Predicting the Compositionality of Multiword Expressions Using Translations in Multiple Languages

Bahar Salehi♣♡ **and Paul Cook**♡
♣ NICTA Victoria Research Laboratory
♡ Department of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
bsalehi@student.unimelb.edu.au, paulcook@unimelb.edu.au

Abstract

In this paper, we propose a simple, language-independent and highly effective method for predicting the degree of compositionality of multiword expressions (MWEs). We compare the translations of an MWE with the translations of its components, using a range of different languages and string similarity measures. We demonstrate the effectiveness of the method on two types of English MWEs: noun compounds and verb particle constructions. The results show that our approach is competitive with or superior to state-of-the-art methods over standard datasets.

1 Compositionality of MWEs

A multiword expression (MWE) is any combination of words with lexical, syntactic or semantic idiosyncrasy (Sag et al., 2002; Baldwin and Kim, 2009), in that the properties of the MWE are not predictable from the component words. For example, with *ad hoc*, the fact that neither *ad* nor *hoc* are standalone English words, makes *ad hoc* a lexically-idiosyncratic MWE; with *shoot the breeze*, on the other hand, we have semantic idiosyncrasy, as the meaning of "to chat" in usages such as *It was good to shoot the breeze with you*[1] cannot be predicted from the meanings of the component words *shoot* and *breeze*.

Semantic idiosyncrasy has been of particular interest to NLP researchers, with research on binary compositional/non-compositional MWE classification (Lin, 1999; Baldwin et al., 2003), or a three-way compositional/semi-compositional/non-compositional distinction (Fazly and Stevenson, 2007). There has also been research to suggest that MWEs span the entire continuum from full compositionality to full non-compositionality (McCarthy et al., 2003; Reddy et al., 2011).

Investigating the degree of MWE compositionality has been shown to have applications in information retrieval and machine translation (Acosta et al., 2011; Venkatapathy and Joshi, 2006). As an example of an information retrieval system, if we were looking for documents relating to *rat race* (meaning "an exhausting routine that leaves no time for relaxation"[2]), we would not be interested in documents on rodents. These results underline the need for methods for broad-coverage MWE compositionality prediction.

In this research, we investigate the possibility of using an MWE's translations in multiple languages to measure the degree of the MWE's compositionality, and investigate how literal the semantics of each component is within the MWE. We use Panlex to translate the MWE and its components, and compare the translations of the MWE with the translations of its components using string similarity measures. The greater the string similarity, the more compositional the MWE is.

Whereas past research on MWE compositionality has tended to be tailored to a specific MWE type (McCarthy et al., 2007; Kim and Baldwin, 2007; Fazly et al., 2009), our method is applicable to any MWE type in any language. Our experiments

[1] The example is taken from http://www.thefreedictionary.com

[2] This definition is from WordNet 3.1.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 266–275, Atlanta, Georgia, June 13-14, 2013. ©2013 Association for Computational Linguistics

over two English MWE types demonstrate that our method is competitive with state-of-the-art methods over standard datasets.

2 Related Work

Most previous work on measuring MWE compositionality makes use of lexical, syntactic or semantic properties of the MWE. One early study on MWE compositionality was Lin (1999), who claimed that the distribution of non-compositional MWEs (e.g. *shoot the breeze*) differs significantly from the distribution of expressions formed by substituting one of the components with a semantically similar word (e.g. *shoot the wind*). Unfortunately, the method tends to fall down in cases of high statistical idiosyncrasy (or "institutionalization"): consider *frying pan* which is compositional but distributionally very different to phrases produced through word-substitution such as *sauteing pan* or *frying plate*.

Some research has investigated the syntactic properties of MWEs, to detect their compositionality (Fazly et al., 2009; McCarthy et al., 2007). The assumption behind these methods is that non-compositional MWEs are more syntactically fixed than compositional MWEs. For example, *make a decision* can be passivised, but *shoot the breeze* cannot. One serious problem with syntax-based methods is their lack of generalization: each type of MWE has its own characteristics, and these characteristics differ from one language to another. Moreover, some MWEs (such as noun compounds) are not flexible syntactically, no matter whether they are compositional or non-compositional (Reddy et al., 2011).

Much of the recent work on MWEs focuses on their semantic properties, measuring the semantic similarity between the MWE and its components using different resources, such as WordNet (Kim and Baldwin, 2007) or distributional similarity relative to a corpus (e.g. based on Latent Semantic Analysis: Schone and Jurafsky (2001), Bannard et al. (2003), Reddy et al. (2011)). The size of the corpus is important in methods based on distributional similarity. Unfortunately, however, large corpora are not available for all languages.

Reddy et al. (2011) hypothesize that the number of common co-occurrences between a given MWE and its component words indicates the degree of compositionality of that MWE. First, the co-occurrences of a given MWE/word are considered as the values of a vector. They then measure the Cosine similarity between the vectors of the MWE and its components. Bannard et al. (2003) presented four methods to measure the compositionality of English verb particle constructions. Their best result is based on the previously-discussed method of Lin (1999) for measuring compositionality, but uses a more-general distributional similarity model to identify synonyms.

Recently, a few studies have investigated using parallel corpora to detect the degree of compositionality (Melamed, 1997; Moirón and Tiedemann, 2006; de Caseli et al., 2010; Salehi et al., 2012). The general approach is to word-align the source and target language sentences and analyse alignment patterns for MWEs (e.g. if the MWE is always aligned as a single "phrase", then it is a strong indicator of non-compositionality). de Caseli et al. (2010) consider non-compositional MWEs to be those candidates that align to the same target language unit, without decomposition into word alignments. Melamed (1997) suggests using mutual information to investigate how well the translation model predicts the distribution of words in the target text given the distribution of words in the source text. Moirón and Tiedemann (2006) show that entropy is a good indicator of compositionality, because word alignment models are often confused by non-compositional MWEs. However, this assumption does not always hold, especially when dealing with high-frequency non-compositional MWEs. Salehi et al. (2012) tried to solve this problem with high frequency MWEs by using word alignment in both directions.[3] They computed backward and forward entropy to try to remedy the problem with especially high-frequency phrases. However, their assumptions were not easily generalisable across languages, e.g., they assume that the relative frequency of a specific type of MWE (light verb constructions) in Persian is much greater than in English.

Although methods using bilingual corpora are intuitively appealing, they have a number of drawbacks. The first and the most important problem

[3] The IBM models (Brown et al., 1993), e.g., are not bidirectional, which means that the alignments are affected by the alignment direction.

is data: they need large-scale parallel bilingual corpora, which are available for relatively few language pairs. Second, since they use statistical measures, they are not suitable for measuring the compositionality of MWEs with low frequency. And finally, most experiments have been carried out on English paired with other European languages, and it is not clear whether the results translate across to other language pairs.

3 Resources

In this research, we use the translations of MWEs and their components to estimate the relative degree of compositionality of a MWE. There are several resources available to translate words into various languages such as Babelnet (Navigli and Ponzetto, 2010),[4] Wiktionary,[5] Panlex (Baldwin et al., 2010) and Google Translate.[6] As we are ideally after broad coverage over multiple languages and MWEs/component words in a given language, we exclude Babelnet and Wiktionary from our current research. Babelnet covers only six languages at the time of writing this paper, and in Wiktionary, because it is constantly being updated, words and MWEs do not have translations into the same languages. This leaves translation resources such as Panlex and Google Translate. However, after manually analysing the two resources for a range of MWEs, we decided not to use Google Translate for two reasons: (1) we consider the MWE out of context (i.e., we are working at the type level and do not consider the usage of the MWE in a particular sentence), and Google Translate tends to generate compositional translations of MWEs out of context; and (2) Google Translate provides only one translation for each component word/MWE. This left Panlex.

Panlex is an online translation database that is freely available. It contains lemmatized words and MWEs in a large variety of languages, with lemma-based (and less frequently sense-based) links between them. The database covers more than 1353 languages, and is made up of 12M lemmas and expressions. The translations are sourced from handmade electronic dictionaries, making it more accurate than translation dictionaries generated automatically, e.g. through word alignment. Usually there are several **direct translations** for a word/MWE from one language to another, as in translations which were extracted from electronic dictionaries. If there is no direct translation for a word/MWE in the database, we can translate indirectly via one or more pivot languages (**indirect translation**: Soderland et al. (2010)). For example, English *ivory tower* has direct translations in only 13 languages in Panlex, including French (*tour d'ivoire*) but not Esperanto. There is, however, a translation of *tour d'ivoire* into Esperanto (*ebura turo*), allowing us to infer an indirect translation between *ivory tower* and *ebura turo*.

4 Dataset

We evaluate our method over two datasets, as described below.

REDDY (Reddy et al., 2011): 90 English (binary) noun compounds (NCs), where the overall NC and each component word has been annotated for compositionality on a scale from 0 (non-compositional) to 5 (compositional). In order to avoid issues with polysemy, the annotators were presented with each NC in a sentential context. The authors tried to achieve a balance of compositional and non-compositional NCs: based on a threshold of 2.5, the dataset consists of 43 (48%) compositional NCs, 46 (51%) NCs with a compositional usage of the first component, and 54 (60%) NCs with a compositional usage of the second component.

BANNARD (Bannard, 2006): 160 English verb particle constructions (VPCs) were annotated for compositionality relative to each of the two component words (the verb and the particle). Each annotator was asked to annotate each of the verb and particle as `yes`, `no` or `don't know`. Based on the majority annotation, among the 160 VPCs, 122 (76%) are verb-compositional and 76 (48%) are particle-compositional.

We compute the proportion of `yes` tags to get the compositionality score. This dataset, unlike REDDY, does not include annotations for the compositionality of the whole VPC, and is also less balanced, containing more VPCs which are verb-compositional than verb-non-compositional.

[4] `http://lcl.uniroma1.it/babelnet/`

[5] `http://www.wiktionary.org/`

[6] `http://translate.google.com/`

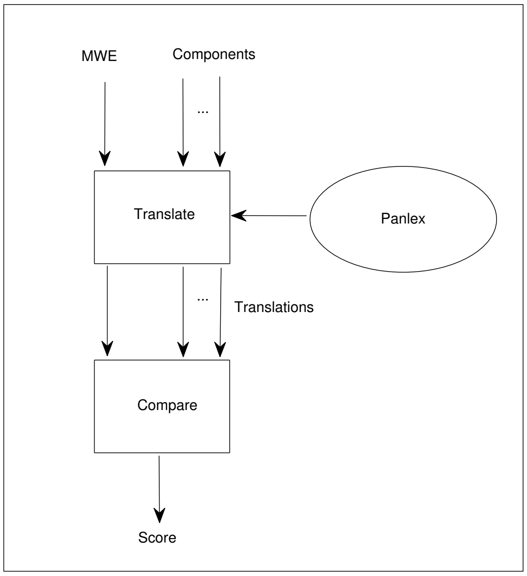

Figure 1: Schematic of our proposed method

5 Method

To predict the degree of compositionality of an MWE, we require a way to measure the semantic similarity of the MWE with its components. Our hypothesis is 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. Our second hypothesis is that the more languages we use as the basis for determining translation similarity between the MWE and its component words, the more accurately we will be able to estimate compositionality. Thus, rather than using just one translation language, we experiment with as many languages as possible.

Figure 1 provides a schematic outline of our method. The MWE and its components are translated using Panlex. Then, we compare the translation of the MWE with the translations of its components. In order to locate the translation of each component in the MWE translation, we use string similarity measures. The score shown in Figure 1 is derived from a given language. In Section 6, we show how to combine scores across multiple languages.

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

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

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).[7] 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.

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

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

language t as $component^t$.

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:

$$\frac{LongestCommonString(MWE^t, component^t)}{\min(len(MWE^t), len(component^t))}$$

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

$$1 - \frac{LEV1(MWE^t, component^t)}{\max(len(MWE^t), len(component^t))}$$

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:

$$1 - \frac{LEV2(MWE^t, component^t)}{len(MWE^t) + len(component^t)}$$

Smith Waterman (SW) This method is based on the Needleman-Wunsch algorithm,[8] and was developed to locally-align two protein sequences (Smith and 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 an in-built 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:

$$\frac{len(alignedSequence)}{\min(len(MWE^t), len(component^t))}$$

6 Computational Model

Given the scores calculated by the aforementioned string similarity measures between the translations for a given component word and the MWE, we need some way of combining scores across component words.[9] First, we measure the compositionality of each component within the MWE (s_1 and s_2):

$$\begin{aligned} s_1 &= f_1(sim_1(w_1, MWE), ..., sim_i(w_1, MWE)) \\ s_2 &= f_1(sim_1(w_2, MWE), ..., sim_i(w_2, MWE)) \end{aligned}$$

where sim is a string 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:

$$s_3 = f_2(s_1, s_2)$$

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 (2.5 for REDDY and 0.5 for BANNARD). Note that word order is not an issue in our method, as we calculate the similarity independently for each MWE component.

In this research, we consider simple functions for f_1 such as mean, median, product, min and max. f_2

[8] The Needleman-Wunsch (NW) algorithm, was designed to align two sequences of amino-acids (Needleman and 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.

[9] Note that in all experiments we only combine scores given by the same string similarity measure.

NC		
Language	Frequency	Family
Czech	100	Slavic
Norwegian	100	Germanic
Portuguese	100	Romance
Thai	99	Kam-thai
French	95	Romance
Chinese	94	Chinese
Dutch	93	Germanic
Romanian	91	Romance
Hindi	67	Indic
Russian	43	Slavic

Table 2: The 10 best languages for REDDY using LCS.

VPC:verb		
Language	Frequency	Family
Basque	100	Basque
Lithuanian	100	Baltic
Slovenian	100	Slavic
Hebrew	99	Semitic
Arabic	98	Semitic
Czech	95	Slavic
Slovak	92	Slavic
Latin	79	Italic
Tagalog	74	Austronesian
Polish	44	Slavic

Table 3: The 10 best languages for the verb component of BANNARD using LCS.

VPC:particle		
Language	Frequency	Family
French	100	Romance
Icelandic	100	Germanic
Thai	100	Kam-thai
Indonesian	92	Indonesian
Spanish	90	Romance
Tamil	87	Dravidian
Turkish	83	Turkic
Catalan	79	Romance
Occitan	76	Romance
Romanian	69	Romance

Table 4: The 10 best languages for the particle component of BANNARD using LCS.

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:

$$f_2(s_1, s_2) = \alpha s_1 + (1 - \alpha) s_2$$

Based on 3-fold cross validation, we chose $\alpha = 0.7$ for REDDY.[10]

Since we do not have judgements for the compositionality of the full VPC in BANNARD (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.

7 Language Selection

Our method is based on the translation of an MWE into many languages. In the first stage, we chose 54 languages for which relatively large corpora were available.[11] 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 (about 0.50 for BANNARD and about 0.80 for REDDY) 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.[12] 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.) In Tables 2, 3

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

[11] In future work, we intend to look at the distribution of translations of the given MWE and its components in corpora for many languages. The present method does not rely on the availability of large corpora.

[12] 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.

f_1	$sim()$	N1	N2	NC
Mean	SW	**0.541**	0.396	0.637
	LCS	0.525	**0.431**	**0.649**
	LEV1	0.405	0.200	0.523
	LEV2	0.481	0.263	0.577
Prod	SW	0.451	0.287	0.410
	LCS	0.430	0.233	0.434
	LEV1	0.299	0.128	0.311
	LEV2	0.294	0.188	0.364
Median	SW	0.443	0.334	0.544
	LCS	0.408	0.365	0.553
	LEV1	0.315	0.054	0.376
	LEV2	0.404	0.134	0.523
Min	SW	0.420	0.176	0.312
	LCS	0.347	0.225	0.307
	LEV1	0.362	0.310	0.248
	LEV2	0.386	0.345	0.338
Max	SW	0.371	0.408	0.345
	LCS	0.406	0.430	0.335
	LEV1	0.279	0.362	0.403
	LEV2	0.380	0.349	0.406

Table 5: Correlation on REDDY (NCs). N1, N2 and NC, are the first component of the noun compound, its second component, and the noun compound itself, respectively.

f_1	$sim()$	Verb	Particle
Mean	SW	0.369	**0.510**
	LCS	**0.406**	0.509
	LEV1	0.335	0.454
	LEV2	0.340	0.460
Prod	SW	0.315	0.316
	LCS	0.339	0.299
	LEV1	0.322	0.280
	LEV2	0.342	0.284
Median	SW	0.316	0.409
	LCS	0.352	0.423
	LEV1	0.295	0.387
	LEV2	0.309	0.368
Min	SW	0.262	0.210
	LCS	0.329	0.251
	LEV1	0.307	0.278
	LEV2	0.310	0.281
Max	SW	0.141	0.288
	LCS	0.268	0.299
	LEV1	0.145	0.450
	LEV2	0.170	0.398

Table 6: Correlation on BANNARD (VPC), based on the best-10 languages for the verb and particle individually

and 4, we show how often each language was selected in the top-10 languages over the combined 100 (10×10) folds of nested 10-fold cross validation, based on LCS.[13] The tables show that the selected languages were mostly consistent over the folds. The languages are a mixture of Romance, Germanic and languages from other families (based on Voegelin and Voegelin (1977)), with no standout language which performs well in all cases (indeed, no language occurs in all three tables). Additionally, there is nothing in common between the verb and the particle top-10 languages.

8 Results

As mentioned before, 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 in the test set for that fold. The scores are compared with human judgements using Pearson's correlation. The results are shown in Tables 5 and 6. Among the five functions we experimented with for f_1, Mean performs much more consistently than the others. Median is less prone to noise, and therefore performs better than Prod, Max and Min, but it is still worse than Mean.

For the most part, LCS and SW perform better than the other measures. There is little to separate these two methods, partly because they both look for a sequence of similar characters, unlike LEV1 and LEV2 which do not consider contiguity of match.

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 REDDY is 0.649, as compared to the best single-language correlation of 0.497 for Portuguese. On BANNARD, the best LCS result for the verb component is 0.406, as compared to the best single-

[13] Since our later results show that LCS and SW have higher results, we only show the best languages using LCS. These largely coincide with those for SW.

language correlation of 0.350 for Lithuanian.

Reddy et al. (2011) reported a correlation of 0.714 on REDDY. Our best correlation is 0.649. Note that Reddy et al. (2011) base their method on identification of MWEs in a corpus, thus requiring MWE-specific identification. Given that this has been shown to be difficult for MWE types including English VPCs (McCarthy et al., 2003; Baldwin, 2005), the fact that our method is as competitive as this is highly encouraging, especially when you consider that it can equally be applied to different types of MWEs in other languages. Moreover, the computational processing required by methods based on distributional similarity is greater than our method, as it does not require processing a large corpus.

Finally, we experimented with combining our method (STRINGSIM$_{\text{MEAN}}$) with a reimplementation of the method of Reddy et al. (2011), based on simple averaging, as detailed in Table 7. The results are higher than both component methods and the state-of-the-art for REDDY, demonstrating the complementarity between our proposed method and methods based on distributional similarity.

In Table 8, we compare our results (STRINGSIM$_{\text{MEAN}}$) with those of Bannard et al. (2003), who interpreted the dataset as a binary classification task. The dataset used in their study is a subset of BANNARD, containing 40 VPCs, of which 29 (72%) were verb compositional and 23 (57%) were particle compositional. By applying a threshold of 0.5 over the output of our regression model, we binarize the VPCs into the compositional and non-compositional classes. According to the results shown in Table 6, LCS is a better similarity measure for this task. Our proposed method has higher results than the best results of Bannard et al. (2003), in part due to their reliance on VPC identification, and the low recall on the task, as reported in the paper. Our proposed method does not rely on a corpus or MWE identification.

9 Error Analysis

We analyse items in REDDY which have a high 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 REDDY. 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")[14] 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.

10 Conclusion and Future Work

In this study, we proposed a method to predict MWE compositionality based on the translation of the MWE and its component words into multiple languages. We used string similarity measures between the translations of the MWE and each of its components to predict the relative degree of compositionality. Among the four similarity measures that we experimented with, LCS and SW were found to be superior to edit distance-based methods. Our best results were found to be competitive with state-of-the-art results using vector-based approaches, and were also shown to complement state-of-the-art methods.

In future work, we are interested in investigating whether alternative ways of combining our proposed method with vector-based models can lead to further enhancements in results. These models could be especially effective when comparing translations which are roughly synonymous but not string-wise similar.

Acknowledgments

We would like to thank Timothy Baldwin, Su Nam Kim, and the anonymous reviewers for their valuable comments and suggestions.

[14] This definition is from Wordnet 3.1.

$sim()$	StringSim$_{\text{MEAN}}$	StringSim$_{\text{MEAN}}$ + Reddy et al.
SW	0.637	0.735
LCS	**0.649**	**0.742**
LEV1	0.523	0.724
LEV2	0.577	0.726

Table 7: Correlation after combining Reddy et al.'s method and our method with Mean for f_1 (StringSim$_{\text{MEAN}}$). The correlation using Reddy et al.'s method is 0.714.

Method	Precision	Recall	F-score ($\beta = 1$)	Accuracy
Bannard et al. (2003)	0.608	0.666	0.636	0.600
StringSim$_{\text{MEAN}}$	0.862	0.718	0.774	0.693

Table 8: Results for the classification task. StringSim$_{\text{MEAN}}$ is our method using Mean for f_1

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Otavio Costa Acosta, Aline Villavicencio, and Viviane P Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, pages 101–109.

Timothy Baldwin and Su Nam Kim. 2009. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, Boca Raton, USA, 2nd edition.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Timothy Baldwin, Jonathan Pool, and 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*, pages 37–40.

Timothy Baldwin. 2005. The deep lexical acquisition of English verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

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

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 65–72.

Colin James Bannard. 2006. *Acquiring Phrasal Lexicons from Corpora*. Ph.D. thesis, University of Edinburgh.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Helena Medeiros de Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1):59–77.

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16.

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

Su Nam Kim and 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)*, pages 40–48.

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

Diana McCarthy, Bill Keller, and 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-Volume 18*, pages 73–80.

Diana McCarthy, Sriram Venkatapathy, and 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)*, pages 369–379.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the Fifth Workshop on Very Large Corpora*. EMNLP.

Begona Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-wordexpressions in a multilingual context*, pages 33–40.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Saul B Needleman and 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.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP*, pages 210–218.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing Computational Linguistics (CICLing-2002)*, pages 189–206. Springer.

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

Patrick Schone and Dan 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)*, pages 100–108.

TF Smith and MS Waterman. 1981. Identification of common molecular subsequences. *Molecular Biology*, 147:195–197.

Stephen Soderland, Oren Etzioni, Daniel S Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, Jeff Bilmes, et al. 2010. Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174(9):619–637.

Sriram Venkatapathy and Aravind K Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20–27.

Charles Frederick Voegelin and Florence Marie Voegelin. 1977. *Classification and index of the world's languages*, volume 4. Elsevier Science Ltd.