of possible links between idiomaticity and conventionalization is needed.

We also calculated the correlation between PMI and the human–prediction difference (diff), to determine if DSMs build less precise vectors for less conventionalized compounds (approximated as those with lower PMI). However, no statistically significant correlation was found for most DSMs (Table 10, column $\rho[\text{diff}, \text{PMI}]$).

9.3 Range-Based Analysis of Predicted Compositionality

Spearman correlation assesses the performance of a given configuration by providing a single numerical value. This facilitates the comparison between configurations, but it hides the internal distribution of predictions. By splitting the data sets into ranges, we obtain a more fine-grained view of possible patterns linked to compositionality prediction.

To determine if the compounds that humans agree more on are also more accurately predicted, we divided *ALL-comp* into three equally sized subsets, according to the standard deviation among human annotators (low, mid-range, and high values of standard deviation, $\sigma_{\mathbf{HM}}$). As high standard deviation indicates disagreement among annotators, it may be an indicator of the difficulty of the annotation. Table 11 presents the best DSMs, according to Spearman's ρ evaluated separately on each of the subsets. Indeed, for the compounds that had low $\sigma_{\mathbf{HM}}$, the Spearman values were the highest (between 0.73 and 0.75), while for those with high $\sigma_{\mathbf{HM}}$, the Spearman correlation with human judgments was the lowest (between 0.35 and 0.43). These results confirm that higher scores are achieved for the compounds for which humans agree more, and suggest that part of the difficulty of this task for automatic systems is also related to difficulties for humans.

To determine if compositional compounds would be more precisely predicted than idiomatic compounds, we divide *ALL-comp* into three equally sized subsets based on the level of human compositionality scores (low, mid-range, and high values of $\text{hc}_{\mathbf{HM}}$). Table 11 presents the correlation obtained on each subset for the best configuration of each DSM. The more idiomatic compounds have the lowest Spearman values (from 0.16 to 0.29) while the more compositional have the highest ones (from 0.32 to 0.37). These results confirm that the predictions are better for compositional than for idiomatic compounds. Moreover, these scores are much lower than those from the full data set

Table 11
Spearman's ρ of best $\text{pc}_{uniform}$ models, separated into 3 ranges according to $\sigma_{\mathbf{HM}}$ and according to $\text{hc}_{\mathbf{HM}}$, all with $p < 0.05$.

DSM	full data set	Ranges of $\sigma_{\mathbf{HM}}$			Ranges of $\text{hc}_{\mathbf{HM}}$		
		low	mid	high	low	mid	high
PPMI–thresh	**0.66**	**0.75**	0.58	0.40	0.29	0.24	**0.37**
glove	**0.63**	**0.73**	0.54	0.35	0.27	0.26	**0.35**
lexvec	**0.64**	**0.73**	0.54	0.36	0.18	0.20	**0.37**
w2v–sg	**0.66**	**0.73**	0.58	0.43	0.16	0.24	**0.32**

(from 0.63 to 0.66), suggesting that it may be harder to make fine-grained distinctions (e.g., between two compositional compounds like *access road* and *subway system*) than to make inter-range distinctions (e.g., between idiomatic and compositional compounds like *ivory tower* and *access road*). However, further investigation would be needed to verify this hypothesis.

10. Conclusions

We proposed a framework for compositionality prediction of multiword expressions, focusing on nominal compounds and using DSMs for meaning representation. We investigated how accurately DSMs capture idiomaticity compared to human judgments and examined the impact of several variables in the accuracy of the predictions. In order to determine how language dependent the results are, we evaluated the compositionality prediction framework in English, French, and Portuguese, using data sets containing human-rated compositionality scores, some of which were specifically constructed as part of this work.[44] Using these data sets, we presented a large-scale evaluation involving 228 DSMs for each language, and we evaluated more than 9,000 framework configurations to determine the impact of possible factors that may influence compositionality prediction.

Our experiments confirmed that our framework is able to capture idiomaticity accurately, obtaining a strong correlation with human judgments for all three languages. Comparing the performance of different DSMs, the particular choice of DSM had a noticeable impact on the results, with differences over 0.10 Spearman ρ points for all languages. For the comparable data sets (*EN-comp*, *FR-comp*, and *PT-comp*), the best models were *w2v* and *PPMI–thresh*.[45] Results differed according to language: although for English *w2v* were the best models, for French and Portuguese, *PPMI–thresh* outperformed the other models. Moreover, the results for the three languages varied considerably, with those for English outperforming by 0.10 and 0.20 Spearman ρ points those for French and Portuguese, respectively. The latter are morphologically richer than the former, and a closer examination of the type of preprocessing adopted for best results reveals that both languages benefit from less sparse representations resulting from lemmatization and stopword removal, while for English no preprocessing was particularly beneficial.

Although corpus size is often assumed to play a fundamental role in the quality of DSMs, so that the bigger the corpus the better the results, prediction quality stabilized at around one billion tokens for all languages. This may reflect the point where the minimum frequency was reached for producing reliable representations for all compounds in these data sets, even the rare cases, and larger corpora did not lead to better predictions. Moreover, for the best models in each language, DSMs with more dimensions resulted in more accurate predictions confirming our hypothesis. We also found a trend for small window sizes leading to better results for the best models in all three languages, contrary to our hypothesis. A typically good configuration used vectors of 750 dimensions built from minimal context windows of one word to each side of the target.

DSMs were also robust regarding the choice of compositionality prediction function, with a uniform combination of the head and modifier producing the best results

[44] The resulting data sets and framework implementation are freely available to the community.

[45] As *Farahmand* is considerably different from the other data sets, a direct comparison is not possible.

for all languages. Other functions like pc_{maxsim} and pc_{geom}, which modify these scores to account for different contributions of each component, produced at best similar results.

A deeper analysis of the predicted compositionality scores revealed that, similarly to human-rated scores, familiarity measured as frequency was positively correlated with predicted compositionality. In the case of conventionalization measured as PMI, no correlation was found with human-rated scores and only a mild correlation was found with some predicted scores, suggesting that our models capture more than compound conventionalization, as they have a strong agreement with human scores. Intra-compound standard deviation on human scores was also found to be related to predicted scores, indicating that DSMs have difficulties on those compounds that humans also found difficult. Moreover, predictions were found to be more accurate for compositional compounds.

Although there are many questions that still need to be solved regarding compositionality, we believe that the results presented here advance significantly its understanding and computational modeling. Furthermore, the proposed framework opens important avenues of research that are ready to be pursued. First, the role of morphological inflection could be clarified by extending this investigation to even more inflected languages, such as Turkish. Moreover, other categories of MWEs such as verb+noun expressions should be evaluated to determine the interplay between compositionality prediction and syntactic flexibility of MWEs. The ultimate test would be to use predicted compositionality scores in downstream applications and tasks involving some degree of semantic processing, ranging from MWE identification to parsing, and word-sense disambiguation. In particular, it would be interesting to predict compositionality in context, in order to distinguish idiomatic from literal usages in sentences.

Appendix A. Glossary

composition function is a function that takes as input a sequence of vectors $\mathbf{v}(w_i)$ to $\mathbf{v}(w_j)$ and outputs a compositionally constructed vector $\mathbf{v}_{\oplus}(w_i \ldots w_j)$ representing the compositional meaning of the sequence, where $\oplus$ indicates the function used to compose the vectors. Example: $\mathbf{v}_{\beta}(w_1, w_2) = \beta \frac{\mathbf{v}(w_1)}{||\mathbf{v}(w_1)||} + (1 - \beta) \frac{\mathbf{v}(w_2)}{||\mathbf{v}(w_2)||}$. 1, 23

compositionality prediction configuration is the combination of a particular DSM configuration with a given compositionality prediction function, fully specifying how a predicted compositionality score is calculated for a given word sequence $w_i \ldots w_j$. 1

compositionality prediction framework is the set of all possible compositionality prediction configurations available. 1, 22

compositionality prediction function is a function that takes as input corpus-based vectors for a sequence of words $\mathbf{v}(w_i \ldots w_j)$ and for the individual words composing that sequence $\mathbf{v}(w_i) \ldots \mathbf{v}(w_j)$, and outputs a predicted compositionality score, usually proportional to the similarity between the corpus-based vector $\mathbf{v}(w_i \ldots w_j)$ and a compositionally constructed vector $\mathbf{v}(w_i)$ to $\mathbf{v}(w_j)$ derived from $\mathbf{v}(w_i) \ldots \mathbf{v}(w_j)$ using a composition function. Example: *maxsim*. 1

compositionally constructed vector is the output of a composition function, that is, a vector $\mathbf{v}_{\oplus}(w_i \ldots w_j)$ derived from the individual words' corpus-derived vectors $\mathbf{v}(w_i)$ to $\mathbf{v}(w_j)$. 1, 23

corpus-derived vector is the output of a DSM for a given element w_i of the vocabulary V, that is, a corpus-derived D-dimensional real-numbered vector $\mathbf{v}(w_i)$ that represents the meaning of w_i. A corpus-derived vector of a word sequence $\mathbf{v}(w_i \ldots w_j)$ is built by treating it as a single token in the corpus. 1, 22

distributional semantic model (DSM) is a function that takes as input a vocabulary V and a (large) corpus, and outputs a corpus-derived vector $\mathbf{v}(w_i)$ for each element w_i of V based on the distributional profile of w_i's occurrences in the corpus. The vocabulary V can be automatically derived from the input corpus. Example: *w2v–cbow*. 1

DSM configuration is a set of DSM parameters and their values, fully specifying how corpus-derived vectors are built from a given corpus. Example: *w2v–cbow* using $lemma_{PoS}.W_8.d_{750}$. 1, 28

DSM parameter is a variable in a DSM whose value influences the way corpus-derived vectors are built from a corpus. Example: WORDFORM. 1, 28

human compositionality score (hc) is a real value representing the compositionality assigned by human annotators to a word sequence $w_i \ldots w_j$. The correlation between predicted compositionality (pc) and human compositionality (hc) scores is used to evaluate a compositionality prediction configuration. When subscripted, indicates the question used to obtain the score. Example: $\text{hc}_{\mathbf{H}}$. 1, 17, 29

predicted compositionality score (pc) is the output of a compositionality prediction function, that is, a real value representing the predicted compositionality of a word sequence $w_i \ldots w_j$. The correlation between predicted compositionality (pc) and human compositionality (hc) scores is used to evaluate a compositionality prediction configuration. When subscripted, indicates the compositionality prediction function used to obtain the score. Example: $\text{pc}_{uniform}$. 1, 29

Appendix B. 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 in this article 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.

B.1 Number of Iterations

Some of the DSMs in consideration on this paper 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

Table 12
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)

number of iterations is inspired by the models found in parts of the literature, where, for example, the number of *glove* iterations can be as high as 50 (Salle, Villavicencio, and Idiart 2016) or even 100 (Pennington, Socher, and 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 12 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,[46] we refrained from building further models with as many iterations in the rest of the article.

B.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.

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

Table 13
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)

Table 13 presents the result from the highest-Spearman configurations along with the results for an identical configuration with a higher occurrence threshold of 50.[47] 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.

B.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 7.1). It is possible that a higher score could be obtained by a configuration in between. While a full exhaustive search would be the ideal solution, an initial 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 shown in Table 14 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.

[47] The threshold used for ρ_{base} depends on the DSM, and is described in Section 5.2.

Table 14
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)

B.4 Higher Number of Dimensions

As seen in Section 7.2, 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 15 presents the result of running *PPMI–thresh* for increasing values of of the DIMENSION parameter. The baseline configuration (indicated as * in Table 15) was the highest-scoring configuration found in Section 7.2: $lemma_{PoS}.W_1.d_{750}$ for *PT-comp* and *FR-comp*, and $surface.W_8.d_{750}$ for *Reddy*. As seen in Section 7.2, results for 250 and 500 dimensions have lower scores than the results for 750 dimensions. Results for 1,000 dimensions were mixed: they are slightly worse for *FR-comp* and *EN-comp*, and slightly better for *PT-comp*. Increasing the number of dimensions generates models that are progressively worse. These results suggest that the maximum vector quality is achieved between 750 and 1,000 dimensions.

B.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 three different runs of every DSM configuration (the original execution ρ_1, used elsewhere in this article, 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 three executions for every model.

Table 15
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 = 1,000	.694	(−0.8)
dim = 2,000	.645	(−5.8)
dim = 5,000	.636	(−6.7)
dim = 30,000	.552	(−15.1)
dim = 999,999	.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 = 1,000	.784	(−0.7)
dim = 2,000	.760	(−3.1)
dim = 5,000	.744	(−4.7)
dim = 30,000	.700	(−9.1)
dim = 999,999	.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 = 1,000	**.609**	(+0.7)
dim = 2,000	.601	(−0.1)
dim = 5,000	.505	(−9.7)
dim = 30,000	.532	(−7.0)
dim = 999,999	.500	(−10.2)

Table 16 reports the highest-Spearman configurations of ρ_{avg} for the *Reddy* and *EN-comp* data sets. 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[48] obtained from the scores of the three executions. Given the high stability of these models, results in the rest of the article were calculated and reported as ρ_1 for all data sets.

B.6 Data Filtering

Along with the verification of parameters, we also evaluate whether data set 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 data set quality. As per Roller, Schulte im Walde, and Scheible (2013), we consider two strategies of data removal: (1) removing individual outlier compositionality judgments through *z*-score

[48] The low standard deviation is not a unique property of high-ranking configurations: The average of deviations for all models was .004 for *EN-comp* and .006 for *Reddy*.

Table 16
Configurations with highest ρ_{avg} for nondeterministic models.

Data set	DSM	configuration	ρ_1	ρ_2	ρ_3	ρ_{avg}	stddev
Reddy	*glove*	$\mathit{lemma}_{PoS}.W_8.d_{250}$	.759	.760	.753	.757	.004
	w2v–cbow	$\mathit{surface}.W_1.d_{500}$	.796	.807	.799	.801	.006
	w2v–sg	$\mathit{surface}.W_1.d_{750}$	.812	.788	.812	.804	.014
EN-comp	*glove*	$\mathit{lemma}_{PoS}.W_8.d_{500}$	.651	.646	.650	.649	.003
	w2v–cbow	$\mathit{surface}^{+}.W_1.d_{750}$	.730	.732	.728	.730	.002
	w2v–sg	$\mathit{surface}^{+}.W_1.d_{750}$	.741	.732	.721	.731	.010

Table 17
Intrinsic quality measures for the raw and filtered data sets.

Data set	$\overline{\sigma}$		$P_{\sigma>1.5}$		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%	—	—

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.[49] These methods allow us to remove accidentally erroneous annotations, as well as annotators whose response deviated too much from the mean (in particular, spammers and non-native speakers).

Table 17 presents the evaluation of raw and filtered data sets regarding two quality measures: The average of the standard deviations for all NCs ($\overline{\sigma}$); and the proportion of NCs in the data set 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 data sets, as seen in the reduction of the proportion of NCs with high standard deviation, as well as in the reduction of the average standard deviation itself. We additionally present the data retention rate (DRR), which is the proportion of NCs that remained in the data set 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.

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, Schulte im Walde, and

[49] 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 $R = 0.5$.

Table 18
Extrinsic quality measures for the raw and filtered data sets.

Data set	*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

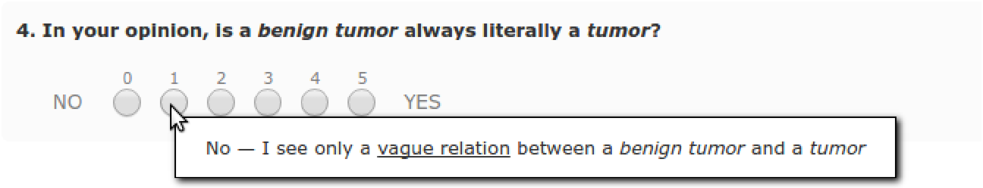
Scheible (2013) on their German data set, where only *R*-filtering was found to improve results under these quality measures. We present our findings in more detail in Ramisch, Cordeiro, and Villavicencio (2016).

We then consider whether filtering can have an impact on the performance of predicted compositionality scores. 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* data sets (using *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 data sets. Looking more specifically at the best configurations for each DSM (see Table 18), we can see that most results do not significantly change when the evaluation is performed on the raw or filtered data sets. 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.

Appendix C. Questionnaire

The questionnaire was structured in five 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 MWE components in their reply (eliciting a paraphrase).
4. Using a Likert scale from 0 to 5, judge how much of the meaning of the compound comes from the modifier and the head separately. Figure 11 shows an example for the judgment of the head.
5. Using a Likert scale from 0 to 5, judge how much of the meaning of the compound comes from its components.



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*

Figure 11
Evaluating compositionality of a compound regarding its head.

We require answers in an even-numbered scale (there are 6 possibilities between 0 and 5), as otherwise the participants could be biased toward the middle score. In order to help participants visualize the meaning of their reply, whenever their mouse hovers over a particular score, we present a guiding tooltip, as can be seen in Figure 11.

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 "non-compositional" for subtask 4 but "compositional" for subtask 5), even if they carefully selected their synonyms in response to subtask 3.

The request for synonyms before the multiple-choice questions prompts the participants to focus on the meaning of the compound. These synonyms can then also be taken into account when considering the semantic contribution of each element of the compound—we leave this for future work.

Appendix D. List of English Compounds

We present below the 90 nominal compounds in *EN-comp*$_{90}$ and the 100 nominal compounds in *EN-comp*$_{Ext}$, along with their human-rated compositionality scores. We refer to Reddy, McCarthy, and Manandhar (2011) for the other 90 compounds belonging to *Reddy* which, together with the former two sets, represent 280 nominal compounds in total.

D.1 Compounds in *EN-comp*$_{90}$

Compounds	hc_{HM}	Compounds	hc_{HM}
ancient history	1.95	closed book	0.68
armchair critic	1.33	computer program	4.50
baby buggy	3.94	con artist	2.10
bad hat	0.62	cooking stove	4.68
benign tumour	4.69	cotton candy	1.79
big fish	0.85	critical review	4.06
birth rate	4.60	dead end	1.32
black cherry	3.11	dirty money	2.21
bow tie	4.25	dirty word	2.48
brain teaser	2.65	disc jockey	1.25
busy bee	0.88	divine service	3.11
carpet bombing	1.24	dry land	3.95
cellular phone	3.78	dry wall	3.33
close call	1.59	dust storm	3.85

Compounds	hc_{HM}	Compounds	hc_{HM}
eager beaver	0.36	market place	3.00
economic aid	4.33	mental disorder	4.89
elbow grease	0.56	middle school	3.84
elbow room	0.61	milk tooth	1.43
entrance hall	4.17	mother tongue	0.59
eternal rest	3.25	narrow escape	1.75
fish story	1.68	net income	2.94
flower child	0.50	news agency	4.39
food market	3.82	noble gas	1.18
foot soldier	1.95	nut case	0.44
front man	1.64	old flame	0.58
goose egg	0.48	old hat	0.35
grey matter	2.39	old timer	0.89
guinea pig	0.45	phone book	4.25
half sister	2.84	pillow slip	3.70
half wit	1.16	pocket book	1.42
health check	4.17	prison guard	4.89
high life	1.67	prison term	4.79
inner circle	1.56	private eye	0.82
inner product	3.00	record book	3.70
insane asylum	3.95	research lab	4.75
insurance company	5.00	sex bomb	0.53
insurance policy	4.15	silver lining	0.35
iron collar	3.88	sound judgement	3.39
labour union	4.76	sparkling water	3.14
life belt	2.84	street girl	3.16
life vest	3.44	subway system	4.63
lime tree	4.61	tennis elbow	2.50
loan shark	1.00	top dog	1.05
loose woman	2.53	wet blanket	0.21
mail service	4.69	word painting	1.62

D.2 Compounds in *EN-comp*$_{Ext}$

Compounds	hc_{HM}	Compounds	hc_{HM}
academy award	3.52	blue blood	0.58
arcade game	3.80	blue print	1.04
baby blues	2.88	box office	0.88
backroom boy	1.48	brain drain	2.08
bad apple	1.13	bull market	1.23
banana republic	0.86	cable car	2.68
bankruptcy proceeding	4.78	calendar month	4.23
basket case	0.42	civil marriage	3.13
beauty sleep	2.96	cocoa butter	3.23
best man	3.12	computer expert	4.46
big cheese	0.36	contact lenses	3.64
big picture	1.48	copy cat	0.74
big wig	0.60	crime rate	4.39
biological clock	2.42	damp squib	0.95
black box	1.29	dark horse	0.65
black operation	1.39	day shift	4.54
blind alley	1.14	disability insurance	4.45
blood bath	1.38	double cross	1.14

Compounds	hc_{HM}	Compounds	hc_{HM}
double dutch	0.29	marketing consultant	4.00
double whammy	2.48	medical procedure	4.83
dream ticket	1.32	music festival	4.58
dutch courage	1.00	music journalist	4.54
fair play	2.59	noise complaint	4.52
fairy tale	1.68	pain killer	2.17
fall guy	1.36	peace conference	4.46
field work	2.10	peace talk	4.13
football season	4.04	pipe dream	0.91
fresh water	4.20	poison pill	0.96
freudian slip	2.35	radioactive material	4.61
ghost town	1.50	radioactive waste	4.58
glass ceiling	0.81	rainy season	4.23
grass root	0.86	rice paper	4.00
hard drive	2.17	shelf life	1.30
hard shoulder	1.52	skin tone	3.88
head hunter	1.50	smoke screen	1.11
health care	4.47	social insurance	2.83
heavy cross	1.17	speed trap	3.71
hen party	1.05	stag night	1.44
home run	2.86	sugar daddy	0.44
honey trap	1.22	tear gas	3.27
hot potato	0.56	time difference	4.41
incubation period	3.92	traffic control	3.69
information age	3.40	traffic jam	3.62
injury time	3.20	travel guide	4.38
insider trading	3.88	wedding anniversary	4.86
jet lag	2.67	wedding day	4.94
job fair	3.50	white noise	1.17
leap year	2.38	white spirit	1.31
love song	4.58	winter solstice	4.55
low profile	2.10	world conference	3.96

Appendix E. List of French Compounds

We present below the 180 nominal compounds in *FR-comp*, along with their human-rated compositionality scores.

Compounds	hc_{HM}	Translation (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/roadside’ (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 living*)
bonne humeur	4.53	‘good mood’ (lit. *good mood*)

Compounds	hc_{HM}	Translation (Gloss)
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	'most important helper/assistant' (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 (sexual connotation)' (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	'problem, difficulty' (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*)
crème fraîche	3.73	'French sour cream' (lit. *cream fresh*)
crème glacée	4.75	'ice cream' (lit. *cream icy*)
dernier cri	0.67	'something trendy' (lit. *last scream*)
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/frustration' (lit. *shower cold*)
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 soft/sweet*)
eau minérale	4.00	'mineral water' (lit. *water mineral*)
eau potable	5.00	'drinking water' (lit. *water potable*)
eau vive	3.44	'jellyfish' (lit. *water lively*)
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 conductor*)

Compounds	$\mathrm{hc}_{\mathrm{HM}}$	Translation (Gloss)
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/fat word*)
gros plan	1.87	'close-up' (lit. *large/fat 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 integration, employability' (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. *matter grey*)
matière première	2.90	'raw material' (lit. *matter primary*)
mauvaise foi	2.38	'bad faith' (lit. *bad faith*)
mauvaise langue	2.21	'gossiper' (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 domestic*)
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*)
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 pastry' (lit. *small oven*)
petit nègre	0.50	'pidgin or 'badly spoken' French' (lit. *little black-person*)
petite annonce	2.69	'classified ad' (lit. *small announcement*)

Compounds	hc_{HM}	Translation (Gloss)
petite nature	0.47	‘sensitive/fragile person’ (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	‘prime 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	‘cold blood, self-control’ (lit. *blood cold*)
second degré	1.40	‘irony, 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, discussion’ (lit. *table round*)
tapis rouge	3.31	‘red carpet, luxurious welcoming’ (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*)

Appendix F. List of Portuguese Compounds

We present below the 180 nominal compounds in *PT-comp*, along with their human-rated compositionality scores.

Compounds	hc_{HM}	Translation (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*)

Compounds	hc_{HM}	Translation (Gloss)
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*)
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*)

Compounds	hc_{HM}	Translation (Gloss)
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*)
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*)

Compounds	hc_{HM}	Translation (Gloss)
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*)

Acknowledgments

This work has been partly funded by projects PARSEME (Cost Action IC1207), PARSEME-FR (ANR-14-CERA-0001), AIM-WEST (FAPERGS-INRIA 1706-2551/13-7), CNPq (312114/2015-0, 423843/2016-8) "Simplificação Textual de Expressões Complexas," sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91. We would like to thank the anonymous reviewers who provided numerous helpful suggestions, Alexis Nasr for reviewing earlier versions of this article, Rodrigo Wilkens and Leonardo Zilio for contributing to the data set creation, and all anonymous annotators who judged the compositionality of compounds.

References

Agirre, Eneko, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31–June 5, 2009*, pages 19–27, Boulder, CO.

Artstein, Ron, and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Baldwin, Timothy, and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2nd edition. CRC Press, Taylor and Francis Group, Boca Raton, FL, pages 267–292.

Bannard, Colin, Timothy Baldwin, and Alex Lascarides. 2003. 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)*, pages 65–72, Stroudsburg, PA.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. 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)*, pages 238–247, Baltimore.

Baroni, Marco, and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bick, Eckhard. 2000. The Parsing System "palavras": *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, University of Aarhus.

Boos, Rodrigo, Kassius Prestes, and Aline Villavicencio. 2014. Identification of multiword expressions in the brWaC. In *Proceedings of the Conference on Language Resources and Evaluation 2014*, pages 728–735, ELRA. ACL Anthology Identifier: L14–1429.

Bride, Antoine, Tim Van de Cruys, and Nicholas Asher. 2015. A generalisation of lexical functions for composition in distributional semantics. In *Association for Computational Linguistics (1)*, pages 281–291.

Bullinaria, John A., and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. 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)*, pages 1–7, Beijing.

Cap, Fabienne, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 19–28, Association for Computational Linguistics, Denver.

Carpuat, Marine, and Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Proceedings of NAACL/HLT 2010*, pages 242–245, Los Angeles.

Church, Kenneth Ward, and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Cordeiro, Silvio, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. 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)*, pages 1986–1997, Berlin.

Cordeiro, Silvio, Carlos Ramisch, and Aline Villavicencio. 2016. mwetoolkit+sem: Integrating word embeddings in the mwetoolkit for semantic MWE processing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1221–1225, European Language Resources Association (ELRA), Paris.

Curran, James R., and Marc Moens. 2002. Scaling context space. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 231–238.

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

Evert, Stefan. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany.

Farahmand, Meghdad, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, Association for Computational Linguistics, Denver.

Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Ferret, Olivier. 2013. Identifying bad semantic neighbors for improving distributional thesauri. In *Association for Computational Linguistics (1)*, pages 561–571.

Finlayson, Mark, and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Association for Computational Linguistics 2011 Workshop on MWEs*, pages 20–24, Portland, OR.

Firth, John R. 1957. A synopsis of linguistic theory, 1930–1955. In F. R. Palmer, ed., *Selected Papers of J. R. Firth*, pages 168–205, Longman, London.

Fleiss, Joseph L., and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Frege, Gottlob. 1892/1960. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50. Translated, as ‘On Sense and Reference,’ by Max Black.

Freitag, Dayne, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 25–32.

Girju, Roxana, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech & Language*, 19(4):479–496.

Goldberg, Adele E. 2015. *Compositionality*, Chapter 24. Routledge, Amsterdam.

Guevara, Emiliano. 2011. Computing semantic compositionality in distributional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 135–144, Association for Computational Linguistics, Stroudsburg, PA.

Harris, Zellig. 1954. Distributional structure. *Word*, 10:146–162.

Hartung, Matthias, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. 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 (Volume 1)*, pages 54–64.

Hendrickx, Iris, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. Semeval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of *SEM 2013 (Volume 2 — SemEval)*, pages 138–143, Association for Computational Linguistics.

Hwang, Jena D., Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. Propbank annotation of multilingual light verb constructions. In *Proceedings of the LAW 2010*, pages 82–90, Association for Computational Linguistics.

Jagfeld, Glorianna, and Lonneke van der Plas. 2015. Towards a better semantic role labelling of complex predicates. In *Proceedings of NAACL Student Research Workshop*, pages 33–39, Denver.

Jurafsky, Daniel, and James H. Martin. 2009. *Speech and Language Processing*, 2nd Edition, Prentice-Hall, Inc., Upper Saddle River, NJ.

Kiela, Douwe, and Stephen Clark. 2014. 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*, pages 21–30.

Köper, Maximilian, and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of German particle verbs. In *HLT-NAACL*, pages 353–362.

Kruszewski, Germán, and Marco Baroni. 2014. Dead parrots make bad pets: Exploring modifier effects in noun phrases. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics, *SEM@COLING 2014, August 23-24, 2014*, pages 171–181, The *SEM 2014 Organizing Committee, Dublin.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Lapesa, Gabriella, and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Lapesa, Gabriella, and Stefan Evert. 2017. Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *EACL 2017*, pages 394–400.

Lauer, Mark. 1995. How much is enough?: Data requirements for statistical NLP. *CoRR*, abs/cmp-lg/9509001.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (Volume 2)*, pages 768–774.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the Association for Computational Linguistics 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Association for Computational Linguistics, Sapporo, Japan.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In

Advances in Neural Information Processing Systems, pages 3111–3119.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Mitchell, Jeff, and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Association for Computational Linguistics*, pages 236–244.

Mitchell, Jeff, and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Mohammad, Saif, and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *CoRR*, abs/1203.1858.

Nakov, Preslav. 2008. Paraphrasing verbs for noun compound interpretation. In *Proceedings of the LREC Workshop Towards a Shared Task for MWEs*, pages 46–49.

Nakov, Preslav. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19:291–330.

Nivre, Joakim, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Conference on Language Resources and Evaluation (Volume 6)*, pages 2216–2219.

Padó, Sebastian, and Mirella Lapata. 2003. Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (Volume 1)*, pages 128–135.

Padó, Sebastian, and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Padró, Muntsa, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. 2014a. Comparing similarity measures for distributional thesauri. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2964–2971, European Language Resources Association, Reykjavik.

Padró, Muntsa, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. 2014b. Nothing like good old frequency: Studying context filters for distributional thesauri. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Short Papers)*, pages 419–424, Doha, Qatar.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Association for Computational Linguistics, Doha, Qatar.

Ramisch, Carlos, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, Aline Villavicencio, and Rodrigo Wilkens. 2016. How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pages 156–161.

Ramisch, Carlos, Silvio Ricardo Cordeiro, and Aline Villavicencio. 2016. Filtering and measuring the intrinsic quality of human compositionality judgments. In *Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016)*, pages 32–37, Berlin.

Reddy, Siva, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011)*, pages 210–218, Chiang Mai, Thailand.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the ACL 2009 Workshop on MWEs*, pages 47–54, Singapore.

Riedl, Martin, and Chris Biemann. 2015. A single word is not enough: Ranking multiword expressions using distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2430–2440, Association for Computational Linguistics.

Roller, Stephen, and Sabine Schulte im Walde. 2014. Feature norms of German noun compounds. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 104–108, Association for Computational Linguistics.

Roller, Stephen, Sabine Schulte im Walde, and Silke Scheible. 2013. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 32–41, Association for Computational Linguistics.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002, Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*. Springer, New York, pages 1–15.

Salehi, Bahar, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden.

Salehi, Bahar, Paul Cook, and Timothy Baldwin. 2015. 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*, pages 977–983, Denver.

Salehi, Bahar, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 54–59, Association for Computational Linguistics, Denver.

Salle, Alexandre, Aline Villavicencio, and Marco Idiart. 2016. 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)*, pages 419–424, Berlin.

Schmid, Helmut. 1995. Treetagger—A language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Schneider, Nathan, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval 2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of SemEval*, pages 546–559, San Diego.

Schone, Patrick, and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing*, pages 100–108, Pittsburgh.

Schulte im Walde, Sabine, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. GhoSt-NN: A representative gold standard of German noun-noun compound. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 2285–2292.

Schulte im Walde, Sabine, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of *SEM 2013 (Volume 1)*, pages 255–265. Association for Computational Linguistics.

Socher, Richard, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211.

Stymne, Sara, Nicola Cancedda, and Lars Ahrenberg. 2013. Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics*, 39(4):1067–1108.

Tsvetkov, Yulia, and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.

Turney, Peter D., and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Van de Cruys, Tim, Laura Rimell, Thierry Poibeau, and Anna Korhonen. 2012. Multiway tensor factorization for unsupervised lexical acquisition. In *COLING 2012*, pages 2703–2720.

Yazdani, Majid, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Association for Computational Linguistics, Lisbon.