# Determining the Semantic Compositionality of Croatian Multiword Expressions

**Petra Almić and Jan Šnajder**

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
petra.almic@gmail.com, jan.snajder@fer.hr

### Abstract

A distinguishing feature of many multiword expressions (MWEs) is their semantic non-compositionality. Being able to automatically determine the semantic (non-)compositionality of MWEs is important for many natural language processing tasks. We address the task of determining the semantic compositionality of Croatian MWEs. We adopt a composition-based approach within the distributional semantics framework. We build a small dataset of Croatian MWE with human-annotated semantic compositionality scores. We build and evaluate a model for predicting the semantic compositionality based on Latent Semantic Analysis. The predicted scores correlate well with human judgments ($\rho$=0.48). When compositionality detection is treated as a classification task, the model achieves an F1-score of 0.65.

### Določanje semantične kompozicionalnosti hrvaških večbesednih enot

Pomembna lastnost številnih večbesednih enot je njihova semantična nekompozicionalnost. Zmožnost avtomatskega določevanja takšne (ne)kompozicionalnosti je pomembna za številne naloge pri obdelavi naravnega jezika. V prispevku obravnavamo določanje semantične kompozicionalnosti hrvaških večbesednih enot. Uporabimo metodo, ki temelji na kompozicionalnosti v okviru distribucijske semantike. Zgradimo majhno podatkovno množico hrvaških večbesednih enot z ročno določenimi vrednostmi njihove semantične kompozicionalnosti. Zgradimo in evalviramo model za napovedovanje semantične kompozicionalnosti, ki temelji na latentni semantični analizi. Napovedane vrednosti dobro korelirajo s človeškimi ocenami ($\rho$ = 0,48). Če detektiranje kompozicionalnosti obravnavamo kot klasifikacijsko nalogo, doseže model za mero F1 vrednost 0,65.

## 1. Introduction

The peculiarity of multiword expressions (MWEs) has long been acknowledged in natural language processing (NLP). According to Sag et al. (2002), 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") (Evert, 2008; Baldwin et al., 2003). 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 (Baldwin, 2006). 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 (Bannard et al., 2003). 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 determine the semantic compositionality of MWEs has been shown to be important for many NLP tasks, ranging from machine translation (Carpuat and Diab, 2010) and information retrieval (Acosta et al., 2011) to word sense disambiguation (Finlayson and Kulkarni, 2011). It is thus not surprising that the task of automatically determining semantic compositionality has gained a lot of attention (Katz and Giesbrecht, 2006; Baldwin, 2006; Biemann and Giesbrecht, 2011; Reddy et al., 2011; Krčmář et al., 2013).

In this paper we address the task of automatically determining the semantic compositionality of Croatian MWEs comprised of two words. We follow up on the work of Katz and Giesbrecht (2006) and Biemann and Giesbrecht (2011) 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 (Harris, 1954). To determine the compositionality of an MWE, we compare its context distribution in a corpus to the context distribution approximated by the composition of its parts.

The contribution of our work is twofold. Firstly, we build a dataset of Croatian MWE annotated with semantic compositionality scores. Second, we build and evaluate a semantic compositionality model based on Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997). Our results show that the compositionality scores produced by the model 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 semantic compositionality detection for the Croatian language.

## 2. Related work

The approaches to determining the 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 (Kim and

Baldwin, 2006). The obvious downside of knowledge-based approaches is that the linguistic resources are unavailable for the most languages and that acquiring them is 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., McCarthy et al. (2007) rely on the hypothesis that non-compositional MWEs tend to be syntactically more fixed than compositional MWEs, while Pedersen (2011) assumes that lexical association correlates with non-compositionality.

Related to the work presented in this paper are the corpus-based approaches that rely on the distributional semantic modeling of MWEs and their constituents. The pioneering work in this direction is that of Lin (1999), 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 be successful, the idea that non-compositional expressions have a "different distributional characteristic" than similar compositional expressions paved a way for other distributional semantics based approaches. Baldwin et al. (2003) 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, Katz and Giesbrecht (2006) 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, Biemann and Giesbrecht (2011) organized a shared task on semantic compositionality detection, 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. Shortly thereafter, Krčmář et al. (2013) 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 Katz and Giesbrecht (2006) to compare the distribution of an MWE to the composition of its parts, but we experiment with different composition functions, proposed by Mitchell and Lapata (2010). To build the dataset, we adopt the methodology of Biemann and Giesbrecht (2011).

## 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 Biemann and Giesbrecht (2011), 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] (Šnajder et al., 2013), a filtered version of the Croatian web corpus *hrWaC* (Ljubešić and Erjavec, 2011). The corpus has been tokenized, lemmatized, POS tagged, and dependency parsed using the the HunPos tagger and the CST lemmatizer for Croatian (Agić et al., 2013), and the MSTParser for Croatian (Agić and Merkler, 2013), respectively. We next describe the construction of the dataset.[2]

### 3.1. MWE extraction

Following the work of Biemann and Giesbrecht (2011), 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 and additionally pre-annotated each as 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, as our focus is on discriminating between the compositional and non-compositional MWEs, balancing the dataset is justified in this case. 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 the process of the candidate selection, we identified various flavors of compositionality. For example, a *yellow card* really is a yellow card, but it has an additional (and a dominant one) figurative meaning (a warning indication). In contrast, *gray economy* is indeed a type of economy, but *gray* does not stand for a color here. Further along these lines, *chain* in a *chain store* is not a chain in its dominant 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:[4]

[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.

[4] Note that our typology is motivated by practical rather than theoretical concerns. In the realm of automatic compositionality detection, type NC3 is arguably more easily determinable than type NC1. From a theoretical perspective, the proposed typology is oversimplified and we make no attempts here to relate it to the different types of figures of speech studied in linguistics (e.g., metaphors, metonyms, synegdochs, etc.).

**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 dominant senses of one or both of its constituents. For example, if we consider a *chain* only as a series of metal rings, then a *mountain chain* is a non-compositional expression.[5]

We (the two authors) annotated the 200 MWEs according to the above types 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.

### 3.3. Annotation

Biemann and Giesbrecht (2011) 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 of this setup, they were not able to estimate the inter-annotator agreement, but 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 semantic compositionality detection of the sort we are addressing here. The reason is that it 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. 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.

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 a 10% overlap among all four groups (20 expressions that were annotated by all 24 annotators).

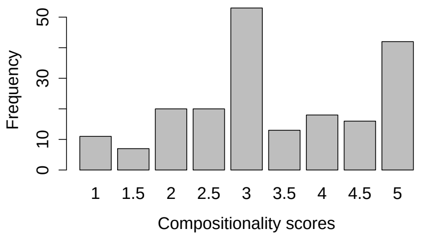

Figure 1: Histogram of MWE compositionality scores.

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

Table 1: Examples from the annotated dataset.

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 (Krippendorff, 2004) for each of the groups as well as the overlapping part of the dataset. We 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. $>$ 1.3).

| 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 $\alpha$).

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

| High agreement | Low agreement |
|---|---|
| *igrati nogomet (play soccer)* | *zabilježiti rast (record growth)* |
| *služiti kaznu (serve sentence)* | *žuti karton (yellow card)* |
| *financijska pomoć (financ. aid)* | *prvi korak (first step)* |
| *pjevati pjesmu (sing song)* | *telefonska linija (phone line)* |
| *nemati sumnje (have no doubt)* | *crveni karton (red card)* |

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

To be able to compare the performance of the models against human judgments as 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.

## 4. Compositionality model

To build our model, we use the fHrWaC corpus, the same corpus we used to build the dataset. To optimize and experiment with the various parameters, we randomly split our dataset into the train and test set, each consisting of 100 MWEs. 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.

**Modeling word meaning.** To model the meaning of constituent words, we use the Latent Semantic Analysis (Landauer and Dumais, 1997). LSA has shown to perform quite good in the task of semantic compositionality detection (Katz and Giesbrecht, 2006; Krčmář et al., 2013). Furthermore, LSA models excelled in the task of identifying synonyms in the Croatian language (Karan et al., 2012). We defined the context as a $\pm 5$ word window around the word, or, in the case of the MWEs, a $\pm 5$ word window around both constituents. For the constituent words, we only considered 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 Katz and Giesbrecht (2006). 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 used 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 (Landauer, 2007) and Local Mutual Information (LMI) (Evert, 2005). We used singular value decomposition to reduce the dimensionality of the matrix from 10000 to 100 dimensions per target.

**Modeling composed meaning.** The second step was to model the composition of the word meanings. Mitchell and Lapata (2010) 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 Mitchell and Lapata (2010), we optimized the weights on the train set to maximize the correlation with human scores. The weights are optimized globally and they are identical for every MWE. In the second one (model Dyn), we calculated the weights dynamically, separately for each MWE, as proposed by Reddy et al. (2011). The two weights, $\alpha$ and $\beta$, are defined as

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

where $\vec{xy}$ 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*.

In addition, we experiment with a linear combination of the additive model, the multiplicative model, and the two individual constituents model (Reddy et al., 2011):

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

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

**Meaning comparison.** Finally, in the third step, we use the cosine similarity measure to compare the vector-represented meaning 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.

## 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 $\rho$) between model-predicted and human-annotated compositionality scores on the test set. Even though we experimented with two weighting functions, here we present only the results for log-entropy because LMI gave consistently worse results. Additive models outperform the multiplicative model. This is in contrast to the conclusions of Mitchell and Lapata (2010), but in accordance with the results of Guevara (2011) and Krčmář et al. (2013). Also, it is noticeable that the AN expressions have better correlation than verb expressions, which goes along the fact that the former had a higher inter-annotator agreement. Best performing model is the linear combination, which suggests that combining the evidence from multiple models is beneficial. Overall, results seem to be comparable to the results in (Biemann and Giesbrecht, 2011; Krčmář et al., 2013) obtained for English.

| Model | AN+SV+VO | AN | SV+VO |
|---|---|---|---|
| Multiplicative | −0.19 | −0.20 | −0.18 |
| Simple additive | 0.45 | 0.54 | 0.35 |
| Weighted additive (Opt) | 0.46 | 0.56 | 0.28 |
| Weighted additive (Dyn) | 0.46 | 0.57 | 0.26 |
| First constituent | 0.41 | 0.50 | 0.19 |
| Second constituent | 0.28 | 0.31 | 0.31 |
| Linear combination ($\lambda$) | **0.48** | **0.56** | **0.34** |
| Annotators | 0.77 | 0.77 | 0.74 |

Table 4: Correlation results on the test set.

| | AN+SV+VO | AN | SV+VO |
|---|---|---|---|
| Precision | 0.58 | 0.74 | 0.43 |
| Recall | 0.73 | 0.65 | 0.77 |
| Accuracy | 0.65 | 0.72 | 0.54 |
| F1-score | 0.65 | 0.69 | 0.56 |

Table 5: Classification results on the test set.

| MWE | Prediction | Error | Level |
|---|---|---|---|
| *nemati sumnje* | 2.48 | 2.85 | C |
| *organizacijski odbor* | 2.66 | 2.56 | C |
| *dati život* | 2.16 | 2.55 | NC3 |
| *optužnica teretiti* | 4.51 | 2.51 | C |
| *spasiti život* | 2.85 | 2.25 | C |
| *uroditi plodom* | 3.85 | 2.24 | NC1 |
| *izvršna vlast* | 2.61 | 2.23 | C |

Table 6: MWEs on which the model performs the worst.

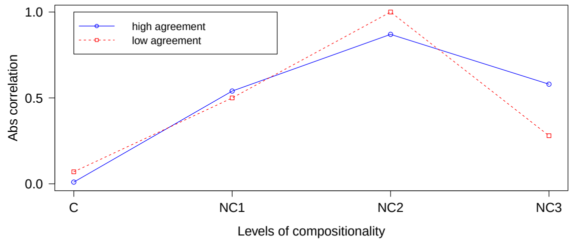

Figure 2: Correlation on the test set for the four compositionality levels and two inter-annotator agreement levels.

### 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). Because the distribution is bimodal, we decided to set the cut-off after the first peak, so that MWEs with the score in the $[1, 3]$ range are labeled as non-compositional (NC), while those with the score in the $\langle 3, 5]$ range are labeled as compositional (C). We consider only the best-performing model from the previous evaluation task (the Linear combination model). The model predicts C if the cosine similarity between the MWE vector and the linear combination vector is above a certain threshold, otherwise it predicts NC. We optimized the threshold on the train set by optimizing the F1-score. The results are shown in Table 5.

The classification task is similar to the one considered by Katz and Giesbrecht (2006). In their experiment, they achieved the F1-score of 0.48, but they only considered the additive model for modeling semantic compositionality.

### 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 model performs the worst. We define the error as an absolute difference in the Z-scores between the model-predicted and human-annotated scores. The results seem to suggest that most errors occur on compositional expressions (C), which happen to be the ones on which the annotators easily agreed about the high degree of compositionality.

To explore this hypothesis a bit further, we divided our test set into the subsets based on the compositionality levels (C – 48%, NC1 – 31%, NC2 – 7%, NC3 – 14%), and then calculated correlation on each subset separately. Fig. 2 shows the (absolute) correlation on each of these subsets, for high and low inter-annotator agreement levels. The plot again suggests that the model performs the worst on the compositional MWEs, while it performs best on partially non-compositional MWEs.

A deeper analysis should be done 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.

## 6. Conclusion

We considered the problem of determining the semantic compositionality of Croatian multiword expressions (MWEs) using a composition-based distributional semantics approach. 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 built an LSA model over the Croatian web corpus. We experimented with the additive and multiplicative compositional models. The best-performing model combines the additive and the multiplicative compositional models and the representations of the two individual words. The model achieves a correlation of 0.48 and an F1-score of 0.65.

For future work we plan to enlarge the dataset to allow for a more reliable analysis. Furthermore, we will consider doing the analysis on an unbalanced and hence a more realistic dataset. We also intend to consider the task of token-based semantic compositionality detection, along the lines of Cook et al. (2007) and Sporleder and Li (2009).

## 7. References

O. C. Acosta, A. Villavicencio, and V. P. Moreira. 2011. 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.

Ž. Agić and D. Merkler. 2013. Three syntactic formalisms for data-driven dependency parsing of Croatian. In *Text, Speech, and Dialogue*, pages 560–567. Springer.

Ž. Agić, N. Ljubešić, and D. Merkler. 2013. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proc. of ACL*.

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. 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.

T. Baldwin. 2006. 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*.

C. Bannard, T. Baldwin, and A. Lascarides. 2003. 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.

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

M. Carpuat and M. Diab. 2010. 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.

P. Cook, A. Fazly, and S. Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proc. of the Workshop on a Broader Perspective on Multiword Expressions*, pages 41–48. ACL.

S. Evert. 2005. *The statistics of word cooccurrences*. Ph.D. thesis, Dissertation, Stuttgart University.

S. Evert. 2008. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2:223–233.

M. A. Finlayson and N. Kulkarni. 2011. Detecting multiword expressions improves word sense disambiguation. In *Proc. of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, pages 20–24.

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

Z. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

M. Karan, J. Šnajder, and B. D. Bašić. 2012. Distributional semantics approach to detecting synonyms in Croatian language. *Information Society*, pages 111–116.

G. Katz and E. Giesbrecht. 2006. 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.

S. N. Kim and T. Baldwin. 2006. Automatic identification of English verb particle constructions using linguistic features. In *Proc. of the Third ACL-SIGSEM Workshop on Prepositions*, pages 65–72. ACL.

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

L. Krčmář, K. Ježek, and P. Pecina. 2013. Determining compositionality of 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.

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

Landauer. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

D. Lin. 1999. 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.

N. Ljubešić and T. Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, pages 395–402. Springer.

D. McCarthy, S. Venkatapathy, and A. K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *EMNLP-CoNLL*, pages 369–379.

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

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

S. Reddy, D. McCarthy, S. Manandhar, and S. Gella. 2011. 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.

I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

J. Šnajder, S. Padó, and Ž. Agić. 2013. 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.

C. Sporleder and L. Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. ACL.