Modeling Semantic Compositionality of Croatian Multiword Expressions

Jan Šnajder and Petra Almić
University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
E-mail: jan.snajder@fer.hr, petra.almic@gmail.com

Keywords: multiword expressions, semantic composition, distributional semantics, Croatian language

Received: March 30, 2015

A distinguishing feature of many multiword expressions (MWEs) is their semantic non-compositionality. Determining the semantic compositionality of MWEs is important for many natural language processing tasks. We address the task of modeling semantic compositionality of Croatian MWEs. We adopt a composition-based approach within the distributional semantics framework. We build and evaluate models based on Latent Semantic Analysis and the recently proposed neural network-based Skip-gram model, and experiment with different composition functions. We show that the compositionality scores predicted by the Skip-gram additive models correlate well with human judgments (ρ=0.50). When framed as a classification task, the model achieves an accuracy of 0.64.

Povzetek: Razvita je metoda za dekompozicijo hrvaškega jezika.

1 Introduction

The peculiarity of multiword expressions (MWEs) has long been acknowledged in natural language processing (NLP). According to [29], MWEs can be defined as idiosyncratic interpretations that cross word boundaries (or spaces). Because of their unpredictable and idiosyncratic behavior, such expressions need to be listed in a lexicon and treated as a single unit ("word with spaces") [10, 5]. One dimension along which the MWEs can be analyzed is their semantic compositionality, sometimes referred to as semantic idiomaticity or semantic transparency. Semantic compositionality is the degree to which the features of the parts of an MWE combine to predict the features of the whole [4]. The meaning of a non-compositional MWE cannot be deduced from the meaning of its parts. In reality, MWEs span a continuum between completely compositional expressions (e.g., *world war*) to non-compositional ones [6]. A prime example of non-compositional MWEs are idioms, such as *kick the bucket (to die)* or *red tape (excessive rules and regulations).*

Being able to model the semantic compositionality of MWEs – and in turn determine whether a given MWE is semantically transparent or opaque – has been shown to be important for many NLP tasks, ranging from machine translation [8] and information retrieval [1] to word sense disambiguation [11]. It is thus not surprising that the task of automatically determining semantic compositionality has gained a lot of attention [15, 4, 7, 28, 18].

In this paper we address the task of modeling semantic compositionality of Croatian MWEs comprised of two words. We follow up on the work of [15] and [7] and adopt a compositionality-based approach: the basic idea is to compare the meaning of an MWE against the meaning of the composition of its parts. To model the meaning of the MWEs and its parts, we use distributional semantics, which represents the word's meaning based on the distribution on its contexts in a corpus, assuming that similar words tend to appear in similar contexts [13]. To determine the compositionality of an MWE, we compare its context distribution in the corpus to the context distribution approximated by the composition of its parts.

The contribution of our work is twofold. First, we build a dataset of Croatian MWE annotated with semantic compositionality scores. Secondly, we build and evaluate a set of semantic compositionality models based on Latent Semantic Analysis (LSA) [20] and the recently proposed neural network-based Skip-gram model [24]. Our results show that the compositionality scores predicted by additive compositional models correlate well with human-annotated scores, thereby confirming similar results for the English language. To the best of our knowledge, this is the first work to consider the modeling of semantic compositionality for the Croatian language.

The remainder of the paper is structured as follows. In the next section we give an overview of related work. We describe the creation of the dataset in Section 3 and the compositionality models in Section 4. In Section 5 we present evaluation results. Section 6 concludes the paper.

2 Related work

The approaches for modeling semantic compositionality can be broadly divided into two groups: knowledge-based approaches and corpus-based approaches. The former rely on linguistic resources (e.g., WordNet) to measure the semantic similarity between an MWE and its parts [16]. An obvious downside of knowledge-based approaches is that for most languages the linguistic resources are not available, while the construction of such resources is labor-intensive and expensive. In contrast, corpus-based approaches rely on statistical properties of MWEs and the constituting words, which can be readily extracted from corpora. E.g., [23] rely on the hypothesis that non-compositional MWEs tend to be syntactically more fixed than compositional MWEs, while [27] assumes that lexical association correlates with non-compositionality.

Related to the work presented in this paper are specifically the corpus-based approaches that make use of distributional semantic modeling of MWEs and their constituents. The pioneering work in this direction is that of [21], who used a statistical association measure to discriminate between compositional and non-compositional MWEs. Lin compared the mutual information of an MWE and of an expression obtained as a slight modification of the original MWE (e.g., *red tape* vs. *orange tape*). Although this method has not shown to work well, the idea that non-compositional expressions have a "different distributional characteristic" than similar compositional expressions paved a way for other distributional semantics based approaches. [5] used LSA to compare the similarity between an MWE and its head, and showed that there exists a correlation between the measured semantic similarity and compositionality. Along the same lines, [15] used LSA to compare the semantic vector of an MWE against the semantic vector of the composition of its constituents, obtained simply as the sum of the corresponding vectors.

To consolidate the research efforts, [7] organized a shared task on Distributional Semantics and Compositionality (DISCo), and provided datasets in English and German with human compositionality judgments. The task was shown to be hard and no clear winner emerged. However, the approaches based on distributional semantics seemed to outperform those based on statistical association measures. Following up on DISCo, [18] performed a systematic evaluation of various distributional semantic approaches to compositionality detection, and showed that LSA-based models perform quite well.

In this paper we adopt the methodology of [15] to compare the distribution of an MWE to the composition of its parts, but we experiment with different composition functions, proposed by [26]. To build the dataset, we adopt the methodology of [7].

3 Annotated dataset

The starting point of our work is a dataset of representative Croatian MWEs annotated with human compositionality judgments. In building this dataset, we adopted the approach of [7], but depart from it in some key aspects that we discuss below. As a source of data, we used the 1.2 billion words corpus fHrWaC[1] [30], a filtered version of the Croatian web corpus *hrWaC* [22]. The corpus has been tokenized, lemmatized, POS tagged, and dependency parsed using the HunPos tagger and CST lemmatizer for Croatian [3], and the MSTParser for Croatian [2], respectively. We next describe the construction of the dataset.[2]

3.1 MWE extraction

Following the work of [7], we restricted ourselves to the following three MWE types:

- **AN**: an adjective modifying a noun, e.g., *žuti karton* (*yellow card*);
- **SV**: a verb with a noun in the subject position, e.g., *podatak govori* (*data says*);
- **VO**: a verb with a noun in the object position, e.g., *popiti kavu* (*drink coffee*).

We extracted all dependency bigrams (i.e., possibly non-contiguous bigrams) from the corpus that match one of these three types and sorted them by frequency in descending order.[3] Going from the top of list, we (the two authors) manually annotated the MWEs (i.e., for each bigram we annotated whether it constitutes an MWE) and additionally pre-annotated each MWE as either compositional (C) or non-compositional (NC). We next selected the bigrams on which both annotators agreed, and then balanced the set so that it contains an equal number of compositional and non-compositional MWEs. The so-obtained dataset does not reflect the true distribution of MWEs, as the compositional MWEs are much more frequent in the corpus. However, balancing the dataset is justified because our focus is on discriminating between the compositional and non-compositional MWEs. The final dataset contains 100 compositional and 100 non-compositional MWEs (125 AN, 10 SV, and 65 VO expressions). Note that the C/NC annotation is preliminary; each of the 200 MWEs has subsequently been annotated with compositionality scores by multiple human annotators other than the authors (cf. Section 3.3).

3.2 Levels of compositionality

During MWE pre-annotation, we identified various flavors of compositionality. For example, a *yellow card* really is a

[1] http://takelab.fer.hr/data/fhrwac/

[2] The dataset is available under the Creative Commons BY-SA license from http://takelab.fer.hr/cromwesc

[3] By considering only the most frequent MWEs, we limit ourselves to MWEs with most reliable distributional representations.

yellow card, but its predominant sense is a figurative one (a warning indication). In contrast, *gray economy* is indeed a type of economy, but *gray* does not have a literate meaning. Further along these lines, *chain* in a *chain store* is not a chain in its predominant sense. One can argue that all these expressions are non-compositional to a certain extent. In an attempt to give an operational account of the different levels of non-compositionality, we propose the following typology:

NC3: Expressions that are completely non-compositional, i.e., the meaning of constituents cannot be combined to give the meaning of the expression. E.g., *žuti karton (yellow card)* and *preliti čašu* (literal meaning: *spill over the cup*; figurative meaning: *the last straw*), *trljati ruke (to rub ones hands)*;

NC2: Partially compositional expressions, i.e., the meaning of one but not both constituents is opaque, e.g., *siva ekonomija (gray economy)*, *bilježiti rast (to record a growth)*, *morski pas* (literal meaning: *sea dog*; compositional meaning: *a shark*);

NC1: The expressions that are non-compositional if we consider only the predominant senses of one or both of its constituents. For example, if we consider the predominant sense of *chain* to be a series of metal rings, then a *mountain chain* is a non-compositional expression.[4]

Note that our typology is motivated by practical rather than theoretical concerns. When concerned with automatic compositionality detection, we expect type NC3 to be more easily determinable than type NC1. However, from a theoretical perspective, the proposed typology is oversimplified and we make no attempt here to relate it to the different types of figures of speech studied in linguistics (e.g., metaphors, metonyms, synegdochs, etc.).

3.3 Annotation

[7] used the crowdsourcing service Amazon Turk to annotate their dataset. For every expression, they provided five different context sentences. For each in-context MWE, they asked the turkers to annotate how literal the MWE is, on a scale from 1 (non-compositional) to 10 (compositional). Because the set of annotators differs across MWEs, they were not able to estimate the inter-annotator agreement. However, they argued that the judgments for the expressions should be reliable because they were averaged over several sentences and several annotators. As the final compositionality scores, they computed the mean score for each MWE.

We departed from the above-described setup for two reasons. Methodologically, we argue that annotating MWEs across contexts is inappropriate for the task of type-based semantic compositionality detection, which is what we are addressing here. The reason is that this setup ignores the fact that MWEs may have different meanings (compositional and non-compositional ones) depending on the context, thus averaging across the contexts will lump together the various senses.[5]On a practical side, in-context annotation is more expensive and would require more resources (we feel that annotating five sentences per MWE would not suffice to reliably capture the sense variability of MWEs). For these reasons, we chose not to annotate MWEs across different contexts.

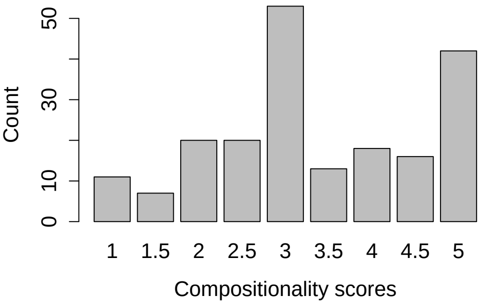


Figure 1: Histogram of MWE compositionality scores.

Our annotation setup was as follows. A total of 24 volunteers (mostly students) participated in the annotation. To reduce the workload, we divided the 200 MWEs into four groups (A, B, C, D) and randomly assigned one group to each annotator. Thus, each MWE was annotated by six annotators. To be able to computer the inter-annotator agreement, we ensured that there is a 10% overlap among all four groups (20 expressions that were annotated by all 24 annotators). We asked our annotators to judge how literal each MWE is on the scale from 1 (non-compositional) to 5 (compositional). For each MWE, we provided one context sentence that instantiates its non-compositional meaning (for non-compositional MWEs) or typical compositional meaning (for compositional MWEs). We did this to ensure that annotators consider the same sense of an MWE, so that the judgments would not diverge because of sense mismatches.

We computed the final compositionality score for each MWE as the median of its compositionality scores. Fig. 1 shows the scores histogram, while Table 1 shows some examples from the annotated dataset.

3.4 Annotation analysis

Table 2 shows the inter-annotator agreement in terms of the Krippendorff's alpha coefficient [17] for each of the groups as well as the overlapping part of the dataset. We

[4] We are aware that the notion of a predominant sense is a problematic one. Many of the NC1 MWEs in our dataset are in fact borderline cases between NC1 and C classes.

[5] In-context MWE compositionality annotation would be adequate for the task of token-based semantic compositionality detection (detection of semantic compositionality of a MWE instance in context). Curiously enough, [7] were also addressing the type-based ask, but used in-context annotation.

MWE	Type	Score
maslinovo ulje (olive oil)	AN	5
krvni tlak (blood pressure)	AN	5
telefonska linije (telephone line)	AN	4
pružiti pomoć (to offer help)	VO	4
kućni ljubimac (a pet)	AN	3.5
crno tržište (black market)	AN	3
voditi brigu (to worry)	VO	3
ostaviti dojam (to leave an impression)	VO	2.5
zeleno svjetlo (green light)	AN	1
hladni rat (cold war)	AN	1

Table 1: Examples from the annotated dataset.

Sample	AN+SV+VO	AN	SV+VO
Group A	0.587	0.620	0.535
Group B	0.506	0.510	0.478
Group C	0.490	0.544	0.337
Group D	0.586	0.505	0.648
Overlap (10%)	0.456	0.452	0.439

Table 2: Inter-annotator agreement (Krippendorff's α).

consider the agreement to be moderate and indicative of the high subjectivity of the task. The agreement on the verb expressions is somewhat lower in comparison to adjective-noun expressions. In Table 3 we present some example MWEs from the dataset where the annotators achieved a high level of agreement (zero standard deviation) and a low level of agreement (st. dev. larger than 1.3).

As an indication of the ceiling performance, we computed the correlation between every annotator's scores and the median scores. The average Spearman's correlation coefficient over 24 annotators is 0.77.

High agreement ($\sigma = 0$)	
igrati nogomet (play soccer)	5.0
služiti kaznu (serve sentence)	3.0
financijska pomoć (financial aid)	5.0
pjevati pjesmu (sing song)	5.0
nemati sumnje (have no doubt)	5.0
Low agreement ($\sigma > 1.3$)	
zabilježiti rast (record growth)	4.5
žuti karton (yellow card)	3.0
prvi korak (first step)	3.0
telefonska linija (phone line)	4.0
crveni karton (red card)	4.5

Table 3: Examples of MWEs and median compositionality scores with high and low inter-annotator agreement.

3.5 Test set

To optimize and experiment with the various parameters, we randomly split our dataset into the train and test set, each consisting of 100 MWEs. The breakdown of MWEs by type in the test set is as follows: 65 AN, 30 VO, 5 VS. Furthermore, we (the two authors) annotated the level of compositionality (cf. Section 3.2) for each MWE from the test set and resolved the disagreements by consensus. Our primarily motivation for this was to be able to investigate how the level of non-compositionality influences the performance of the model. The breakdown of compositionality levels in the test set is as follows: 48 C, 31 NC1, 7 NC2, 14 NC3.

4 Compositionality models

To build our models, we use the fHrWaC corpus, the same corpus we used to build the dataset. To determine the semantic compositionality of a MWE, we carry out the following three steps: (1) model the meaning of the constituent words, (2) model the composition of the meaning, and (3) compare these meanings.

4.1 Modeling word meaning

To model the meaning of constituent words, we use two distributional semantics models. First is the well-known Latent Semantic Analysis [20] model. LSA has shown to perform quite good in the task of semantic compositionality detection [15, 18], and performed very good in the task of identifying synonyms in the Croatian language [14]. We defined the context as a ±5 word window around the word, or, in the case of the MWEs, a ±5 word window around both constituents. For the constituent words, we only consider the contexts in which they appear alone, i.e., not as a part of any MWE from our dataset. Motivation behind this is to emphasize the independent contribution of the constituents in an expression, as proposed by [15]. As context elements (the columns of the LSA matrix), we use the 10k most frequent lemmas from the corpus (excluding stop words). As target elements (the rows of the matrix), we use the MWEs and their constituting words, as well as the 5k most frequent lemmas from the corpus. For weighing the word-context associations, we experimented with two functions: log-entropy [19] and Local Mutual Information (LMI) [9]. Since log-entropy gave consistently better results, we use only log-entropy in subsequent experiments. We use singular value decomposition to reduce the dimensionality of the matrix from 10000 to 100 dimensions per target.

The second model we experiment with is the neural network-based Skip-gram model recently proposed by [24]. Skip-gram model produces low-dimensional, real-valued vector representations of words (also known as word embeddings) by learning to predict the context of each input word. Skip-gram model has been shown to be very effective for predicting the semantic similarity of words and has excelled in the synonymy detection and relation modeling task for Croatian [31], outperforming LSA by a large margin. Furthermore, it has been shown that with Skip-gram model the semantic composition of short phrases can be modeled quite effectively via simple vector addition [25], which makes it well-suited for our task. We build 100-dimensional vector representations of MWEs and their constituting words using the `word2vec` tool,[6] with default parameters (number of negative examples set to 5, no hierarchical softmax, maximum skip length of 5), but without a frequency threshold.

4.2 Modeling composed meaning

The second step was to model the composition of the word meanings. [26] introduced a number of composition models (additive, weighted additive, multiplicative, tensor product, and dilation), which they evaluated on a phrase similarity task (e.g. *vast amount* vs. *large quantity*). In this work, we experiment with additive ($\vec{z} = \vec{x} + \vec{y}$), weighted additive ($\vec{z} = \alpha\vec{x} + \beta\vec{y}$), and the multiplicative model ($\vec{z} = \vec{x} \odot \vec{y}$), where z stands for the composed vector and $\vec{x}$ and $\vec{y}$ stand for vectors of its constituent words.

We experiment with two weighted additive models. In the first one (model Opt), similarly to [26], we optimize the weights on the train set to maximize the correlation with human-assigned scores. The weights are optimized globally and they are identical for every MWE. In the second model (model Dyn), we calculate the weights dynamically, separately for each MWE, as proposed by [28]. The two weights, α and β, are defined as

$$\alpha = \frac{\cos(\overrightarrow{xy}, \vec{x})}{\cos(\overrightarrow{xy}, \vec{x}) + \cos(\overrightarrow{xy}, \vec{y})}, \quad \beta = 1 - \alpha \tag{1}$$

where $\overrightarrow{xy}$ is the MWE vector. The intuition behind this method is that more importance should be given to the constituent that is semantically more similar to the whole MWE, i.e., the constituent whose vector is closer, in terms of the cosine similarity, to the vector of the MWE. For example, in the expression *gray economy*, more importance should given to the word *economy* than the word *gray*.

4.3 Compositionality prediction

Finally, in the third step, we use the cosine similarity measure to compare the vector of the MWE and the vector of its composition-derived meaning. We expected that for the compositional MWEs these two meaning vectors will be similar, i.e., cosine similarity will be closer to 1, while for non-compositional it will be closer to 0. Thus, the model simply predicts the semantic compositionality score as the cosine between the MWE vector and the composed MWE vector.

Additionally, we consider the linear combination model proposed by [28]. Instead of relying on a single compositionality prediction, this model uses the collective evidences from several models. More precisely, the model predicts the semantic compositionality score as a linear combination of the prediction of the additive model, the multiplicative model, and the two individual constituents model:

$$\begin{aligned}\lambda = a_0 + a_1 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x+y}) + a_2 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x \odot y}) \\ + a_3 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x}) + a_4 \cdot \cos(\overrightarrow{xy}, \overrightarrow{y})\end{aligned} \tag{2}$$

We optimized the parameters a_0–a_4 using least squares regression on the train set.

5 Evaluation

The task of determining semantic compositionality can be framed as a regression problem (prediction of compositionality scores) or a classification problem (compositionality vs. non-compositionality). We consider both settings.

5.1 Predicting compositionality scores

In Table 4 we show the correlation (Spearman's ρ) between model-predicted and human-assigned compositionality scores on the test set. Generally, additive models outperform the other considered composition models. This is in contrast to the conclusions of [26], but in accordance with the results of [12] and [18]. For both LSA and Skip-gram, correlation for verbal MWEs is much worse then for adjective-noun MWEs. This is expected, as it has been observed that the semantics of verbs is not fully covered by distributional spaces (cf. [31, 14]) Skip-gram model outperforms LSA, confirming the findings from other tasks [31]. The difference in performance is most prominent for verbal (SV+VE) MWEs. Overall, the best performing models are the Skip-gram additive and linear combination models ($\rho = 0.50$). Specifically, if one considers the AN and SV+VO subsets, Skip-gram linear combination model emerges as the clear winner, suggesting that combining the evidence from multiple models is beneficial. We consider the obtained correlation of 0.50 to be satisfactory, considering that the average correlation of human annotators is 0.77. Our results seem to be comparable to those of [7, 18] obtained for English.

5.2 Compositionality classification

For the compositionality classification task, we converted the compositionality scores to binary labels. To this end, we analyzed the distribution of the scores in the dataset (Fig. 1). The distribution is bimodal, so we chose to set the cut-off after the first mode: MWEs with a score in the $[1, 3]$ range are labeled as non-compositional (NC), while

[6] `http://code.google.com/p/word2vec/`

	LSA			Skip-gram		
Composition model	AN+SV+VO	AN	SV+VO	AN+SV+VO	AN	SV+VO
Multiplicative	-0.19	-0.20	-0.18	0.01	-0.14	0.38
Simple additive	0.45	0.54	$\mathbf{0.35}$	$\mathbf{0.50}$	0.55	0.40
Weighted additive (Opt)	0.46	0.56	0.28	$\mathbf{0.50}$	0.55	0.40
Weighted additive (Dyn)	0.46	$\mathbf{0.57}$	0.26	$\mathbf{0.50}$	0.55	0.40
First constituent	0.41	0.50	0.19	0.37	0.43	0.21
Second constituent	0.28	0.31	0.31	0.41	0.49	0.36
Linear combination (λ)	$\mathbf{0.48}$	0.56	0.34	$\mathbf{0.50}$	$\mathbf{0.58}$	$\mathbf{0.47}$

Table 4: Spearman's correlation coefficient on the test set for LSA and Skip-gram model and different composition functions.

	AN+SV+VO	AN	SV+VO
Precision	0.56	0.63	0.44
Recall	0.82	0.84	0.92
F1-score	0.67	0.72	0.60
Accuracy	0.64	0.69	0.54

Table 5: Classification results for the Skip-gram linear combination model.

those with a score in the $\langle 3, 5]$ range are labeled as compositional (C). This gave us 44 compositional (C) and 56 non-compositional MWEs in the test set. We consider only the best-performing model from the previous experiment (the Skip-gram linear combination model). The model predicts C (positive class) if the prediction of the linear combination model defined by (2) is above a certain threshold, otherwise it predicts NC (negative class). We set the threshold to $t = 3.11$, obtained by optimizing the F1-score on the train set. The results are shown in Table 5. The overall classification accuracy is 0.64. The accuracy is higher for adjective-noun MWEs (0.72) than for verbal MWEs (0.54), which is in line with the results from the previous experiment. Precision is substantially lower than recall (0.56 vs. 0.82), indicating that the model more often predicts compositionality for a non-compositional MWE than the other way around, i.e., the predictions for non-compositional MWEs are often higher than they ought to be. Our model outperforms the accuracy of a majority class (NC) baseline, which is 0.56, but not the F1-score, which is 0.72.

The classification task is similar to the one considered by [15]. In their experiment, they achieved an F1-score of 0.48, but they only considered the additive model.

5.3 Result analysis

In this section we give some insights about the model performance. Results show moderate level of correlation, so we are interested in investigating on what MWEs the model fails. We are also interested in relating the model performance to the levels of compositionality introduced in Section 3.2 and the inter-annotator agreement levels.

In Table 6 we list the MWEs on which the Skip-gram linear combination model performs the worst. We define the error as an absolute difference between z-scored model-predicted and human-annotated compositionality scores. The results suggest that most errors occur on compositional expressions (C).

To explore this hypothesis a bit further, we divide our test set into the subsets based on the compositionality levels and analyze compositionality scores and correlation on these subsets. Fig. 2a shows z-scored human-assigned compositionality scores and z-scored model predictions across different compositionality levels. Both human-assigned and predicted scores increase with the level of compositionality, however the model tends to assign lower scores to compositional MWEs (C) and higher scores to completely non-compositional MWEs (NC3); the latter is in line with the classification results (low precision).

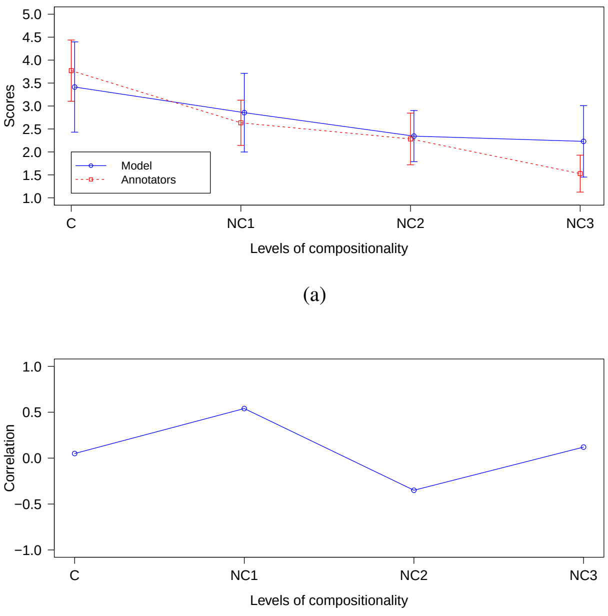
Fig. 2b shows the correlation between human-assigned and model-predicted scores across different compositionality levels. The plot shows that the model performs best on non-compositional MWEs of type NC1 (non-compositional in the predominant reading) and much worse on other non-compositional MWEs as well as compositional MWEs.[7] While this is in line with the previous analysis (model underestimates C scores and overestimates NC3 scores), it remains unclear why the model performs better on NC1. The results seems counterintuitive, as one would expect NC3 (completely non-compositional) MWEs to be more easily detectable than NC1 (non-compositional in the predominant reading).

A more systematic analysis, which is out of the scope of this paper, would be required to determine the underlying causes. One of the possible reasons could be the low quality of vector representations for some (rare) words. The low quality of the individual words propagates to the low quality of compositional representations, which in turn makes the composed vector too dissimilar to the MWE vector. A further problem might stem from the polysemy, another weakness of distributional semantic models.

[7] Note that NC2 and NC3 have few data instances, hence their correlation results should be taken with caution.

MWE	Type	Level	Score	Prediction	Error
oglasna ploča (announcement board)	AN	C	4.5	1.69	3.33
organizacijski odbor (organizing committee)	AN	C	5	2.76	2.26
motorno vozilo (motor vehicle)	AN	C	5	2.79	2.22
nemati sumnje (have no doubt)	VO	C	5	2.82	2.18
optužnica tereti (charged with)	SV	C	2.5	4.35	1.96
životno djelo (lifework)	AN	C	3	1.81	1.94
novi val (new wave)	AN	NC3	1	3.33	1.78

Table 6: MWEs on which the Skip-gram linear model performs the worst (human-assigned scores and model predictions are not scaled, while the error is between z-scored values).



(a)

(b)

Figure 2: Analysis across four compositionality levels: (a) z-scored human-assigned scores and model predictions, (b) correlations.

6 Conclusion

In this paper we modeled of semantic compositionality of Croatian multiword expressions (MWEs) using composition-based distributional semantics. We built a small dataset of Croatian MWEs, manually annotated with semantic compositionality scores. To represent the meaning of the MWEs and their constituents, we build two kinds of models (LSA and Skip-gram), and experimented with additive and multiplicative composition functions. The best-performing model combines the predictions of the additive and the multiplicative models, and achieves a correlation of 0.50 and a classification accuracy of 0.64. The model tends to underestimate scores for compositional MWEs and overestimate scores for non-compositional MWEs. Surprisingly, the model works best on MWEs that that are non-compositional if one considers the predominant reading of MWE constituents.

Future work might address a more systematic analysis. This implies annotating a larger dataset (possibly one that is unbalanced and hence more realistic) and accounting for confounding factors such as MWE frequency and ambiguity.

References

[1] Otavio Costa Acosta, Aline Villavicencio, and Viviane P. Moreira. Identification and treatment of multiword expressions applied to information retrieval. In *Proc. of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, pages 101–109. ACL, 2011.

[2] Željko Agić and Danijela Merkler. Three syntactic formalisms for data-driven dependency parsing of Croatian. In *Text, Speech, and Dialogue*, pages 560–567. Springer, 2013.

[3] Željko Agić, Nikola Ljubešić, and Danijela Merkler. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proc. of ACL*, 2013.

[4] Timothy Baldwin. Compositionality and multiword expressions: Six of one, half a dozen of the other. In

Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, 2006.

[5] Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. An empirical model of multiword expression decomposability. In *Proc. of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. ACL, 2003.

[6] Colin Bannard, Timothy Baldwin, and Alex Lascarides. A statistical approach to the semantics of verb-particles. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 65–72. ACL, 2003. doi: 10.3115/1119282.1119291.

[7] Chris Biemann and Eugenie Giesbrecht. Distributional semantics and compositionality 2011: Shared task description and results. In *Proc. of the Workshop on Distributional Semantics and Compositionality*, pages 21–28. ACL, 2011.

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

[9] Stefan Evert. *The statistics of word cooccurrences*. PhD thesis, Dissertation, Stuttgart University, 2005.

[10] Stefan Evert. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2:223–233, 2008.

[11] Mark Alan Finlayson and Nidhi Kulkarni. Detecting multi-word expressions improves word sense disambiguation. In *Proc. of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, pages 20–24, 2011.

[12] Emiliano Guevara. Computing semantic compositionality in distributional semantics. In *Proc. of the Ninth International Conference on Computational Semantics*, pages 135–144. ACL, 2011.

[13] Zellig Harris. Distributional structure. *Word*, 10(23): 146–162, 1954.

[14] Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. Distributional semantics approach to detecting synonyms in Croatian language. *Information Society*, pages 111–116, 2012.

[15] Graham Katz and Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proc. of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. ACL, 2006.

[16] Su Nam Kim and Timothy Baldwin. Automatic identification of English verb particle constructions using linguistic features. In *Proc. of the Third ACL-SIGSEM Workshop on Prepositions*, pages 65–72. ACL, 2006.

[17] Klaus Krippendorff. Reliability in content analysis. *Human Communication Research*, 30(3):411–433, 2004.

[18] Lubomír Krčmář, Karel Ježek, and Pavel Pecina. Determining compositionality of word expresssions using various word space models and methods. In *Proc. of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 64–73. ACL, 2013.

[19] Landauer. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007. ISBN 0805854185.

[20] Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240, 1997.

[21] Dekang Lin. Automatic identification of non-compositional phrases. In *Proc. of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324. ACL, 1999.

[22] Nikola Ljubešić and Tomaž Erjavec. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, pages 395–402. Springer, 2011.

[23] Diana McCarthy, Sriram Venkatapathy, and Aravind K Joshi. Detecting compositionality of verb-object combinations using selectional preferences. In *EMNLP-CoNLL*, pages 369–379, 2007.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proc. of ICLR*, Scottsdale, AZ, USA, 2013a.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013b.

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

[27] Ted Pedersen. Identifying collocations to measure compositionality: shared task system description. In *Proc. of the Workshop on Distributional Semantics and Compositionality*, pages 33–37. ACL, 2011.

[28] Siva Reddy, Diana McCarthy, Suresh Manandhar, and Spandana Gella. Exemplar-based word-space model for compositionality detection: Shared task system description. In *Proc. of the Workshop on Distributional Semantics and Compositionality*, pages 54–60. ACL, 2011.

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

[30] Jan Šnajder, Sebastian Padó, and Željko Agić. Building and evaluating a distributional memory for Croatian. In *In Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789. ACL, 2013.

[31] Leo Zuanovic, Mladen Karan, and Jan Šnajder. Experiments with neural word embeddings for croatian. In *Proc. of the Ninth Language Technologies Conference, Information Society (IS-JT 2014)*, pages 69–72, 2014.