Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time

Silvio Cordeiro[1,2], **Carlos Ramisch**[1], **Marco Idiart**[3], **Aline Villavicencio**[2]
[1] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[3] Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
silvioricardoc@gmail.com carlos.ramisch@lif.univ-mrs.fr
marco.idiart@gmail.com avillavicencio@inf.ufrgs.br

Abstract

Distributional semantic models (DSMs) are often evaluated on artificial similarity datasets containing single words or fully compositional phrases. We present a large-scale multilingual evaluation of DSMs for predicting the degree of semantic compositionality of nominal compounds on 4 datasets for English and French. We build a total of 816 DSMs and perform 2,856 evaluations using *word2vec*, *GloVe*, and PPMI-based models. In addition to the DSMs, we compare the impact of different parameters, such as level of corpus preprocessing, context window size and number of dimensions. The results obtained have a high correlation with human judgments, being comparable to or outperforming the state of the art for some datasets (Spearman's ρ=.82 for the *Reddy* dataset).

1 Introduction

Distributional semantic models (DSMs) use context information to represent the meaning of lexical units as vectors. They normally focus on the accurate semantic representation of single words. It is based on single words that many optimizations for these models have been proposed (Lin, 1999; Erk and Padó, 2010; Baroni and Lenci, 2010). This is particularly true for *word embeddings*, that is, a type of DSM where distributional vectors are obtained as a by-product of training a neural network to learn a function between words and their contexts (Mikolov et al., 2013a).

Simultaneously, there has been intensive research on models to compose individual word vectors in order to create representations for larger units such as phrases, sentences and even whole documents (Mitchell and Lapata, 2010; Mikolov et al., 2013a). Larger units can often be assumed to have their meanings derived from their parts according to the language's grammar, but this is not always the case (Sag et al., 2002). Many multiword units are associated with idiomatic interpretations, unrelated to the meaning of the component words (e.g. *silver bullet, eager beaver*).

Precision-oriented NLP applications need to be able to identify partly-compositional and idiomatic cases and ensure meaning preservation during processing. Compositionality identification is a first step towards complete semantic interpretation in tasks such as machine translation (to translate non-compositional compounds as a unit), word sense disambiguation (to avoid assigning a sense to parts of non-compositional compounds), and semantic parsing (to identify complex predicates and their arguments).

Even when larger units are explicitly represented in DSMs (McCarthy et al., 2003; Reddy et al., 2011; Mikolov et al., 2013c; Ferret, 2014), it is not clear whether the quality of these representations is comparable to the representations of single words. In particular, when building vectors for larger units, their generally lower frequencies in corpora (Kim and Baldwin, 2006) may combine with morphosyntactic phenomena to increase sparsity even further, often requiring non-trivial preprocessing (lemmatization and word reordering) to conflate variants.

This paper presents a large-scale multilingual evaluation of DSMs and their parameters for the task of compositionality prediction of nominal compounds in French and English. We examine parameters like the level of corpus preprocessing, the size of the context window and the number of dimensions for context representation. Additionally, we compare standard DSMs based on positive pointwise mutual information (PPMI)

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1986–1997,
Berlin, Germany, August 7-12, 2016.

against widely used word embedding tools such as *word2vec*, henceforth *w2v* (Mikolov et al., 2013c), and *GloVe* (Pennington et al., 2014). We start with a discussion of related work (§2) and the materials and methods used (§3). We report on the evaluations performed (§4) and finish with conclusions and future work (§5).

2 Related Work

We define *nominal compounds* as conventional noun phrases composed by two or more words, such as *science fiction* (Nakov, 2013). In English, they are often expressed as *noun compounds* but their syntactic realization may vary for different languages. For instance, one of the equivalent forms in French involves a denominal adjective used as modifier (e.g. *cell death* and the corresponding *mort cellulaire*).[1] In this paper, we focus on 2-word nominal compounds involving modifiers that are nouns (e.g. *word embedding*) or adjectives (e.g. *hard time*).

Semantically, nominal compounds may display a wide range of idiomaticity, from compositional cases like *access road* to idiomatic or non-compositional cases like *gravy train*, whose meaning is unrelated to its parts.[2] Even when there is a level of compositionality in the compound, the contribution of each word may vary considerably, independently from its status as a syntactic head or modifier, as *cash* in *cash cow* versus *tears* in *crocodile tears*. Indeed, various annotation scales have been proposed as means to collect human judgments about compositionality. Particularly for nominal compounds, Reddy et al. (2011) used a 6-point scale to collect judgments on the literal or figurative use of nominal compounds and its components in English. Similar judgments have also been collected for 244 German compounds, for which an average of 30 judgments on a scale from 1 to 7 were gathered through crowdsourcing (Roller et al., 2013). An alternative to multi-point scales is the binary judgment adopted by Farahmand et al. (2015), for a dataset of English nominal compounds.

There has been much interest in creating semantic representations of larger units, such as phrases (Mikolov et al., 2013b), sentences and documents (Le and Mikolov, 2014), and in examining whether it is possible to accurately derive the semantics of a compound or multiword expression from its parts (McCarthy et al., 2003; Baldwin et al., 2003; Tratz and Hovy, 2010; Reddy et al., 2011). For the latter, proposals include using additive and multiplicative functions to combine vector representations of component words (Mitchell and Lapata, 2008; Reddy et al., 2011), calculating the overlap between the components and the expression (McCarthy et al., 2003) and looking at the literality of translations into multiple languages (Salehi et al., 2014). Other proposals to explicitly represent the semantics of nominal compounds include the use of paraphrases (Lauer, 1995; Nakov, 2008; Hendrickx et al., 2013), and inventories of semantic relations (Girju et al., 2005).

The ability of DSMs for accurately capturing semantic information may be affected by a number of factors involved in constructing the models, such as the source *corpus*, *context* representation, and parameters of the *model*. Relevant corpus parameters include size (Ferret, 2013; Mikolov et al., 2013c) and quality (Lapesa and Evert, 2014). Factors related to context representation include the context window size and the number of context dimensions adopted for a model (Lapesa and Evert, 2014); the choice of contexts to be used with targets (syntactic dependencies vs. bag-of-words) (Agirre et al., 2009); the use of morphosyntactic information (Padó and Lapata, 2003; Padó and Lapata, 2007); context filtering (Riedl and Biemann, 2012; Padró et al., 2014a); and dimensionality reduction methods (van de Cruys et al., 2012). Important model parameters that have been studied include the choice of association and similarity measures (Curran and Moens, 2002) and the use of subsampling and negative sampling techniques (Mikolov et al., 2013c). However, the particular effects may be heterogeneous and depend on the task and model (Lapesa and Evert, 2014). In this paper, we examine the impact of both corpus and context parameters for a variety of models, for the task of nominal compound compositionality prediction in English and French.

For the choice of particular DSM, contradictory results have been published showing the superiority of neural models (Baroni et al., 2014) and of more traditional but carefully designed models (Levy et al., 2015). The former were also reported as a better fit to behavioral data on semantic prim-

[1] In French, one can also use a preposition and optional determiner, like *cancer du poumon* (*lung cancer*).

[2] It refers to an initiative that provides money to many people without much effort.

ing tasks (Mandera et al., 2016). Moreover, these evaluations are often performed on single-word similarity tasks (Freitag et al., 2005; Camacho-Collados et al., 2015) and little has been said about the use of word embeddings for the compositionality prediction of multiword expressions. Two notable exceptions are the recent works of Salehi et al. (2015) and Yazdani et al. (2015). Salehi et al. (2015) show that word embeddings are more accurate in predicting compositionality than a simplistic count-based DSM. Yazdani et al. (2015) focus on the composition function, using a lightly supervised neural network to learn the best combination strategy for individual word vectors. In order to consolidate previous punctual results, we present a large-scale and systematic evaluation, comparing DSMs and their parameters, on several compositionality datasets.

3 Materials and Methods

We examine the impact of corpus parameters related to the target language and the degree of corpus preprocessing adopted. We also investigate context parameters related to the size of the context window and the number of dimensions used to represent context.

3.1 Corpora Preprocessing

We use the lemmatized and POS-tagged versions of the ukWaC for English (∼2 billion tokens) and frWaC (∼1.6 billion tokens) for French (Baroni et al., 2009) to train the models and build vector representations of words and compounds. For each corpus, we re-tokenize all target compounds as a single word with a separator (e.g. *monkey business* → *monkey_business*) and re-tag them using a single manually selected tag per compound to handle POS-tagging errors.[3] 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[4] is applied to normalize words across English spelling variants (e.g. *color* → *colour*).

To test the influence of preprocessing in model accuracy, for each corpus, we generate four variants with different degrees of abstraction:

1. *surface*$^+$: the original corpus with no preprocessing, containing surface forms.
2. *surface*: stopword removal; generating a corpus of surface forms of content words.
3. *lemma*: stopword removal and lemmatization; generating a corpus of lemmas of content words.
4. *lemmaPOS*: stopword removal, lemmatization and POS-tagging; generating a corpus of content words, represented as lemma/tag.

The operation of stopword removal eliminates from the corpus all function words, leaving only nouns, adjectives, adverbs and verbs. In lemmatized corpora, the lemmas of proper names are replaced by placeholders.

3.2 Compositionality Datasets

For evaluation, we use nominal compound compositionality datasets for English (*Reddy*, *Reddy++* and *Farahmand*) and for French (*FR-comp*). They provide annotations as to whether a given compound is more idiomatic or more compositional.

Reddy contains compositionality judgments for 90 compounds and their individual word components, in a scale of literality from 0 (idiomatic) to 5 (literal), collected with Mechanical Turk (Reddy et al., 2011). For each compound, compositionality scores are averaged over its annotators. Compounds included in the dataset were selected to balance frequency range and degree of compositionality (low, middle and high). We use only the global compositionality score, ignoring individual word judgments. With a few exceptions (e.g. *sacred cow*), most compounds are formed exclusively by nouns.

Reddy++ is a new resource created for this evaluation (Ramisch et al., 2016). It extends the *Reddy* set with an additional 90 English nominal compounds, in a total of 180 entries. Scores also range from 0 to 5 and were collected through Mechanical Turk and averaged over the annotators. The extra 90 entries include some adjective-noun compounds and are balanced with respect to frequency and compositionality. We focus our evaluation on this combined dataset, since it includes *Reddy*. However, to allow comparison with state of the art, we also report results individually for *Reddy*.

Farahmand contains 1042 English compounds extracted from Wikipedia with binary non-compositionality judgments by four experts (Farahmand et al., 2015). We consider a compound as non-compositional if at least two judges agree that it is non-compositional, following Yaz-

[3] We use a simplified tag set (e.g. `v` instead of `vvz`).

[4] `https://hunspell.github.io`

dani et al. (2015). In our evaluations, we use the sum of all judgments in order to have a single numeral compositionality score, ranging from 0 (compositional) to 4 (idiomatic).

FR-comp is also a new resource created for this evaluation (Ramisch et al., 2016). It contains 180 adjective-noun and noun-adjective compounds in French, such as *belle-mère* (*mother-in-law*, lit. *beautiful-mother*) and *carte bleue* (*credit card*, lit. *blue card*). This dataset was constructed in the same manner as the extension to *Reddy*, that is, using crowdsourcing and average numerical scores. Special care was taken to guarantee that annotators were native speakers by asking them to provide paraphrases along with compositionality scores.

The new datasets *Reddy++* and *FR-comp* are similar to *Reddy*. For instance, the average standard deviation of compound scores given by different annotators is $\sigma = 1.17$ for the new compounds in *Reddy++*, $\sigma = 1.15$ for *FR-comp* and $\sigma = 0.99$ for *Reddy*. Their detailed evaluation is presented by Ramisch et al. (2016).

3.3 DSM Models

We build three types of DSMs: models based on sparse PPMI cooccurrence vectors, as well as those constructed with *word2vec* and *GloVe*.

PPMI For each target word or compound, we extract from the corpus its neighboring nouns and verbs in a symmetric sliding window of w words to the left/right[5], using a linear decay weighting scheme with respect to its distance d to the target (Levy et al., 2015). In other words, each cooccurrence count of target-context pairs is incremented by $w + 1 - d$ instead of 1. The representation of a target is a vector containing the positive pointwise mutual information (PPMI) association scores between the target and its contexts.[6]

In *PPMI-thresh*, we follow Padró et al. (2014b) to select the top k most relevant contexts (highest PPMI) for each target. No further dimensionality reduction is applied.

In *PPMI-TopK*, we use a fixed global list of 1000 contexts, built by looking at the most frequent words in the corpus: the top 50 are skipped, and the next 1000 are taken (Salehi et al., 2015). No further dimensionality reduction is applied.

In *PPMI-SVD*, for each target, contexts that appear less than 1000 times are discarded.[7] We then use the Dissect toolkit[8] (Dinu et al., 2013) in order to build a PPMI matrix and reduce its dimensionality using singular value decomposition (SVD) to factorize the matrix.

w2v Uses the *word2vec* toolkit based on neural networks to predict target/context cooccurrence (Mikolov et al., 2013a). We build models from two variants of word2vec: CBOW (*w2v-cbow*) and skipgram (*w2v-sg*). In both cases, the configurations are the default ones, except for the following: no hierarchical softmax; negative sampling of 25; frequent-word downsampling weight of 10^{-6}; runs 15 training iterations. We use the default minimum word count threshold of 5.

glove We use the count-based DSM of Pennington et al. (2014), which implements a factorization of the co-occurrence count matrix. The configurations are the default ones, except for the following: internal cutoff parameter $x_{max} = 75$; builds co-occurrence matrix in 15 iterations. Due to the large vocabulary size, we use a minimum word count threshold of 5 for lemma-based models, 15 for *surface* and 20 for *surface*$^+$.

For each DSM, we evaluate the influence of a set of parameters. By varying the values of these parameters, we build a total of 408 models per language. The parameters are:

- WORDFORM: Refers to one of the four variants of each corpus: *surface*$^+$, *surface*, *lemma*, and *lemmaPOS*.
- WINDOWSIZE: Indicates within how many words to the left/right we are searching for target-context co-occurrence pairs. In this work we explore windows of sizes of 1, 4 and 8.
- DIMENSION: Each model is constructed to have a maximum number of final dimensions for each vector. We generate models with 250, 500 and 750 dimensions.

3.4 Compositionality Prediction

To predict the compositionality of a nominal compound w_1w_2 using the DSMs, we use as a measure the cosine similarity between the compound

[5] Syntactic context definition is planned as future work.

[6] *PPMI* vectors are built using minimantics `https://github.com/ceramisch/minimantics`.

[7] Aggressive filtering was required because SVD seems quite sensitive to low-frequency contexts.

[8] `http://clic.cimec.unitn.it/composes/toolkit/index.html`

vector representation $v(w_1w_2)$ and the sum of the vector representations of the component words:

$$\cos(\ v(w_1w_2),\ v(w_1 + w_2)\ )$$

where for $v(w_1 + w_2)$ we use the normalized sum

$$v(w_1 + w_2) = \frac{v(w_1)}{||v(w_1)||} + \frac{v(w_2)}{||v(w_2)||}.$$

In this framework, a compound is compositional if the compound representation is close to the sum of its components representations (cosine is close to 1), and it is idiomatic otherwise.

One possible improvement of the predictive model would consist in using more sophisticated composition functions instead of sum, such as the multiplicative model of Mitchell and Lapata (2008). However, we want to first assess the performance of a simple additive function. Other optimized functions like the ones proposed by Yazdani et al. (2015) could also be verified, but are out of the scope of this paper, since they are based on supervised learning.

3.5 Evaluation Setup

We evaluate the compositionality models and their parameters on the datasets described in Section 3.2. For *Reddy*, *Reddy++* and *FR-comp*, we report Spearman's ρ correlation between the ranking provided by humans and those calculated from the models. We follow Yazdani et al. (2015) and report the best F1 score (BF1) obtained for the *Farahmand* dataset, by calculating the F1 score for the top k compounds classified as positive (non-compositional), for all possible values of k.

Given the high number of experiments we performed, we report the best performance of each model type. For instance, the performances reported for *w2v-cbow* using different values of WINDOWSIZE are the best configurations across all possible values of other parameters such as DIMENSION and WORDFORM. This avoids reporting local maxima that can arise if one fixes all other parameters when evaluating a given one (Lapesa and Evert, 2014).

For *Reddy++* and *Farahmand*, we distinguish between *strict* evaluation, reported in the form of wider bars in the figures, and *loose* evaluation, shown as narrow blue bars in the figures. Strict evaluation corresponds to the performance of the model only on those compounds that have a vector representation in all underlying DSMs, 175 out of 180 for *Reddy++* and 913 out of 1042 for *Farahmand*. Loose evaluation considers the full dataset, using a fallback strategy for the imputation of missing values, assigning the average compositionality score to absent compounds (Salehi et al., 2015). This is particularly important for *Farahmand*, which contains more rare compounds such as *universe human* and *mankind instruction* so that 129 compounds are missing in the corpus. Only strict evaluation is reported for *FR-comp*, as all compounds are frequent enough in FRWaC.

The vectors generated by *w2v* and *glove* have some non-determinism due to random initialization. To assess its impact on results, we report the average of 3 runs using identical configurations and use error bars in the graphics.[9]

4 Results

We report results on each dataset separately and then discuss findings that hold for all datasets.

4.1 *Reddy++* and *Reddy* Datasets

Figure 1 summarizes the results for *Reddy++* dataset.[10] Overall, *w2v-cbow* ($\rho = 0.73$), *w2v-sg* ($\rho = 0.73$), *PPMI-SVD* ($\rho = 0.72$) and *PPMI-thresh* ($\rho = 0.71$) obtain similar results. In spite of this, except for the two best *w2v* models, all differences were deemed statistically significant (Wilcoxon rank correlation test, $p < 0.05$).

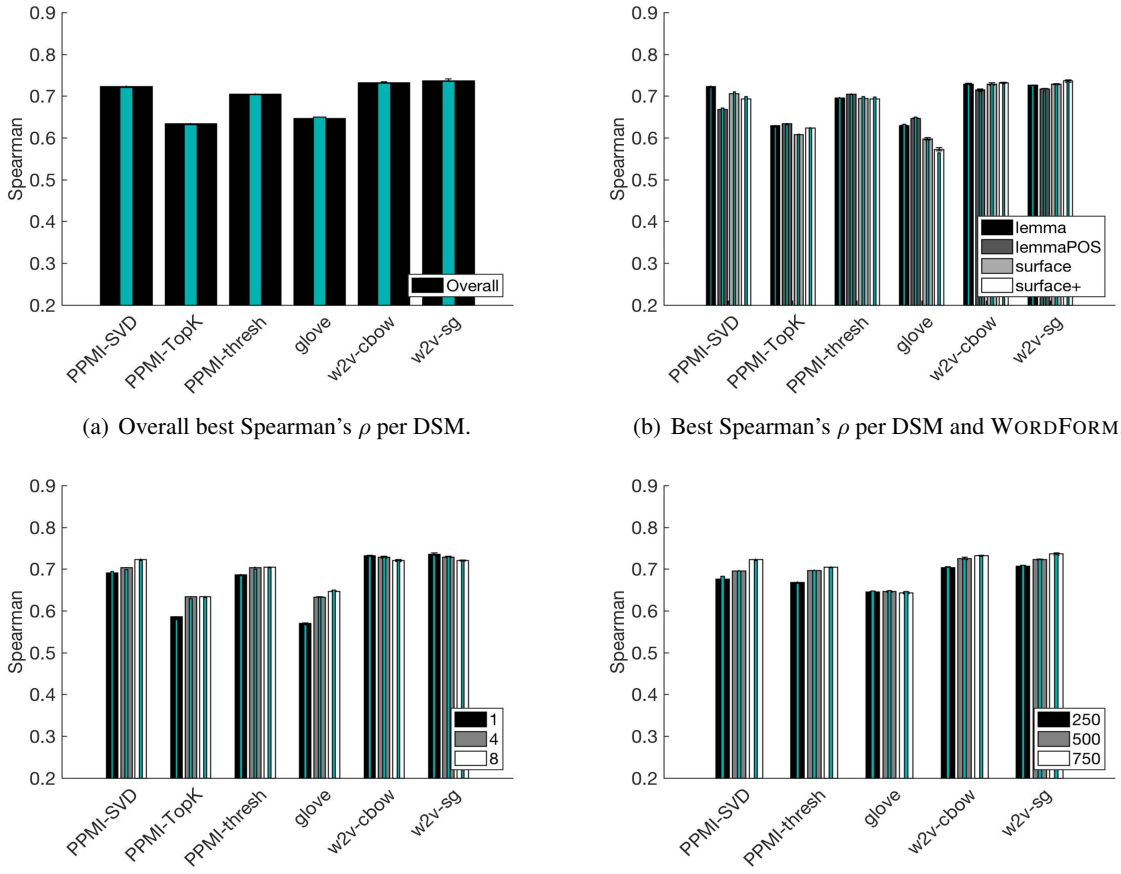
Figure 1(b) shows the influence of the degree of corpus preprocessing (shown as WORDFORM in these figures). The results are heterogeneous, as the best *w2v* models seem to profit from the presence of stopwords, unlike the other models for which more preprocessing (*lemma* and *lemma-POS*) leads to better results. One exception is *PPMI-SVD* for which the use of *lemmaPOS* drastically reduces performance.[11]

For WINDOWSIZE, Figure 1(c), although increasing context size seems to help DSMs (at least up to 4), for the best *w2v* models, a better result is obtained with limited context of 1 word left/right. Probably the interaction between the subsampling strategy and randomized window size explains why increasing this value does not improve the

[9] Error bars are barely visible because results are stable.

[10] In the remainder of this section, we will discuss strict evaluation results (outer bars).

[11] Further investigation must be done to determine the cause of this reduction as an increase in vocabulary size alone is insufficient to explain the effect, given that both surface forms outperform it.



(a) Overall best Spearman's ρ per DSM.

(b) Best Spearman's ρ per DSM and WORDFORM.

(c) Best Spearman's ρ per DSM and WINDOWSIZE.

(d) Best Spearman's ρ per DSM and DIMENSION.

Figure 1: Spearman's ρ for different DSM parameters on *Reddy++* dataset.

results. *PPMI-SVD* can use extra information from larger window sizes (WINDOWSIZE=8) better than models based on context filtering. This is probably related to the aggressive context filter, which keeps only very salient cooccurrences even in large windows.

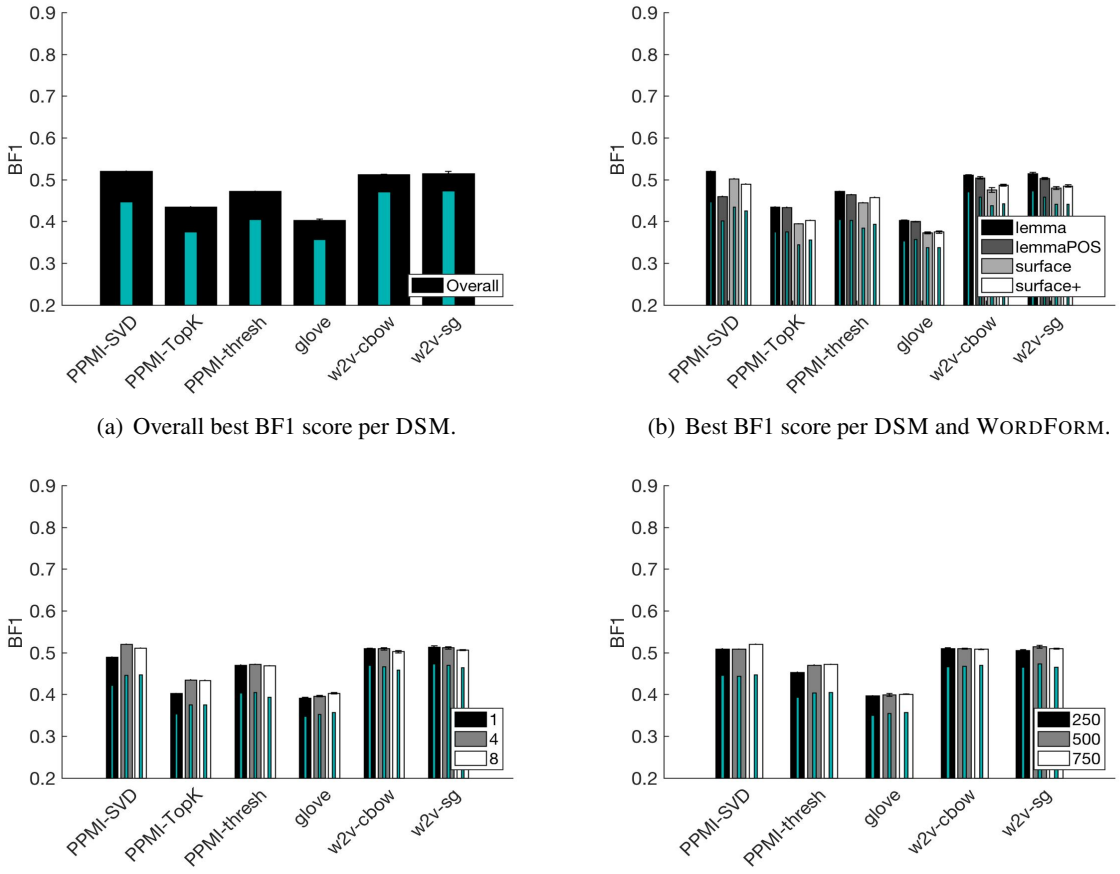
The results for context vector dimensionality, Figure 1(d), show, as expected, that the best results are obtained with larger dimensions (DIMENSION=750) for all models, except for *glove*, which displays very similar results independently of the number of dimensions.

Examining the *Reddy* dataset alone, the same trends for all parameters were found, but with higher results. The overall best performances on *Reddy* were quite similar: *w2v-cbow* ($\rho = 0.82$), *w2v-sg* ($\rho = 0.81$), *PPMI-SVD* ($\rho = 0.80$) and *PPMI-thresh* ($\rho = 0.79$), and the differences are significant except for the two best *w2v* models. The 90 compounds added to *Reddy++* seem to be more difficult to assess than the original ones, probably because they include many adjectives, which have been found harder to judge for compositionality than nouns (Ramisch et al., 2016).

4.2 *Farahmand* Dataset

Figure 2(a) shows the overall best model for the *Farahmand* dataset. *PPMI-SVD* reached a BF1 score of 0.52, with DIMENSION=750, WINDOWSIZE=4, using *lemma*, and both *w2v* (BF1=0.51) obtain comparable results with similar configurations.

These results show a marked difference between the loose (the narrower bars in the figures) and the strict evaluation (wider bars). The former uses a fallback strategy for the imputation of missing values that does not accurately reflect how the compositionality scores vary. Indeed, we observed that compounds that do not appear very often in our corpora tend to be non-compositional, whereas most of the compound occurrences are compositional, increasing average compositionality. For instance, the 10 most compositional compounds in *Reddy++* occur an average of 26551



(a) Overall best BF1 score per DSM.

(b) Best BF1 score per DSM and WORDFORM.

(c) Best BF1 score per DSM and WINDOWSIZE.

(d) Best BF1 score per DSM and DIMENSION.

Figure 2: BF1 scores for different DSM parameters on *Farahmand* dataset.

times in the UKWaC vs 1096 times for the 10 least compositional ones. Spearman rank correlation between frequency and compositionality in *Reddy++* is $\rho = 0.43$.[12] In short, even if a fallback strategy is adopted as the means to obtain a lower-bound for performance, it may be unrelated to the real performance for the missing compounds.

For most models, corpus preprocessing resulted in better scores, with WORDFORM=*lemma* outperforming all other forms of preprocessing, especially for French. Concatenating lemmas and POS tags does not seem to help, probably due to decreasing word frequencies without substantial gain in informativeness (Figure 2(b)).

The impact of WINDOWSIZE has a similar trend to the one found for the *Reddy++* and *Reddy* datasets (Figure 2(c)). That is, the larger window was preferred by most models, but the average difference between the best and the worst size for each DSM is only 0.01. For DIMENSION, a larger number resulted in better scores, as expected, with 750 being the best for all models in Figure 2(d). Nonetheless, here too the average difference in scores between DIMENSION=750 and 250 is 0.01.

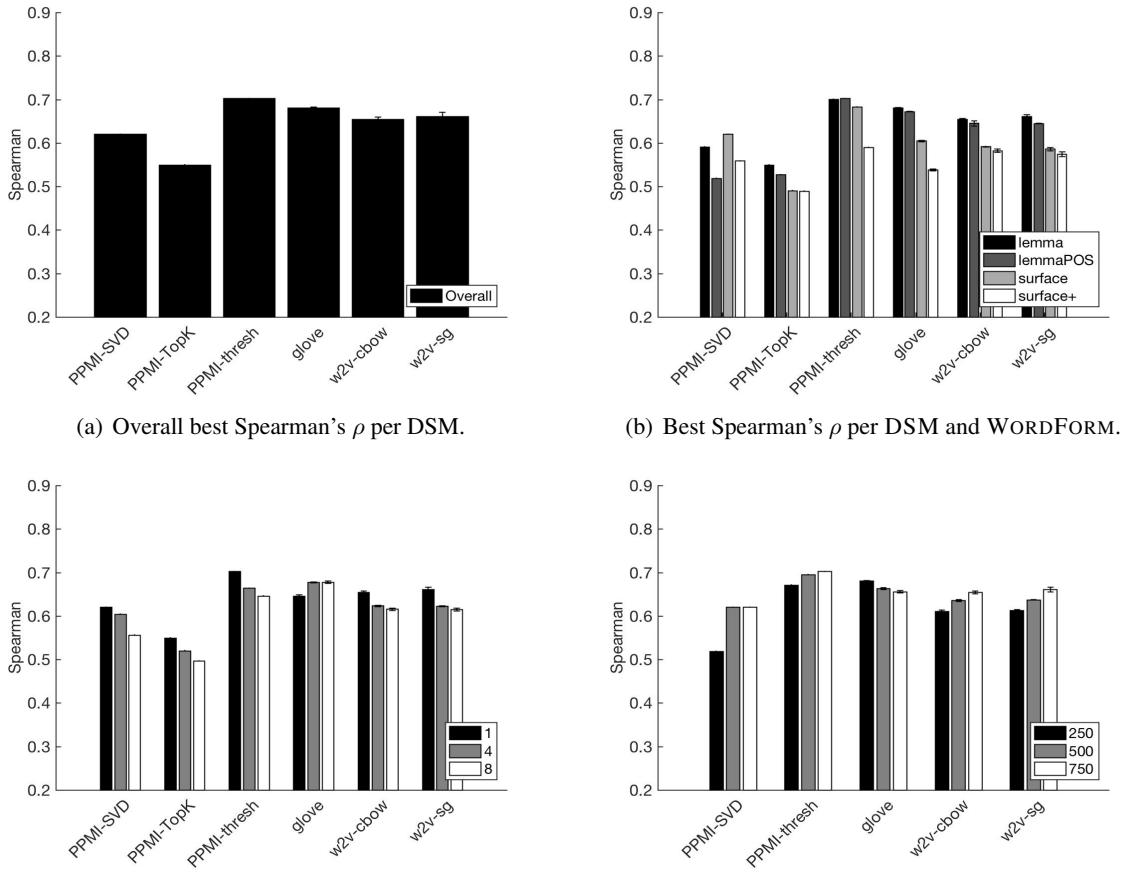
4.3 *FR-comp* Dataset

Globally, for the *FR-comp* dataset, *PPMI-thresh* ($\rho = 0.70$) outperforms *glove* ($\rho = 0.68$) and *w2v* ($\rho = 0.66$), as can be seen in Figure 3(a).[13]

For morphologically rich languages like French, Figure 3(b) indicates that working on lemmatized data often yields better results than working on surface forms. Lemmas conflate the frequencies for all the many morphologically inflected variants which would otherwise be dispersed in different surface forms. Therefore, it is not surprising that the best results concerning WORDFORM are achieved by *lemma*. These results differ from English, where a corpus without any preprocess-

[12] We report these figures for *Reddy++* because *Farahmand* has many ties, given the binary nature of compositionality annotations.

[13] As all compounds in the dataset occur in the corpus, only strict evaluation results are reported.



(a) Overall best Spearman's ρ per DSM.

(b) Best Spearman's ρ per DSM and WORDFORM.

(c) Best Spearman's ρ per DSM and WINDOWSIZE.

(d) Best Spearman's ρ per DSM and DIMENSION.

Figure 3: Spearman's ρ for different DSM parameters on *FR-comp* dataset.

ing yields more accurate results. Moreover, a smaller WINDOWSIZE leads to better results for most models, as shown in Figure 3(c). But just as in English, all models except *glove* benefit from an increase in dimension, as shown in Figure 3(d).[14]

4.4 Discussion

When comparing DIMENSION across languages and datasets, larger values often bring better performance. Likewise, the *lemma* is usually the better WORDFORM. The recommended WINDOWSIZE depends on the model and language, but for the best models in all datasets, a window of 1 outperforms the others. This may be a consequence of the linear decay context weighting process, which assigns higher weights to closer words as window size increases. As an overall conclusion, in combination with a large dimension and a small window size, investing in preprocessing provides a good balance of a small vocabulary (of lemmas) and good accuracy. This is especially clear for a morphologically richer language like French, where lemmatization is homogeneously better for all models, even in *w2v*, for which surface forms were better for English.

In terms of models, the *w2v* models performed better than *PPMI* for *Reddy++*, both were in a tie for *Farahmand*, and *w2v* was outperformed by *PPMI-thresh* for French. The performance of *glove* for English was underwhelming, probably because we did not perform parameter tuning. As shown by (Salehi et al., 2015), *PPMI-TopK* is not an appropriate DSM for this task, as it does not model relevant cooccurrence very well.

The average Spearman's ρ for *Reddy* over all tested parameter configurations was 0.71 for both *w2v* models and 0.67 for *PPMI-SVD* and *PPMI-*

[14]For *w2v*, the same parameters used for English were adopted also for French. As a sanity check, we tested a range of negative sampling values [5, 15, 25, 35, 50], as well as subsampling rates for powers of 10 in [10^{-3} to 10^{-7}]. Variations in ρ are minor and do not show any clear trend.

[14]DSM parameters: WF: WORDFORM, D: DIMENSION, W: WINDOWSIZE. Results in parentheses for loose evaluation, using fallback.

Model & Parameters	Result
Reddy et al. (2011)	.71
Salehi et al. (2014)	.74
Salehi et al. (2015)	.80
Best *w2v* (*sg, WF=surface, D=750, W=1*)	.82 (.80)
Best *PPMI* (*thresh, WF=surface, D=750, W=8*)	.80 (.80)
Best *glove* (*WF=lemmapos, D=250, W=8*)	.76 (.76)

Table 1: Comparison of our best models with state-of-the-art ρ for *Reddy*.[14]

Model & Parameters	Result
Yazdani et al. (2015)	.49
Best *w2v* (*sg, WF=lemma, D=500, W=1*)	.51 (.47)
Best *PPMI* (*svd, WF=lemma, D=750, W=4*)	.52 (.45)
Best *glove* (*WF=lemma, D=500, W=8*)	.40 (.36)

Table 2: Comparison of our best models with state-of-the-art BF1 for *Farahmand*.[14]

thresh, and this was also observed for the other datasets. In short, both types of models can obtain good results. While *PPMI-thresh* is a simple, fast and inexpensive model to build, *w2v* has a free and push-button implementation, and requires less hyper-parameter tuning, as is it seems more robust to parameter variation. More generally, the best results obtained for *Reddy* and *Farahmand* are comparable and even outperform the state of the art, as shown in Tables 1 and 2, when strict evaluation is adopted (that is, when not using a fallback strategy for missing compounds).

5 Conclusions

In this paper we presented a multilingual, large-scale evaluation of DSMs for compound compositionality prediction. We have built 816 DSMs and performed 2,856 evaluations, examining the impact of corpus and context parameters, namely the level of corpus preprocessing, the context window size and the number of dimensions. Evaluation on 3 English datasets and a French one revealed that a large dimension is consistently better, and corpus preprocessing is usually beneficial. The choice of window size varies according to language and dataset, but a small window can often provide a good performance. The DSMs *w2v* and *PPMI* alternated in providing the best results. Moreover, the results obtained were comparable and even outperformed the state-of-the-art.

As future work, we plan to examine the use of a voting scheme for combining the output of complementary DSMs. Moreover, we also plan to combine additional sources of information for building the models, such as multilingual resources or translation data, to improve even further the compositionality prediction. We would also like to propose and evaluate more sophisticated compositionality functions that take into account the unbalanced contribution of individual words to the global meaning of a compound.

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 482520/2012-4, 312114/2015-0, "Simplificação Textual de Expressões Complexas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

References

Eneko Agirre, 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, Boulder, Colorado, USA*, pages 19–27. The Association for Computational Linguistics.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 89–96, Sapporo, Japan, Jul. ACL.

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

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Lang. Res. & Eval.*, 43(3):209–226, Sep.

Marco Baroni, 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, Maryland, June. Association for Computational Linguistics.

José Camacho-Collados, 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, China, July. Association for Computational Linguistics.

James Curran and Marc Moens. 2002. Scaling context space. In *Proc. of the 40th ACL (ACL 2002)*, pages 231–238, Philadelphia, PA, USA, Jul. ACL.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. DISSECT - DIStributional SEmantics composition toolkit. In *Proc. of the ACL 2013 System Demonstrations*, pages 31–36, Sofia, Bulgaria, Aug. ACL.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, Short Papers*, pages 92–97. The Association for Computer Linguistics.

Meghdad Farahmand, 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, Denver, Colorado, June. Association for Computational Linguistics.

Olivier Ferret. 2013. Identifying bad semantic neighbors for improving distributional thesauri. In *Proc. of the 51st ACL (Volume 1: Long Papers)*, pages 561–571, Sofia, Bulgaria, Aug. ACL.

Olivier Ferret. 2014. Compounds and distributional thesauri. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 2979–2984. European Language Resources Association (ELRA).

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In Ido Dagan and Dan Gildea, editors, *Proc. of the Ninth CoNLL (CoNLL-2005)*, pages 25–32, University of Michigan, MI, USA, Jun. ACL.

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

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. Semeval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In James Curran, editor, *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 491–498, Sidney, Australia, Jul. ACL.

Gabriella Lapesa 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.

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

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Proceedings*, pages 1188–1196. JMLR.org.

Omer Levy, 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.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proc. of the 37th ACL (ACL 1999)*, pages 317–324, College Park, MD, USA, Jun. ACL.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2016. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 73–80, Sapporo, Japan, Jul. ACL.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 746–751. The Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proc. of the 46th ACL: HLT (ACL-08: HLT)*, pages 236–244, Columbus, OH, USA, Jun. ACL.

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

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

Preslav Nakov. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Nat. Lang. Eng. Special Issue on Noun Compounds*, 19(3):291–330.

Sebastian Padó and Mirella Lapata. 2003. Constructing semantic space models from parsed corpora. In *Proc. of the 41st ACL (ACL 2003)*, pages 128–135, Sapporo, Japan, Jul. ACL.

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

Muntsa Padró, 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)*, Reykjavik, Iceland, May. European Language Resources Association.

Muntsa Padró, 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 (EMNLP 2014) - short papers*, Doha, Qatar, Oct.

Jeffrey Pennington, 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, Doha, Qatar, October. Association for Computational Linguistics.

Carlos Ramisch, 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 *Proc. of the 55th ACL (Volume 2: Short Papers)*, Berlin, Germany, Aug. ACL.

Siva Reddy, 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)*, Chiang Mai, Thailand, November.

Martin Riedl and Chris Biemann. 2012. Topictiling: A text segmentation algorithm based on LDA. In *Proc. of the ACL 2012 SRW*, pages 37–42, Jeju, Republic of Korea, Jul. ACL.

Stephen Roller, Sabine Schulte im Walde, and Silke Scheible. 2013. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proc. of the 9th Workshop on MWEs (MWE 2013)*, pages 32–41, Atlanta, GA, USA, Jun. ACL.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico, Feb. Springer.

Bahar Salehi, 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, April. Association for Computational Linguistics.

Bahar Salehi, 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, Colorado, May–June. Association for Computational Linguistics.

Stephen Tratz and Eduard Hovy. 2010. ISI: Automatic classification of relations between nominals using a maximum entropy classifier. In Katrin Erk and Carlo Strapparava, editors, *Proc. of the 5th SemEval (SemEval 2010)*, pages 222–225, Uppsala, Sweden, Jul. ACL.

Tim van de Cruys, Laura Rimell, Thierry Poibeau, and Anna Korhonen. 2012. Multi-way tensor factorization for unsupervised lexical acquisition. In *Proc. of the 24th COLING (COLING 2012)*, pages 2703–2720, Mumbai, India, Dec. The Coling 2012 Organizing Committee.

Majid Yazdani, 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, Lisbon, Portugal, September. Association for Computational Linguistics.