Chapter 13

Exploiting multilingual lexical resources to predict MWE compositionality

Bahar Salehi
The University of Melbourne

Paul Cook
University of New Brunswick

Timothy Baldwin
The University of Melbourne

Semantic idiomaticity is the extent to which the meaning of a multiword expression (MWE) cannot be predicted from the meanings of its component words. Much work in natural language processing on semantic idiomaticity has focused on compositionality prediction, wherein a binary or continuous-valued compositionality score is predicted for an MWE as a whole, or its individual component words. One source of information for making compositionality predictions is the translation of an MWE into other languages. This chapter extends two previously-presented studies – Salehi & Cook (2013) and Salehi et al. (2014) – that propose methods for predicting compositionality that exploit translation information provided by multilingual lexical resources, and that are applicable to many kinds of MWEs in a wide range of languages. These methods make use of distributional similarity of an MWE and its component words under translation into many languages, as well as string similarity measures applied to definitions of translations of an MWE and its component words. We evaluate these methods over English noun compounds, English verb-particle constructions, and German noun compounds. We show that the estimation of compositionality is improved when using translations into multiple languages, as compared to simply using distributional similarity in the source language. We further find that string similarity complements distributional similarity.

Bahar Salehi, Paul Cook & Timothy Baldwin. 2018. Exploiting multilingual lexical resources to predict MWE compositionality. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 343–373. Berlin: Language Science Press. DOI:10.5281/zenodo.1469573

1 Compositionality of MWEs

Multiword expressions (hereafter MWEs) are combinations of words which are lexically, syntactically, semantically or statistically idiosyncratic (Sag et al. 2002; Baldwin & Kim 2010). Much research has been carried out on the extraction and identification of MWEs[1] in English (Schone & Jurafsky 2001; Pecina 2008; Fazly et al. 2009) and other languages (Dias 2003; Evert & Krenn 2005; Salehi et al. 2012). However, considerably less work has addressed the task of predicting the meaning of MWEs, especially in non-English languages. As a step in this direction, the focus of this study is on predicting the compositionality of MWEs.

An MWE is fully compositional if its meaning is predictable from its component words, and it is non-compositional (or idiomatic) if not. For example, *stand up* "rise to one's feet" is compositional, because its meaning is clear from the meaning of the components *stand* and *up*. However, the meaning of *strike up* "to start playing" is largely unpredictable from the component words *strike* and *up*.

In this study, following McCarthy et al. (2003) and Reddy et al. (2011), we consider compositionality to be graded, and aim to predict the *degree* of compositionality. For example, in the dataset of Reddy et al. (2011), *climate change* is judged to be 99% compositional, while *silver screen* is 48% compositional and *ivory tower* is 9% compositional. Formally, we model compositionality prediction as a regression task.

An explicit handling of MWEs has been shown to be useful in NLP applications (Ramisch 2012). As an example, Carpuat & Diab (2010) proposed two strategies for integrating MWEs into statistical machine translation. They show that even a large scale bilingual corpus cannot capture all the necessary information to translate MWEs, and that in adding the facility to model the compositionality of MWEs into their system, they could improve translation quality. Acosta et al. (2011) showed that treating non-compositional MWEs as a single unit in information retrieval improves retrieval effectiveness. For example, while searching for documents related to *ivory tower*, we are almost certainly not interested in documents relating to elephant tusks.

Our approach is to use a large-scale multi-way translation lexicon to source translations of a given MWE and each of its component words, and then model the semantic similarity between each component word and the MWE.[2] We consider similarity measures based on distributional similarity from monolingual

[1] In this chapter, we follow Baldwin & Kim (2010) in considering MWE "identification" to be a token-level disambiguation task, and MWE "extraction" to be a type-level lexicon induction task.

[2] Note that we will always assume that there are two component words, but the method is easily generalisable to MWEs with more than two components.

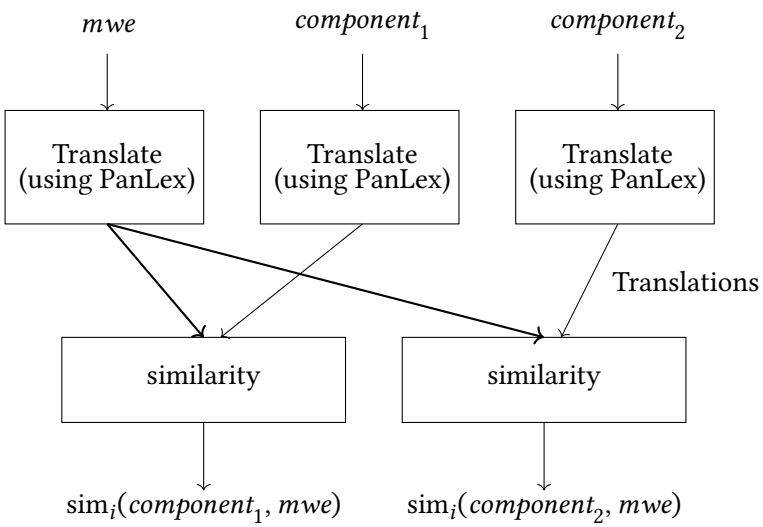


Figure 1: Outline of our approach to computing the similarity of translations of an MWE with each of its component words, for a given target language. sim_i is the similarity between the first or second component of the MWE, and the MWE itself, based on either string or distributional similarity, as measured using language *i*.

corpora for the source language and each of the target languages, as well as string similarity measures applied to definitions of translations of an MWE and its component words as shown in Figure 1. We then consider a variety of approaches to combining similarity scores from the various languages to produce a final compositionality score for the source language expression, as illustrated in Figure 2. We hypothesise that by using multiple translations we will be able to better predict compositionality, and that string similarity measures will complement distributional similarity. Our results confirm our hypotheses, and we further achieve state-of-the-art results over two compositionality prediction datasets.

This chapter combines two previous works – Salehi & Cook (2013) and Salehi et al. (2014) – and extends them in the following ways:

- two new string similarity measures in §4.1.1;
- updated results in §4.2 for the method of Salehi & Cook (2013) such that they are now comparable with the results of the method of Salehi et al. (2014) in §6 – previously these results were not comparable because they used different cross-validation folds during evaluation;
- new results for a dataset of German noun compounds based on the string similarity methods in §4.2;

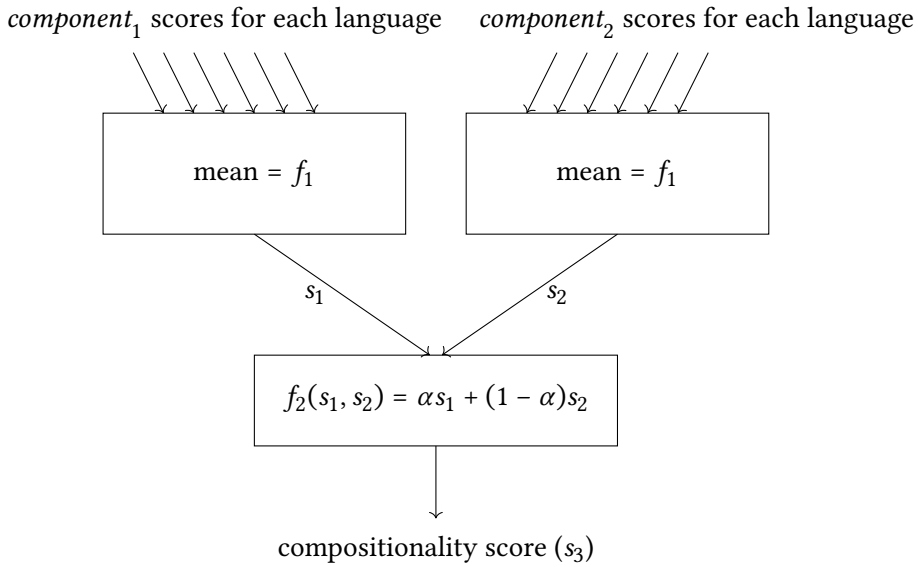


Figure 2: Outline of the method for combining similarity scores from multiple languages, across the components of the MWE.

- additional error analysis in §4.2.1 for English verb-particle constructions;
- two new translation-based similarity approaches, and results for these methods, in §4.2.2;
- experiments considering an alternative translation dictionary in §5;
- analysis of the impact of window size on the distributional similarity approach in §6.1.1.

2 Related work

Most recent work on predicting the compositionality of MWEs can be divided into two categories: language/construction-specific and general-purpose. This can be at either the token-level (over token occurrences of an MWE in a corpus) or type-level (over the MWE string, independent of usage). The bulk of work on compositionality has been language/construction-specific and operated at the token-level, using dedicated methods to identify instances of a given MWE, and specific properties of the MWE in that language to predict compositionality (Lin 1999; Kim & Baldwin 2007; Fazly et al. 2009).

General-purpose token-level approaches such as distributional similarity have

been commonly applied to infer the semantics of a word/MWE (Schone & Jurafsky 2001; Baldwin et al. 2003; Reddy et al. 2011). These techniques are based on the assumption that the meaning of a word is predictable from its context of use, via the neighbouring words of token-level occurrences of the MWE. In order to predict the compositionality of a given MWE using distributional similarity, the different contexts of the MWE are compared with the contexts of its components, and the MWE is considered to be compositional if the MWE and component words occur in similar contexts.

Identifying token instances of MWEs is not always easy, especially when the component words do not occur sequentially. For example, consider *put on* in ***put*** *your jacket* ***on***, and ***put*** *your jacket* ***on*** *the chair*. In the first example *put on* is an MWE, while in the second example, *put on* is a simple verb with prepositional phrase and not an instance of an MWE. Moreover, if we adopt a conservative identification method, the number of token occurrences will be limited and the distributional scores may not be reliable. Additionally, for morphologically-rich languages, it can be difficult to predict the different word forms a given MWE type will occur across, posing a challenge for our requirement of no language-specific preprocessing.

Pichotta & DeNero (2013) proposed a token-based method for identifying English phrasal verbs based on parallel corpora for 50 languages. They show that they can identify phrasal verbs better when they combine information from multiple languages, in addition to the information they get from a monolingual corpus. This finding lends weight to our hypothesis that using translation data and distributional similarity from each of a range of target languages, can improve compositionality prediction. Having said that, the general applicability of their method is questionable – there are many parallel corpora involving English, but for other languages, this tends not to be the case.

In the literature, compositionality has been viewed as either compositionality of the whole MWE as one unit (McCarthy et al. 2003; Venkatapathy & Joshi 2005; Katz 2006; Biemann & Giesbrecht 2011; Farahmand et al. 2015), or compositionality relative to each component (Reddy et al. 2011; Hermann et al. 2012; Schulte im Walde et al. 2013). There have also been studies which focus only on one component of the MWE. For example, Korkontzelos & Manandhar (2009) induce the most probable sense of an MWE first, and then measure the semantic similarity between the MWE and its semantic head. This approach of considering only the head component has been shown to be quite accurate for English verb-particle constructions (Bannard et al. 2003). However, this might not always be the case. For example, as shown in Reddy et al. (2011), the compositionality of the first

noun (the modifier) has more impact than the second noun (the head) for English noun compounds.

Elsewhere, a lot of work has been done on specific types of MWE in specific languages. In English, studies have been done specifically on VPCs (McCarthy et al. 2003; Bannard et al. 2003), verb+noun MWEs (Venkatapathy & Joshi 2005; McCarthy et al. 2007; Fazly et al. 2009), noun compounds (Reddy et al. 2011), and adjective+noun compounds (Vecchi et al. 2011). There have also been studies focusing on a specific language other than English, such as Arabic (Saif et al. 2013) and German (Schulte im Walde et al. 2013). This chapter investigates language independent approaches applicable to any type of MWE in any language.

3 Resources

In this section, we describe the datasets used to evaluate our method and the multilingual dictionary it requires. These are the same resources as used by Salehi & Cook (2013) and Salehi et al. (2014).

3.1 Datasets

We evaluate our proposed method over three datasets (two English, one German), as described below.

3.1.1 English noun compounds (ENC)

Our first dataset is made up of 90 binary English noun compounds, from the work of Reddy et al. (2011). Each noun compound was annotated by multiple annotators using the integer scale 0 (fully non-compositional) to 5 (fully compositional). A final compositionality score was then calculated as the mean of the scores from the annotators. If we simplistically consider 2.5 as the threshold for compositionality, the dataset is relatively well balanced, containing 48% compositional and 52% non-compositional noun compounds.

Spearman correlation was used to get an estimate of inter-annotator agreement. The average correlation for compound compositionality was $\rho = 0.522$. This score was slightly higher for the compositionality of components ($\rho = 0.570$ for the first component and $\rho = 0.616$ for the second component).

3.1.2 English verb-particle constructions (EVPC)

The second dataset contains 160 English verb-particle constructions (VPCs), from the work of Bannard (2006). In this dataset, a verb-particle construction consists of a verb (the head) and a prepositional particle (e.g. *hand in*, *look up* or *battle on*). For each component word (the verb and particle, respectively), multiple annotators were asked whether the VPC entails the component word. In order to translate the dataset into a regression task, we calculate the overall compositionality as the number of annotations of entailment for the verb, divided by the total number of verb annotations for that VPC. That is, following Bannard et al. (2003), we only consider the compositionality of the verb component in our experiments. The Kappa score between the multiple annotators is 0.372 for verb and 0.352 for the particle component.

3.1.3 German noun compounds (GNC)

Our final dataset is made up of 246 German noun compounds (von der Heide & Borgwaldt 2009; Schulte im Walde et al. 2013). Multiple annotators were asked to rate the compositionality of each German noun compound on an integer scale of 1 (non-compositional) to 7 (compositional). The overall compositionality score is then calculated as the mean across the annotators. Note that the component words are provided as part of the dataset, and that there is no need to perform decompounding. This dataset is significant as it is non-English and because of the fact that German has relatively rich morphology, which we expect to impact on the identification of both the MWE and the component words.

3.2 Multilingual dictionary

To translate the MWEs and their components, we use PanLex (Baldwin et al. 2010). This online dictionary is massively multilingual, covering more than 1353 languages. The translations are sourced from handmade electronic dictionaries. It contains lemmatised words and MWEs in a large variety of languages, with lemma-based (and less frequently sense-based) links between them.

For each MWE dataset (see §3.1), we translate each MWE, and its component words, from the source language into many target languages. These translations will be used in §4 and §6. In instances where there is no direct translation in a given language for a term, we use a pivot language to find translation(s) in the target language. For example, the English noun compound *silver screen* has direct translations in only 13 languages in PanLex, including Vietnamese (*màn bac*) but

not French. There is, however, a translation of *màn bac* into French (*cinéma*), allowing us to infer an indirect translation between *silver screen* and *cinéma*. In this way, if there are no direct translations into a particular target language, we search for a single-pivot translation via each of our other target languages, and combine them all together as our set of translations for the target language of interest.

4 String similarity

In this section we present our string similarity-based method for predicting compositionality, followed by experimental results using this method. This section extends Salehi & Cook (2013) as described in §1.

4.1 Compositionality prediction based on string similarity

We hypothesize that compositional MWEs are more likely to be word-for-word translations in a given language than non-compositional MWEs. Hence, if we can locate the translations of the components in the translation of the MWE, we can deduce that it is compositional. As an example of our method, consider the English-to-Persian translation of *kick the bucket* as a non-compositional MWE and *make a decision* as a semi-compositional MWE (Table 1).[3] By locating the translation of *decision* (*tasmim*) in the translation of *make a decision* (*tasmim gereftan*), we can deduce that it is semi-compositional. However, we cannot locate any of the component translations in the translation of *kick the bucket*. Therefore, we conclude that it is non-compositional. Note that in this simple example, the match is word-level, but that due to the effects of morphophonology, the more likely situation is that the components don't match exactly (as we observe in the case of *khadamaat* and *khedmat* for the *public service* example), which motivates our use of string similarity measures which can capture partial matches.

4.1.1 String similarity measures

We consider the following string similarity measures to compare the translations. In each case, we normalize the output value to the range $[0, 1]$, where 1 indicates identical strings and 0 indicates completely different strings. We will indicate the translation of the MWE in a particular language t as mwe^t, and the translation of a given component in language t as $component^t$.

[3]Note that the Persian words are transliterated into English for ease of understanding.

Table 1: English MWEs and their components with their translation in Persian. Direct matches between the translation of an MWE and its components are shown in **bold**; partial matches are shown in *italics*.

English	Persian translation
kick the bucket	mord
kick	zad
the	–
bucket	satl
make a decision	**tasmim** gereft
make	sakht
a	yek
decision	**tasmim**
public service	*khadamaat* **omumi**
public	**omumi**
service	*khedmat*

Longest common substring (LCS): The LCS measure finds the longest common substring between two strings. For example, the LCS between `ABABC` and `BABCAB` is `BABC`. We calculate a normalized similarity value based on the length of the LCS as follows:

$$\frac{\text{LCS}(mwe^t, component^t)}{\min(\text{len}(mwe^t), \text{len}(component^t))} \tag{13.1}$$

Levenshtein (LEV1): The Levenshtein distance calculates the number of basic edit operations required to transform one word into the other. Edits consist of single-letter insertions, deletions or substitutions. We normalize LEV1 as follows:

$$1 - \frac{\text{LEV1}(mwe^t, component^t)}{\max(\text{len}(mwe^t), \text{len}(component^t))} \tag{13.2}$$

Levenshtein with substitution penalty (LEV2): One well-documented feature of Levenshtein distance (Baldwin 2009) is that substitutions are in fact the combination of an addition and a deletion, and as such can be considered to be two edits. Based on this observation, we experiment with a variant of LEV1 with this penalty applied for substitutions. Similarly to LEV1, we normalize as follows:

$$1 - \frac{\text{LEV2}(mwe^t, component^t)}{\text{len}(mwe^t) + \text{len}(component^t)} \tag{13.3}$$

Smith Waterman (SW): This method is based on the Needleman-Wunsch algorithm,[4] and was developed to locally-align two protein sequences (Smith & Waterman 1981). It finds the optimal similar regions by maximizing the number of matches and minimizing the number of gaps necessary to align the two sequences. For example, the optimal local sequence for the two sequences below is `AT--ATCC`, in which "–" indicates a gap:

Seq1: **ATGCATCC**CATGAC
Seq2: TCT**ATATCC**GT

As the example shows, it looks for the longest common string but has a built-in mechanism for including gaps in the alignment (with penalty). This characteristic of SW might be helpful in our task, because there may be morphophonological variations between the MWE and component translations (as seen above in the *public service* example). We normalize SW similarly to LCS:

$$\frac{\text{len(alignedSequence)}}{\min(\text{len}(mwe^t), \text{len}(component^t))} \tag{13.4}$$

The aligned sequence is the combination of the common characters in the optimal local sequence we found using SW. In the above example, the aligned sequence is `ATATCC`.

Jaccard and Dice similarity: For further analysis, we experiment with Jaccard and Dice similarity, which are well-known for measuring the similarity between two sentences or bodies of text (Gomaa & Fahmy 2013). Both methods view the texts as sets of words, with similarity based on the size of the intersection between the sets, but differ in the way they are normalized. In our case, we expect relatively low overlap at the word level due to morphophonology, and therefore

[4]The Needleman-Wunsch (NW) algorithm was designed to align two sequences of amino-acids (Needleman & Wunsch 1970). The algorithm looks for the sequence alignment which maximizes the similarity. As with the LEV score, NW minimizes edit distance, but also takes into account character-to-character similarity based on the relative distance between characters on the keyboard. We exclude this score because it is highly similar to the LEV scores and we did not obtain encouraging results using NW in our preliminary experiments.

calculate Jaccard (J) and Dice (D) at the character- instead of word-level as follows:

$$J = \frac{|mwe^t \cap component^t|}{|mwe^t| + |component^t| - |mwe^t \cap component^t|} \tag{13.5}$$

$$D = \frac{2 * |mwe^t \cap component^t|}{|component^t| + |mwe^t|} \tag{13.6}$$

4.1.2 Calculating compositionality

Given the string similarity scores calculated between the translations for a given component word and the MWE, we need some way of combining scores across component words. First, we measure the compositionality of each component within the MWE (s_1 and s_2):

$$s_1 = f_1(\text{sim}_1(w_1, mwe), ..., \text{sim}_i(w_1, mwe)) \tag{13.7}$$

$$s_2 = f_1(\text{sim}_1(w_2, mwe), ..., \text{sim}_i(w_2, mwe)) \tag{13.8}$$

where sim is a similarity measure, sim_i indicates that the calculation is based on translations in language i, and f_1 is a score combination function.

Then, we compute the overall compositionality of the MWE (s_3) from s_1 and s_2 using f_2:

$$s_3 = f_2(s_1, s_2) \tag{13.9}$$

Since we often have multiple translations for a given component word/MWE in PanLex, we exhaustively compute the similarity between each MWE translation and component translation, and use the highest similarity as the result of sim_i. If an instance does not have a direct/indirect translation in PanLex, we assign a default value, which is the mean of the highest and lowest annotation score for the dataset under consideration. Note that word order is not an issue in our method, as we calculate the similarity independently for each MWE component.

We consider simple functions for f_1 such as mean, median, product, minimum and maximum. f_2 was selected to be the same as f_1 in all situations, except when we use mean for f_1. Here, following Reddy et al. (2011), we experimented with weighted mean:

$$f_2(s_1, s_2) = \alpha s_1 + (1 - \alpha)s_2 \tag{13.10}$$

Based on 3-fold cross-validation, we chose $\alpha = 0.7$ for ENC.[5] We found $\alpha = 0.7$ is also optimal for GNC.

Since we do not have judgements for the compositionality of the full VPC in EVPC (we instead have separate judgements for the verb and particle), we cannot use f_2 for this dataset. Bannard et al. (2003) observed that nearly all of the verb-compositional instances were also annotated as particle-compositional by the annotators. In line with this observation, we use s_1 (based on the verb) as the compositionality score for the full VPC.

4.1.3 Language selection

Our method is based on the translation of an MWE into many languages. First, we chose 54 languages for which relatively large corpora were available.[6] The coverage, or the number of instances which have direct/indirect translations in PanLex, varies from one language to another. In preliminary experiments, we noticed that there is a high correlation (between roughly $r = 0.6$ and 0.8 across the three datasets) between the usefulness of a language and its translation coverage on MWEs. Therefore, we excluded languages with MWE translation coverage of less than 50%. Based on nested 10-fold cross-validation in our experiments, we select the 10 most useful languages for each cross-validation training partition, based on the Pearson correlation between the given scores in that language and human judgements.[7] The 10 best languages are selected based only on the training set for each fold. (The languages selected for each fold will later be used to predict the compositionality of the items in the testing portion for that fold.)

4.2 Results

As mentioned above, we perform nested 10-fold cross-validation to select the 10 best languages on the training data for each fold. The selected languages for a given fold are then used to compute s_1 and s_2 (and s_3 for NCs) for each instance

[5] We considered values of α from 0 to 1, incremented by 0.1.

[6] In §6 these corpora will be used to compute distributional similarity. Note that the string similarity methods of interest here do not rely on the availability of large corpora.

[7] Note that for VPCs, we calculate the compositionality of only the verb part, because we don't have the human judgements for the whole VPC.

Table 2: Correlation (r) on each dataset, for each string similarity measure. The best correlation for each dataset is shown in boldface.

Method	ENC	EVPC	GNC
SW	**0.644**	0.349	0.379
LCS	**0.644**	**0.385**	0.372
LEV1	0.502	0.328	0.318
LEV2	0.566	0.327	**0.389**
Jaccard	0.474	0.335	0.299
Dice	0.557	0.331	0.370
Unsupervised (family)	0.556	0.257	0.164
Unsupervised (coverage)	0.642	0.323	0.343

in the test set for that fold. The scores are compared with human judgements using Pearson's correlation.

We experimented with five functions for f_1, namely mean, median, product, maximum and minimum. Among these functions, mean performed consistently better than the others, and as such we only present results using mean in Table 2.

For ENC, LCS and SW perform best, while for EVPC, LCS performs best with SW being the next best measure. Both LCS and SW look for a sequence of similar characters, unlike LEV1 and LEV2, which are not affected by match contiguity. For GNC, LEV2, SW and LCS perform better than LEV1. However, unlike the other two datasets, LEV2 is the best performing method, and SW is slightly better than LCS.

For all datasets, Jaccard and Dice perform worse than SW and LCS. This shows that, despite being useful in measuring the similarity between sentences, these two measures do not perform well in this compositionality prediction task. The relatively poor performance of these measures could be because, unlike the other measures, Jaccard and Dice are calculated independently of the order of characters. Dice performs better than Jaccard for ENC and GNC, while Jaccard performs slightly better than Dice for EVPC.

The results support our hypothesis that using multiple target languages rather than one, results in a more accurate prediction of MWE compositionality. Our best result using the 10 selected languages on ENC is $r = 0.644$, as compared to the best single-language correlation of $r = 0.543$ for Portuguese. On EVPC, the best LCS result for the verb component is $r = 0.385$, as compared to the

best single-language correlation of $r = 0.342$ for Lithuanian. For GNC, the best correlation of $r = 0.389$ is well above the highest correlation of a single language of roughly $r = 0.32$.

In §6 we will combine this string similarity approach with an approach based on distributional similarity, and compare it against a baseline and state-of-the-art approaches.

4.2.1 Error analysis

We analysed items in ENC which have a high absolute difference (more than 2.5) between the human annotation and our scores (using LCS and mean). The words are *cutting edge*, *melting pot*, *gold mine* and *ivory tower*, which are non-compositional according to ENC. After investigating their translations, we came to the conclusion that the first three MWEs have word-for-word translations in most languages. Hence, they disagree with our hypothesis that word-for-word translation is a strong indicator of compositionality. The word-for-word translations might be because of the fact that they have both compositional and non-compositional senses, or because they are calques (loan translations). However, we have tried to avoid such problems with calques by using translations into several languages.

For *ivory tower* ("a state of mind that is discussed as if it were a place")[8] we noticed that we have a direct translation into 13 languages. Other languages have indirect translations. By checking the direct translations, we noticed that, in French, the MWE is translated to *tour* and *tour d'ivoire*. A noisy (wrong) translation of *tour* "tower" resulted in wrong indirect translations for *ivory tower* and an inflated estimate of compositionality.

We repeat the same error analysis for the EVPC dataset. The items with a high difference between the human annotation and our scores are: *carry out*, *drop out*, *get in*, *carry away*, *wear down* and *turn on*. All of these items are annotated as non-compositional. These VPCs also have a compositional sense beside the non-compositional meaning. Also, as with the ENC dataset, we have problems of calques. For example, *drop out* when translated to German (*ausfallen*) includes the word *fallen*, which is one of the translations of *drop*.

4.2.2 Unsupervised approach

The proposed translation-based string similarity approach has been supervised so far, in that the best target languages are selected based on training data. In this

[8]This definition is from Wordnet 3.1.

section, we propose two unsupervised approaches in which: (1) only the target languages of the same language family as the source language are considered; and (2) only the 10 target languages with the highest translation coverage are considered.

Languages of the same family: We hypothesize that translations into target languages in the same language family as the source language might be particularly useful for compositionality prediction for MWEs in the source language. To test this hypothesis, we consider an unsupervised approach in which only languages in the same family as the source language are used when computing the compositionality scores.

In this unsupervised approach, LCS scores of the languages of the same family as the source language (here Germanic, for both the English and German datasets) are considered. The Germanic languages among our 54 languages are: English, German, Danish, Dutch, Icelandic, Luxembourgish, Norwegian and Swedish.

Results for this unsupervised approach are shown in Table 2 ("Unsupervised (family)"). This approach performs substantially worse than the corresponding supervised approach based on LCS, for each dataset. This drop in performance could be because almost none of the 10 best languages selected in the supervised approach are in the same language family as the source language. The shared languages between the supervised approach and this approach are Dutch and Norwegian for ENC, English for GNC. There is no shared language between the two approaches when using EVPC.

Languages with the highest translation coverage: In the proposed supervised setup, the best target languages are those whose scores have the highest correlation with gold-standard annotations. According to our experiments, we showed that there is a strong correlation between being a good language for this compositionality prediction task and its coverage in PanLex (in the range of roughly $0.6 < r < 0.8$ across the three datasets). In other words, the target languages to which most of the source language MWEs have a translation in PanLex, result in higher correlation for compositionality prediction.

We now consider an unsupervised approach, in which only the 10 target languages with the highest translation coverage are considered. The results of this unsupervised approach, again using LCS, are shown in Table 2 ("Unsupervised (coverage)"). According to the results, despite the lower correlation scores for the proposed unsupervised method, this method is comparable to the supervised

Table 3: The 10 languages with the highest translation coverage for ENC, EVPC and GNC. Languages also selected by the supervised approach are shown in **boldface**.

ENC	EVPC	GNC
German	German	**English**
Finnish	Finnish	Japanese
French	French	French
Italian	Italian	Italian
Russian	Japanese	Russian
Spanish	Hungarian	Hungarian
Portuguese	Dutch	Dutch
Japanese	**Polish**	Turkish
Chinese	Chinese	Chinese
Czech	**Czech**	**Czech**

approach. Therefore, in the case of not having a training set for a group of MWEs (no matter in what language or what type of MWE), we suggest using the target languages to which the majority of those MWEs have a translation.

The 10 languages with highest correlation for ENC, EVPC and GNC are shown in Table 3. There is some overlap between the list of languages with the highest coverage and the 10 best languages selected in our supervised approach, as shown in boldface for each dataset.

5 An alternative multilingual dictionary

In this section we consider the same string similarity-based approach to predicting compositionality as in §4.1, but using an alternative multilingual dictionary to PanLex, specifically dict.cc.[9]

dict.cc is a translation dictionary that provides translations for both English and German into 26 languages spoken in Europe. It is a crowd-sourced dictionary, with translations being contributed, and refined, by users. Due to the relatively small number of languages it covers, relying on dict.cc goes against our goals of developing compositionality prediction methods that are applicable to any language; we could not use dict.cc to predict the compositionality of, for example, a French MWE, because translations are not available for French into many languages (only English and German). Nevertheless, by considering the

[9]https://www.dict.cc/

Table 4: Correlation (r) on each dataset, for each string similarity measure, using dict.cc and PanLex as the translation dictionary. The best correlation for each dataset is shown in boldface.

Dictionary	Method	ENC	EVPC	GNC
dict.cc				
	SW	.269	.217	.514
	LCS	.251	.262	**.523**
	LEV1	.181	.161	.482
	LEV2	.163	.189	.474
	Jaccard	.158	.127	.442
	Dice	.230	.192	.420
PanLex				
	SW	**.559**	**.294**	.270
	LCS	.551	.276	.290
	LEV1	.388	.274	.276
	LEV2	.512	.281	.262
	Jaccard	.459	.241	.267
	Dice	.541	.235	.197

use of an alternative translation dictionary (which is applicable to the English and German datasets we use for evaluation) we can learn whether our approach to predicting compositionality implicitly relies on information particular to PanLex, or whether an alternative dictionary can be substituted in its place.

We chose target languages available in dict.cc that overlap with the set of 54 target languages used in experiments with PanLex in §4.1. This resulted in 22 target languages. We introduced this restriction, as opposed to using all languages available in dict.cc, to allow us to compare PanLex and dict.cc when using the exact same set of target languages.

Results for the string similarity-based approach to predicting compositionality, using dict.cc and PanLex, each with the same 22 target languages, are shown in Table 4. The 10 best languages are selected using the same method as in §4.1.3.

For each translation dictionary and dataset, the best method is always one of either SW or LCS, and in many cases these are the top two methods (with the exceptions being EVPC and GNC using PanLex). These methods were also found to perform well in §4.2 when using PanLex and 54 target languages. This

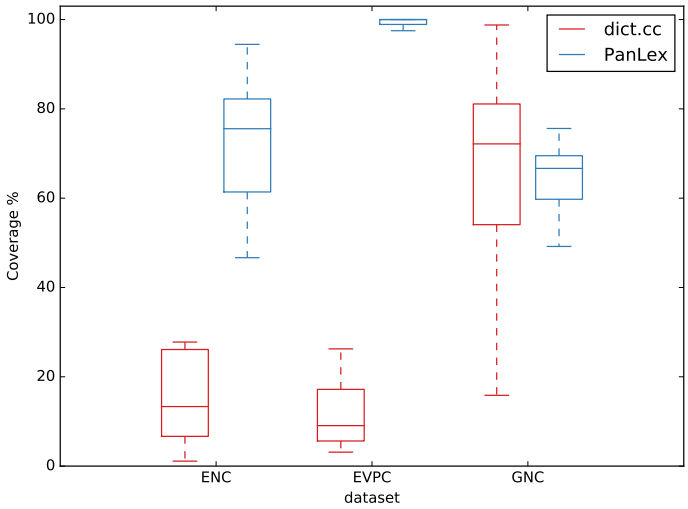


Figure 3: Boxplots showing the percentage of expressions in each dataset covered by dict.cc and PanLex, over the 22 target languages.

demonstrates that the methods are robust to the choice of specific translation dictionary, and when the number of target languages is substantially reduced.

There are, however, substantial differences between the results using different translation dictionaries. For any combination of dataset and method, the results using PanLex are always better than those using dict.cc for ENC and EVPC, while for GNC, the results using dict.cc are always better. To understand why this is the case, for each dataset and dictionary, and for each of the 22 target languages, we computed the proportion of expressions for which translations are available. Boxplots illustrating these findings are shown in Figure 3. On average across the target languages, many more expressions are covered by PanLex than dict.cc for ENC and EVPC, while for GNC the coverage is higher for dict.cc. For example, according to Figure 3, for EVPC the coverage for almost all of the 22 target languages is close to 100% in PanLex.

Because it in keeping with our goal of building methods for compositionality prediction that are applicable to any language, and because it gives the best results in two out of three cases for the datasets used for evaluation, we will only consider PanLex as the translation dictionary for the remainder of this chapter.

6 Distributional similarity

In this section we describe a method for predicting compositionality based on the same framework as in §4, but using distributional similarity instead of string similarity. This section extends Salehi et al. (2014) as described in §1.

6.1 Compositionality prediction based on distributional similarity

To predict the compositionality of a given MWE, we first measure the semantic similarity between the MWE and each of its component words using distributional similarity based on a monolingual corpus in the source language. We then repeat the process for translations of the MWE and its component words into each of a range of target languages, calculating distributional similarity using a monolingual corpus in the target language. We additionally use supervised learning to identify which target languages (or what weights for each language) optimise the prediction of compositionality. We hypothesise that by using multiple translations – rather than only information from the source language – we will be able to better predict compositionality. We further optionally combine our proposed approach with the LCS-based string similarity method from §4.

Below, we detail our method for calculating distributional similarity in a given language, the different methods for combining similarity scores into a single estimate of compositionality, and finally the method for selecting the target languages to use in calculating compositionality.

6.1.1 Calculating distributional similarity

We collected monolingual corpora for each of the 52 languages (51 target languages + 1 source language) from XML dumps of Wikipedia. These languages are based on the 54 target languages used in §4, excluding Spanish because we happened not to have a dump of Spanish Wikipedia, and also Chinese and Japanese because of the need for a language-specific word tokeniser. The raw corpora were preprocessed using the WP2TXT toolbox[10] to eliminate XML tags, HTML tags and hyperlinks, and then tokenisation based on whitespace and punctuation was performed. The corpora vary in size from roughly 750M tokens for English, to roughly 640K tokens for Marathi.

In order to be consistent across all languages and to be as language-independent as possible, we calculate distributional similarity in the following manner for a given language.

[10] http://wp2txt.rubyforge.org/

Table 5: Results of distributional similarities using 10 best languages on ENC dataset (N is window size)

Context window	Correlation (r)
Sentence	0.425
Window (N=3)	0.175
Window (N=3, with positional index)	0.031

Tokenisation is based on whitespace delimiters and punctuation; no lemmatisation or case-folding is carried out. Token instances of a given MWE or component word are identified by full-token n-gram matching over the token stream. We assume that all full stops and equivalent characters for other orthographies are sentence boundaries, and chunk the corpora into (pseudo-)sentences on the basis of them. For each language, we identify the 51st–1050th most frequent words, and consider them to be content-bearing words, in the manner of Schütze (1997). This is based on the assumption that the top-50 most frequent words are stop words, and not a good choice of word for calculating distributional similarity over. That is not to say that we can't calculate the distributional similarity for stop words, however (as we will for the EVPC dataset) they are simply not used as the dimensions in our calculation of distributional similarity.

We form a vector of content-bearing words across all token occurrences of the target word, on the basis of these 1000 content-bearing words. Our preliminary results on selecting the best context window size are shown in Table 5. According to this table, for predicting the compositionality using the best 10 languages, the sentence context window results in a higher correlation. We use sentence boundaries as the context window in the rest of our experiments. According to Weeds (2003) and Padó & Lapata (2007), using dependency relations with the neighbouring words of the target word can better predict the meaning of the target word. However, in line with our assumption of no language-specific preprocessing, we just use word co-occurrence. Finally, distributional similarity is calculated over these context vectors using cosine similarity.

6.1.2 Calculating compositionality

The procedure of calculating the compositionality is similar to what we used in §4.1.2: after translating the MWE and its components into multiple languages and measuring the distributional similarity between the translations of the MWE and

its components (Figure 1), we find the best languages according to the training set. Then, we combine the scores from those best languages and finally calculate a combined compositionality score from the individual distributional similarities between each component word and the MWE. Based on our findings in §4.1.2, we combine the component scores using the weighted mean (Figure 2):

$$\text{Compositionality} = \alpha s_1 + (1 - \alpha) s_2 \tag{13.11}$$

where s_1 and s_2 are the scores for the first and the second component, respectively. We use different α settings for each dataset, based on the settings from §4.1.2.

We experiment with a range of methods for calculating compositionality, as follows:

CS_{L1}: calculate distributional similarity using only distributional similarity in the source language corpus. (This is the approach used by Reddy et al. (2011), as discussed in §2.)

CS_{L2N}: exclude the source language and compute the mean of the distributional similarity scores for the best-N target languages. The value of N is selected according to training data, as detailed in §6.1.3.[11]

CS_{L1+L2N}: calculate distributional similarity over both the source language (CS_{L1}) and the mean of the best-N languages (CS_{L2N}), and combine via the arithmetic mean.[12] This is to examine the hypothesis that using multiple target languages is better than just using the source language.

$CS_{SVR(L1+L2)}$: train a support vector regressor (SVR: Smola & Schölkopf (2004)) over the distributional similarities for all 52 languages (source and target languages).

CS_{string}: calculate string similarity using the LCS-based method of §4. LCS is chosen because, in general, it performs better than the other string similarity measures.

[11] In the case that no translation (direct or indirect) can be found for a given source language term into a particular target language, the compositionality score for that target language is set to the average across all target languages for which scores can be calculated for the given term. If no translations are available for any target language (e.g. the term is not in PanLex) the compositionality score for each target language is set to the average score for that target language across all other source language terms.

[12] We also experimented with taking the mean over all the languages – target and source – but found it best to combine the scores for the target languages first, to give more weight to the source language.

$CS_{string+L1}$: calculate the mean of the string similarity (CS_{string}) and distributional similarity in the source language.

CS_{all}: calculate the mean of the string similarity (CS_{string}) and distributional similarity scores (CS_{L1} and CS_{L2N}).

6.1.3 Selecting target languages

We experiment with two approaches for combining the compositionality scores from multiple target languages.

First, in CS_{L2N} (and CS_{L1+L2N} and CS_{all} that build off it), following the approach from §4.1.3, we use training data to rank the target languages according to Pearson's correlation between the predicted compositionality scores and the gold-standard compositionality judgements. However, in this case, based on this ranking, we take the best-N languages (instead of the best-10 languages as in §4.1.3) and again combine the individual compositionality scores by taking the arithmetic mean. We select N by determining the value that optimises the correlation over the training data. In other words, the selection of N and accordingly the best-N languages are based on nested cross-validation over training data, independently of the test data for that iteration of cross-validation.

Second in $CS_{SVR(L1+L2)}$, we take the compositionality scores from the source and all 51 target languages, combine them into a feature vector, and train an SVR over the data using LIBSVM.[13]

6.2 Results

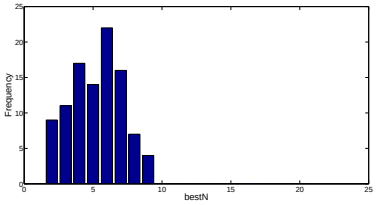
All experiments are carried out using 10 iterations of 10-fold cross validation, randomly partitioning the data independently on each of the 10 iterations, and averaging across all 100 test partitions in our presented results (Table 6). In the case of CS_{L2N} and other methods that make use of it (i.e. CS_{L1+L2N} and CS_{all}), the languages selected for a given training fold are then used to compute the compositionality scores for the instances in the test set.

Figure 4 shows histograms of the number of times each N is selected over 100 folds on ENC, EVPC and GNC datasets, respectively. From the histograms, $N = 6$, $N = 15$ and $N = 2$ are the most commonly selected settings for ENC, EVPC and GNC, respectively. That is, multiple languages are generally used, but more languages are used for English VPCs than either of the compound noun datasets.

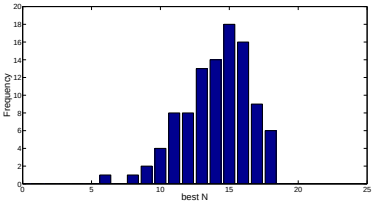
[13] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Table 6: Pearson's correlation on the ENC, EVPC and GNC datasets

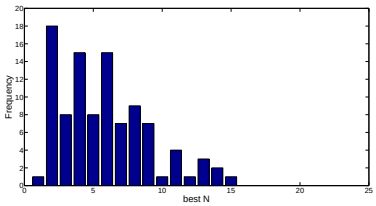
Method	Summary of the Method	ENC	EVPC	GNC
CS_{L1}	Source language	0.700	0.177	0.141
CS_{L2N}	Best-N target languages	0.434	0.398	0.113
CS_{L1+L2N}	Source + best-N target languages	0.725	0.312	0.178
$CS_{SVR(L1+L2)}$	SVR (Source + all 51 target languages)	**0.744**	0.389	0.085
CS_{string}	String Similarity	0.644	0.385	**0.372**
$CS_{string+L1}$	CS_{string} +CS_{L1}	0.739	0.360	0.353
CS_{all}	CS_{L1} + CS_{L2N} + CS_{string}	0.732	**0.417**	0.364



(a) ENC



(b) EVPC



(c) GNC

Figure 4: Histograms displaying how many times a given N is selected as the best number of languages over each dataset. For example, according to the GNC chart, there is a peak for $N = 2$, which shows that over 100 folds, the best-2 languages achieved the highest correlation on 18 folds.

Further analysis reveals that 32 (63%) target languages for ENC, 25 (49%) target languages for EVPC, and only 5 (10%) target languages for GNC have a correlation of $r \geq 0.1$ with gold-standard compositionality judgements. On the other hand, 8 (16%) target languages for ENC, 2 (4%) target languages for EVPC, and no target languages for GNC have a correlation of $r \leq -0.1$.

6.2.1 ENC results

English noun compounds are relatively easy to identify in a corpus,[14] because the components occur sequentially, and the only morphological variation is in noun number (singular vs. plural). In other words, the precision for our token matching method is very high, and the recall is also acceptably high. Partly as a result of the ease of identification, we get a high correlation of $r = 0.700$ for CS_{L1} (using only source language data). Using only target languages (CS_{L2N}), the results drop to $r = 0.434$, but when we combine the two (CS_{L1+L2N}), the correlation is higher than using only source or target language data, at $r = 0.725$. When we combine all languages using SVR, we achieve our best results on this dataset of $r = 0.744$, an improvement over the previous state of the art of Reddy et al. (2011) ($r = 0.714$). These last two results support our hypothesis that using translation data can improve the prediction of compositionality. The results for string similarity on its own (CS_{string}, $r = 0.644$) are slightly lower than those using only source language distributional similarity, but when combined with CS_{L1+L2N} (i.e. CS_{all}) there is a slight rise in correlation (from $r = 0.725$ to $r = 0.732$).

6.2.2 EVPC results

English VPCs are hard to identify. As discussed in §2, VPC components may not occur sequentially, and even when they do occur sequentially, they may not be a VPC. As such, our simplistic identification method has low precision and recall (hand analysis of 927 identified VPC instances would suggest a precision of around 74%). There is no question that this is a contributor to the low correlation for the source language method (CS_{L1}; $r = 0.177$). When we use target languages instead of the source language (CS_{L2N}), the correlation jumps substantially to $r = 0.398$.

When we combine English and the target languages (CS_{L1+L2N}), the results are actually lower than just using the target languages, because of the high weight on the target language, which is not desirable for VPCs, based on the source language results. Even for $CS_{SVR(L1+L2)}$, the results ($r = 0.389$) are slightly below the target language-only results. This suggests that when predicting the compositionality of MWEs which are hard to identify in the source language, it may actually be better to use target languages only. The results for string similarity (CS_{string}: $r = 0.385$) are similar to those for CS_{L2N}. However, as with the ENC

[14] Although see Lapata & Lascarides (2003) for discussion of the difficulty of reliably identifying low-frequency English noun compounds.

dataset, when we combine string similarity and distributional similarity (CS_{all}), the results improve, and we achieve the state of the art for the dataset.

In Table 7, we present classification-based evaluation over a subset of EVPC, binarising the compositionality judgements in the manner of Bannard et al. (2003). Our method achieves state-of-the-art results in terms of overall F-score and accuracy.

Table 7: Results (%) for the binary compositionality prediction task on the EVPC dataset

Method	Precision	Recall	F-score ($\beta = 1$)	Accuracy
Bannard et al. (2003)	60.8	66.6	63.6	60.0
CS_{string}	**86.2**	71.8	77.4	69.3
CS_{all}	79.5	**89.3**	**82.0**	**74.5**

6.2.3 GNC results

German is a morphologically-rich language, with marking of number and case on nouns. Given that we do not perform any lemmatisation or other language-specific preprocessing, we inevitably achieve low recall for the identification of noun compound tokens, although the precision should be nearly 100%. Partly because of the resultant sparseness in the distributional similarity method, the results for CS_{L1} are low ($r = 0.141$), although they are lower again when using target languages ($r = 0.113$). However, when we combine the source and target languages (CS_{L1+L2N}) the results improve to $r = 0.178$. The results for $CS_{SVR(L1+L2)}$, on the other hand, are very low ($r = 0.085$). Ultimately, simple string similarity achieves the best results for the dataset ($r = 0.372$), and this result actually drops slightly when combined with the distributional similarities.

To better understand the reason for the lacklustre results using SVR, we carried out error analysis and found that, unlike the other two datasets, about half of the target languages return scores which correlate negatively with the human judgements. When we filter these languages from the data, the score for SVR improves appreciably. For example, over the best-3 languages overall, we get a correlation score of $r = 0.179$, which is slightly higher than CS_{L1+L2N}.

We further investigated the reason for getting very low and sometimes negative correlations with many of our target languages. We noted that about 24% of the German noun compounds in the dataset do not have entries in PanLex.

This contrasts with ENC where only one instance does not have an entry in PanLex, and EVPC where all VPCs have translations in at least one language in PanLex. We experimented with using string similarity scores in the case of such missing translations, as opposed to the strategy described in §3.2. The results for $CS_{SVR(L1+L2)}$ rose to $r = 0.269$, although this is still below the correlation for just using string similarity.

Our results on the GNC dataset using string similarity to measure the compositionality of the whole compound are competitive with the state-of-the-art results ($r = 0.45$) using a window-based distributional similarity approach over monolingual German data by adding the modifier and head predictions (Schulte im Walde et al. 2013).[15] Note, however, that their method used part-of-speech information and lemmatisation, where ours does not, in keeping with the language-independent philosophy of this research. Furthermore, as shown in §5, our string similarity measure can be substantially improved on GNC by using a multilingual dictionary with higher coverage for the expressions in this dataset.

7 Conclusion

This chapter presented an extension of two previous studies – Salehi & Cook (2013) and Salehi et al. (2014) – that proposed supervised and unsupervised methods to predict the compositionality of MWEs based on measures of string similarity between the translations of an MWE, and translations of its component words, into many target languages, and based on distributional similarity between an MWE and its component words, both in the original source language and under translation.

In experiments using the string similarity approach, we showed that information from translations into multiple target languages can be effectively combined to give improvements over using just a single target language. We also showed that string similarity measures which capture information about character sequences perform better than measures that do not. From the experiments on unsupervised approaches, we learned that languages of the same family as the source language cannot predict the compositionality of MWEs as well as the languages for which we have good translations coverage.

For distributional similarity, our experimental results showed that incorporating information from translations into target languages improved over using

[15] Additionally, Schulte im Walde et al. (2013) showed that their method achieves the state-of-the-art results ($r = 0.65$) in predicting the compositionality of each individual component within the compound.

distributional similarity in just the source language. Furthermore, we learned that there is a strong complementarity between approaches based on string and distributional similarity.

Abbreviations

MWE	multiword expression
ENC	English Noun Compound dataset of Reddy et al. (2011)
EVPC	English Verb-Particle Construction dataset of Bannard et al. (2003)
GNC	German Noun Compound dataset of Schulte im Walde et al. (2013)
LCS	longest common substring
LEV1	Levenshtein
LEV2	Levenshtein with substitution penalty
SW	Smith Waterman algorithm

Acknowledgements

We thank the anonymous reviewers for their valuable comments, and the editors for their time and effort in compiling this volume.

References

Acosta, Otavio, Aline Villavicencio & Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE' 11), 101–109. Association for Computational Linguistics.

Baldwin, Timothy. 2009. The hare and the tortoise: Speed and accuracy in translation retrieval. *Machine Translation* 23(4). 195–240.

Baldwin, Timothy, Colin James Bannard, Takaaki Tanaka & Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (MWE '03), 89–96. Association for Computational Linguistics. DOI:10.3115/1119282.1119294

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing, Second edition*, 267–292. Boca Raton: CRC Press.

Baldwin, Timothy, Jonathan Pool & Susan M. Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations* (COLING '10), 37–40. Association for Computational Linguistics.

Bannard, Colin James. 2006. *Acquiring phrasal lexicons from corpora*. University of Edinburgh dissertation.

Bannard, Colin James, Timothy Baldwin & Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (MWE '03 1), 65–72. Association for Computational Linguistics. DOI:10.3115/1119282.1119291

Biemann, Chris & Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Distributional Semantics and Compositionality Workshop* (DISCo 2011) *in conjunction with ACL 2011*, 21–28.

Carpuat, Marine & Mona 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*, 242–245. Association for Computational Linguistics. http://www.aclweb.org/anthology/N10-1029.

Dias, Gaël. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (MWE '03), 41–48. Association for Computational Linguistics.

Evert, Stefan & Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language* 19(4). 450–466.

Farahmand, Meghdad, Aaron Smith & 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* (MWE '15), 29–33. Association for Computational Linguistics.

Fazly, Afsaneh, Paul Cook & Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1). 61–103. http://aclweb.org/anthology/J09-1005.

Gomaa, Wael H & Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68(13). 13–18.

Hermann, Karl Moritz, Phil Blunsom & Stephen Pulman. 2012. An unsupervised ranking model for noun-noun compositionality. In *Proceedings of the First Joint*

*Conference on Lexical and Computational Semantics (*SEM)*, 132–141. June 7-8, 2012.

Katz, Graham. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (MWE '06), 12–19. Association for Computational Linguistics.

Kim, Su Nam & Timothy Baldwin. 2007. Detecting compositionality of English verb-particle constructions using semantic similarity. In *Proceedings of the 7th meeting of the Pacific association for computational linguistics* (PACLING 2007), 40–48.

Korkontzelos, Ioannis & Suresh Manandhar. 2009. Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference-Short papers*, 65–68. August 4, 2009.

Lapata, Mirella & Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics* (EACL-2003), 235–242.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* (ACL 1999), 317–324.

McCarthy, Diana, Bill Keller & John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on multiword expressions: Analysis, acquisition and treatment* (MWE '03), 73–80. Association for Computational Linguistics. DOI:10.3115/1119282.1119292

McCarthy, Diana, Sriram Venkatapathy & Aravind K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), 369–379. Association for Computational Linguistics. June 28–30, 2007.

Needleman, Saul B. & Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3). 443–453.

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

Pecina, Pavel. 2008. *Lexical association measures: Collocation extraction*. Prague, Czech Republic: Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic dissertation.

Pichotta, Karl & John DeNero. 2013. Identifying phrasal verbs using many bilingual corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2013). October 18-21, 2013.

Ramisch, Carlos. 2012. A generic framework for multiword expressions treatment: From acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, 61–66.

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

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Saif, Abdulgabbar, Mohd Juzaiddin Ab Aziz & Nazlia Omar. 2013. Measuring the compositionality of Arabic multiword expressions. In *Proceedings of the second international multi-conference on artificial intelligence technology*, 245–256.

Salehi, Bahar, Narjes Askarian & Afsaneh Fazly. 2012. Automatic identification of Persian light verb constructions. In *Proceedings of the 13th International Conference on Intelligent Text Processing Computational Linguistics* (CICLing '12)), 201–210.

Salehi, Bahar & Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics*, vol. 1 (* SEM 2013), 266–275. June 13-14, 2013.

Salehi, Bahar, Paul Cook & 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* (EACL 2014), 472–481. Gothenburg. http://aclweb.org/anthology/E/E14/E14-1050.pdf.

Schone, Patrick & Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing* (EMNLP 2001), 100–108. http://pascasarjana.mercubuana.ac.id/49/W01-0513.pdf.

Schulte im Walde, Sabine, Stefan Müller & Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, 255–265. Association for Computational Linguistics. June 13-14, 2013.

Schütze, Hinrich. 1997. *Ambiguity resolution in language learning*. Stanford, USA: CSLI Publications.

Smith, Temple F. & Michael S. Waterman. 1981. Identification of common molecular subsequences. *Molecular Biology* 147. 195–197.

Smola, Alex J. & Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14(3). 199–222.

Vecchi, Eva Maria, Marco Baroni & Roberto Zamparelli. 2011. Linear maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 1–9.

Venkatapathy, Sriram & Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (HLT-EMNLP 2005), 771–778.

von der Heide, Claudia & Susanne Borgwaldt. 2009. Assoziationen zu Unter, Basis und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th norddeutsches linguistisches Kolloquium*, 51–74.

Weeds, Julie Elizabeth. 2003. *Measures and applications of lexical distributional similarity*. University of Sussex dissertation.