DE GRUYTER MOUTON DOI 10.1515/phras-2016-0008 — YoP 2016; 7: 149–166

Ruslan Mitkov

Computational Phraseology light: automatic translation of multiword expressions without translation resources

Abstract: This paper describes the first phase of a project whose ultimate goal is the implementation of a practical tool to support the work of language learners and translators by automatically identifying multiword expressions (MWEs) and retrieving their translations for any pair of languages. The task of translating multiword expressions is viewed as a two-stage process. The first stage is the extraction of MWEs in each of the languages; the second stage is a matching procedure for the extracted MWEs in each language which proposes the translation equivalents.

This project pursues the development of a knowledge-poor approach for any pair of languages which does not depend on translation resources such as dictionaries, translation memories or parallel corpora which can be time consuming to develop or difficult to acquire, being expensive or proprietary. In line with this philosophy, the methodology developed does not rely on any dictionaries or parallel corpora, nor does it use any (bilingual) grammars. The only information comes from comparable corpora, inexpensively compiled. The first proof-of-concept stage of this project covers English and Spanish and focuses on a particular subclass of MWEs: verb-noun expressions (collocations) such as *take advantage*, *make sense*, *prestar atención* and *tener derecho*.

The choice of genre was determined by the fact that newswire is a widespread genre and available in different languages. An additional motivation was the fact that the methodology was developed as language independent with the objective of applying it to and testing it for different languages. The ACCURAT toolkit (Pinnis et al. 2012; Skadina et al. 2012; Su and Babych 2012a) was employed to compile automatically the comparable corpora and documents only above a specific threshold were considered for inclusion. More specifically, only pairs of English and Spanish documents with comparability score (cosine similarity) higher 0.45 were extracted.[1]

1 However, see section 6 which discusses experiments with different comparability scores.

Ruslan Mitkov, Research Institute in Information and Language Processing, University of Wolverhampton, R.Mitkov@wlv.ac.uk

Statistical association measures were employed to quantify the strength of the relationship between two words and to propose that a combination of a verb and a noun above a specific threshold would be a (candidate for) multiword expression. This study focused on and compared four popular and established measures along with frequency: Log-likelihood ratio, T-Score, Log Dice and Salience.

This project follows the distributional similarity premise which stipulates that translation equivalents share common words in their contexts and this applies also to multiword expressions. The Vector Space Model is traditionally used to represent words with their co-occurrences and to measure similarity. The vector representation for any word is constructed from the statistics of the occurrences of that word with other specific/context words in a corpus of texts. In this study, the word2vec method (Mikolov et al. 2013) was employed. Mikolov et al.'s method utilises patterns of word co-occurrences within a small window to predict similarities among words.

Evaluation results are reported for both extracting MWEs and their automatic translation. A finding of the evaluation worth mentioning is that the size of the comparable corpora is more important for the performance of automatic translation of MWEs than the similarity between them as long as the comparable corpora used are of minimal similarity.

Keywords: multiword expressions (MWEs), extraction of MWEs, translation of MWEs, comparable corpora, association measures

1 Rationale

Multiword expressions (MWEs) represent a major linguistic phenomenon and are important and inseparable part of both general and specialised languages with MWEs reported to be of the same order as single words (Jackendoff, 2007). One significant impediment to all researchers focusing on this pervasive phenomenon or language users in general, is that the coverage of MWEs in terms of dictionaries or lexicographical resources is far from ideal. With particular reference to specialised languages, new (multiword) terms are coined on a daily basis and dictionaries or other resources however up-to-date they are, cannot keep up with the speed of emergence of new terms. Therefore, the best way forward is to develop tools for automatically extracting terms and their translations from corpora.

Arguably parallel corpora would be the best source for extracting (multiword) terms along with their translations. However, parallel corpora are not widely

available and possibly do not cover all domains. An alternative and more promising approach would be to use comparable corpora for this task as comparable corpora can be compiled from the web in a comparatively straightforward way, making use of available purpose-built tools.

This paper describes the first phase of a project whose ultimate goal is the implementation of a practical tool to support the work of language learners and translators by automatically identifying multiword expressions (MWEs) and retrieving their translations for any pair of languages. The philosophy of this project lies in the development of a knowledge-poor approach which does not depend on translation resources such as dictionaries, translation memories or parallel corpora which can be time consuming to develop or difficult to acquire, being expensive or proprietary. In line with this vision, the methodology developed does not rely on any dictionaries or parallel corpora, nor does it use any (bilingual) grammars. The only supporting information comes from comparable corpora, inexpensively compiled. The first proof-of-concept stage of this project covers English and Spanish and focuses on a particular subclass of MWEs: verb-noun expressions (collocations) such as *take advantage*, *make sense*, *prestar atención* and *tener derecho*.

In this study the task of translating multiword expressions is viewed as a two-stage process. The first stage is the extraction of MWEs in each of the languages; the second stage is a matching procedure for the extracted MWEs in each language which proposes the translation equivalents.

The rest of the paper is structured as follows. First, the study is contextualised by outlining related work. The comparable corpora which were compiled as the key resource for the project are then presented. Next, the task of annotation and the related annotation agreement are discussed. This is followed by a description of the methodology developed to extract candidate MWEs and evaluation of this approach. After this, the methodology developed to identify translation equivalents of the extracted MWEs is presented and evaluated. The paper concludes by summarising the main outcomes of the study.

2 Related work

There is a large body of work describing different properties of various MWEs (e.g. Baldwin and Kim 2010). Different approaches have been proposed for identifying MWEs in one language (Baldwin and Villavicencio 2002). However, the cross-lingual analysis of these expressions and automatic extraction of their translation equivalents is still an under-researched topic (Bouamor et al. 2012).

Most work on translation of MWEs consists first of identifying monolingual MWE candidates and then applying alignment methods to establish bilingual correspondence. Previous work covers the extraction of MWEs from dictionaries, including our own work (Mendoza et al. 2013). As aforementioned, such dictionaries do not have a wide coverage so a better alternative would be to develop methodology for finding translations of MWEs automatically. Bouamor et al. (2012) use distributional models to align MWEs in order to improve the performance of machine translation systems. However, their method relies exclusively on a sentence-aligned corpus. Rapp and Sharoff (2014) also investigate the use of the patterns of word-co-occurrence across languages to extract MWE translations. While their approach delivers good results in establishing translations of single words, they do not report good results for MWEs. In the case of single words, however, their methodology only covers words which are keywords according to frequency.

3 Comparable corpora as the only resource for translating multiword expressions

The objective of this study was to experiment with the most productive and widely used verbs in Verb + Noun combinations. To this end for each language the 20 most frequent verbs occurring before nouns were identified. In order to obtain an accurate and representative picture as to which are the most frequent verbs preceding nouns, large corpora in both languages were chosen. For English this study benefited from the large and representative the British National Corpus, whereas for Spanish the Spanish Web corpus supplied with the SketchEngine, was used. For English the following verbs were retrieved as the most frequent ones: *take*, *have*, *make*, *give*, *use*, *get*, *raise*, *hold*, *become*, *find*, *keep*, *pay*, *call*, *show*, *play*, *put*, *receive*, *cause* and *lose, offer*.[2] The Spanish list consisted of the following verbs: *tener*, *dar*, *ver*, *hacer*, *formar*, *poner*, *tomar*, *recibir*, *realizar*, *pedir*, *crear*, *encontrar*, *ofrecer*, *buscar*, *citar*, *existir*, *presentar*, *dejar*, *llevar* and *mostrar*. The above corpora were also used for computing the frequency and association measures for the identified patterns verb-nouns (see below). Recall that in this study the task of translating multiword expressions is regarded as a two stage process. The first

2 In this experiment the verb *say* was removed from the data as a number of invalid occurrences *say + SUBJECT* (such as *said police*, *say reporters*) were extracted. The verbs *see*, *need* and *include* were equally excluded from the experiment as they did not form part of any MWEs according to annotators. For the same reason, the following Spanish verbs were removed *decir*, *ver*, *recibir*, *incluir* and *existir*.

stage is the extraction of MWEs in each of the languages; the second stage is a matching procedure for the extracted MWEs which aims at proposing translation equivalents. This proof-of-concept study does not employ bilingual resources but only comparable corpora; to this end comparable data featuring MWEs had to be collected for both English and Spanish (see below).

Most approaches to bilingual term or word extraction operate on parallel corpora which are then aligned and while the word-to-word or phrase alignment does not normally enjoy high performance, the reality is that parallel corpora are not very easy to get and as a result, not readily available. Often parallel data are proprietary as translation memories but even the available parallel corpora are neither large enough for current NLP applications, nor cover all domains. Hence a more viable and pragmatic approach would be to build large quantities of comparable data from the practically unlimited texts on the web. In this study the view was taken that compiling large comparable data for the explicit need of the project, would be the best way forward. The choice of genre was determined by the fact that newswire is a widespread genre and available in different languages. An additional motivation was the fact that the methodology was developed as language independent and is to be tested for different languages. This proof-of-concept first study covers English and Spanish.

The ACCURAT toolkit (Pinnis et al. 2012; Skadina et al. 2012; Su and Babych 2012a), a product of the project with the same name, was employed to compile automatically the comparable corpora for this study. The original goal of ACCURAT was to find, analyse and evaluate novel methods which exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for under-resourced languages and narrow domains. This toolkit was used in this study to download news articles from the web from the RSS feeds of ABC news, Yahoo news, CNN news, Sport news and Euronews in both Spanish and English were collected. In addition, RSS feeds of Ultimahora and Europapress for Spanish were also added to ensure the Spanish date is more balanced.

The downloaded data from online news (1.5 GB) consisted of 200,000 documents in English and 112,000 documents in Spanish. These documents were classified with a view to building a corpus of English texts and another of Spanish texts which are comparable. The criteria for comparability in this project were based on the notion of similarity. Each monolingual corpus was designed to feature documents paired with documents in the other language in terms of the similarity between them. Similarity was automatically computed with the help of the ACCURAT tool, *DictMetric*, which compares documents by employing cosine similarity. More specifically, DictMetric converts text into index vectors and then computes a comparability score of document pairs by applying cosine similarity

measure on the index vectors. This tool was developed to assist the extraction of translation equivalents in Machine Translation. In order to measure the comparability of two documents in different languages, one of the documents is translated in the language of the other. DictMetric translates non-English texts into English by using lexical mapping from the available GIZA++ based bilingual dictionaries.

In this study comparable documents only above a specific threshold were considered for inclusion. More specifically, only pairs of English and Spanish documents with comparability score (cosine similarity) higher 0.45 were extracted.[3] This resulted in 16,436 English documents (around 11,000,000 tokens) and 11,468 Spanish documents (6,000,000 tokens). Each English document was paired with at least one Spanish document; equally, there was at least one paired English document for every Spanish document. Due to the higher number of English documents, the overall number of the English documents paired with Spanish documents was higher than the overall number of Spanish documents paired with English ones. The average cosine similarity of all paired documents was 0.54 which was regarded as a comparability measure between English corpus and the Spanish corpus.

4 Annotation and interannotation agreement

The English BNC corpus and the Spanish Web corpus come with information on POS tags and lemmas for all their words as part of the SketchEngine. On the basis of this information, all possible patterns *verb-noun* (with *verb* being the lemmatised form of this verb) of the aforementioned 20 most frequent verbs, were returned as candidate MWEs.[4]

The so extracted candidates were annotated by human annotators as being MWEs or non-MWEs (candidates with less than 5 occurrences in English and less than 4 in Spanish were ignored in line with the observation that low frequency does not provide meaningful statistical information). The annotators were native speakers (English or Spanish) with background in linguistics. With a view to measuring the interannotator agreement, for each language there was a main annotator and a second annotator who annotated a sample of the data. Once the

3 See also the experiments with different similarity scores in section 6.

4 For English the most frequent candidate pairs were *take place* (10,753 occurrences), *take part*, *give rise*, *take advantage* and *make sense*. Other examples of extracted pairs included *have children*, *become president*, *pay tax*, *pay attention* and *say good-bye*. For Spanish the most frequent candidate pairs were *dar cuenta*, *formar parte*, *tener lugar* and *hacer falta*. Further examples of pairs extracted in Spanish include *prestar atención*, *tomar decision*, *tener hijo* and *tener más*.

annotation task was completed, the Kappa value and the value of observed agreement were computed on these samples (see below).

Annotators were asked to mark up candidates as multiword expressions (MWEs) only if they exhibit a sufficient degree of idiomaticity or in other words, which do not convey literal meaning in that the verb is delexicalised. In addition, a candidate would be annotated as MWE only if it made sense (the program would return 'candidates' which were not meaningful as MWEs due to the preprocessing errors). Following this rule, annotators were instructed to mark up *take a decision* or *tomar decisión* as MWEs as in this construction the light verb *take* (tomar) does not carry literal meaning and refers to the process of arriving at a specific decision (possible, after deliberation). By the same token, *take break* and *have break* would be marked as MWEs and so would be *hacer descanso* and *tomar descanso* (the verbs are not used in their literal senses). Equally, *make use* and *hacer uso* as well as *pay attention* and *prestar atención* would be annotated as MWEs. On the other hand, *have a coffee* or *tomar café* would not be marked as MWEs as in some sense these expressions could be interpreted literally. Further, *pay tax*, *become president* or *have visitors* would not be annotated as MWEs (literal meaning) and neither would be *get people* or *dar Silva* in Spanish (examples of extracted candidates which are not meaningful).

The annotation task triggered interesting discussion with often opposite views expressed by the annotators. By way of example, for the candidate *say good-bye* the view prevailed that it was a multiword expression as this expression goes beyond its literal meaning of 'saying' and describes the 'farewell act'.

This annotation exercise consisted of marking up lists of candidate MWEs extracted by the program. At this stage annotation in context would have been unrealistic due to the prohibitively time-consuming and labour-intensive nature of the task. However, a more comprehensive and linguistically valid exercise of marking up words in context will follow in the very near future.

Given the above constraint, certain candidate expressions were ambiguous in that they could can act as MWEs or not as MWEs depending on the context (e.g. *tener hijo(s)* which could mean *have child*/*children* or *give birth*) and hence annotators were asked to assign the tag 2 to such cases. On the other hand, candidates which were believed to be always MWEs, were allocated the tag 1; those which were deemed never be multiword expressions, were marked up as 0.

For English the main annotator marked up 2,389 candidates; a random sample of 604 candidates from these (about 25%) were annotated by a second annotator. Kappa statistic was applied to measure the agreement which in this case was established to be 0.6.

For Spanish the main annotator marked up 2,958 candidates; a sample which was annotated by a second annotator and on which interannotator agreement

was computed, represented about 20% of the whole set, amounting to 580 candidates. The Kappa measure computed a 0.44 level of agreement for Spanish.

The Kappa measure was computed for each verb separately. English verbs with Kappa value less than 0.4 and Spanish verbs the Kappa value is less than 0.3 were not considered. As a result, this experiment included 9 English verbs (*take*, *give*, *make*, *pay*, *lose*, *find*, *keep*, *put*, *hold*) and 6 Spanish verbs (*tener*, *dar*, *tomar*, *formar*, *crear*, *presentar*). Since the data was considerably skewed and the Kappa measure is known to be affected by the skewed data, the observed agreement value (the percentage of the cases where the annotators agree and mark up candidates with the same tag) was also computed and was 0.85 for English and 0.7 for Spanish.

Table 1: below provides details on the annotation data and illustrates the interannotator agreement figures.

Table 1: Annotation and interannotator agreement figures

	Candidate expressions	Sample tagged by the 2nd annotators	Kappa	Observed agreement
English	2,389	604	0.6	0.85
Spanish	2,958	580	0.44	0.70

5 Extraction of multiword expressions: methodology and evaluation

The selected *verb-noun* expressions, which are annotated with tags 1 and 0, were ranked according to their frequency and/or the association measures computed from BNC for English and Spanish Web Corpus for Spanish. The methodology is underpinned by the premise that patterns with a frequency or association measure above a specific threshold, are good candidates for being MWEs. In this study different statistical association measures were employed to quantify the strength of the relationship between two variables, which are words in this context (see below); these associate measures model the likelihood that a combination of a verb and a noun above a specific threshold is a (candidate for) multiword expression through the frequency of occurrence and strength of association.

While a high number of association measures have been proposed and employed in Natural Language Processing and Lexicography (Pecina 2005 lists 57 association measures), this study employs and compares four popular and established measures: Log-likelihood ratio, T-Score, Log Dice and Salience.

The *log-likelihood ratio* (Dunning, 1993) has been widely used for extracting collocations. It enables direct comparison of the significance of common and rare phenomena and works well with both large and small sizes of corpora. The *T-Score measure* or *t-test* (Krenn and Evert, 2001) addresses the low-frequency bias in the popular pointwise mutual information measure (Church and Hanks, 1990). *Log Dice*, which is the logarithmic form of the Dice formula (Dice 1945) is reported as one of the best performing association measures in MWE induction in Schone and Jurafsky (2001). Finally, *Salience* (also referred as MI.log – **f** in Kilgariff et al. 2004) was a recently proposed association measure for collocation extraction and is a combination of several statistical measures. More specifically, Silence is an adjustment to point-wise mutual information and estimated as the product of mutual information and log frequency. The above measures employed in the experiments were computed via the SketchEngine suite (Kilgariff et al. 2004) which includes all above measures.

Frequency of occurrence has been widely used to retrieve collocations, including in our previous work (Mendoza Rivera et al., 2013). In this study frequency was also experimented with a view to establishing to what extent it can be helpful in ranking a specific type of MWEs (in this case patterns *verb + noun*). Frequency was employed as an informed baseline and evaluated comparing its performance with that of other ranking measures.

Expressions were ranked according to frequency and setting thresholds at different points, the number of true MWEs above the threshold (true positives) and the number of non-MWEs above the threshold (false positives) were counted. In addition, with a view to computing Precision-Recall curves, the false negatives (MWEs under the threshold) and the true negatives (non-MWEs below the threshold) were counted as well.

The lack of balance in the evaluation set due to the higher number of non-MWEs compared to MWEs, especially among low-frequency expressions, is conjectured as a valid reason that would cause comparatively low precision values. In order to smoothen this negative effect, further experiments were conducted with more balanced data where low-frequency expressions (with frequencies lower than 10) are excluded. As a result, the maximum F-measure value (which is a combination of Precision and Recall) for ranking the English expressions according to Frequency, went up from 0.54 to 0.64. These results are based on an evaluation set of 906 English and 1,575 Spanish expressions, excluding the expressions with frequencies lower than 10.

The retrieval performance is reported for each measure on the basis of the ranking Verb + Noun pairs according to the association measures. The approach is rank-based, in other words the candidate expressions with higher ranks are deemed to be more likely MWEs. The retrieval performance is estimated through

Interpolated Precision (IP) curves. They reflect the efficiency of a measure in ranking the relevant items (in this study, MWEs) before the irrelevant ones.

While presenting the full precision-recall curve could be informative, a numerical summary provided by the 11-point Interpolated Precision curve (Manning et al. 2008) which compares the different ranking approaches, could be equally helpful. To this end, candidates are ranked according to the score that a measure assigns to them, and the interpolated precision at the 11 recall values of 0, 10%, ..., 100% is calculated. As detailed in Manning et al. (2008), the interpolated precision at a certain recall level, *r*, is defined as the highest precision found for any recall level r′ >= r (see Manning et al. 2008). A composite precision-recall curve showing 11 points can then be graphed. Following the graph, one can see what would be the precision of every approach for different levels of recall. The results of these graphs which compare the different rankings are presented in Figure 1 for English and Figure 2 for Spanish.

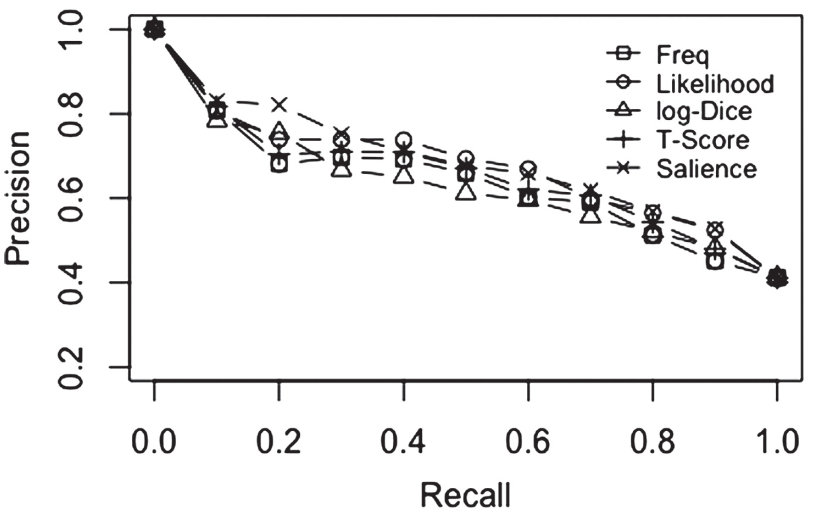

Figure 1: Comparing the performance of different rankings according to 11-point Interpolated Average Precision for English Expressions

As seen from Figure 1, all statistical measures tend to rank (most of) the MWEs higher than/before Verb Noun sequences which are not MWEs with the best precisions being among the higher ranks (lower recalls). In other words, when recall is 10% which corresponds to the extraction of 10% of the truly

positive MWEs (the expressions annotated as MWEs), the precision is very high, since most of the extracted expressions are truly positive. It is worth noting that T-Score works comparably to Frequency in the performance of ranking expressions; Log-likelihood does slightly better with better precisions for high ranked expressions and Salience appears to perform comparably or sometimes better than Log-likelihood in ranking MWEs higher than non-MWEs.

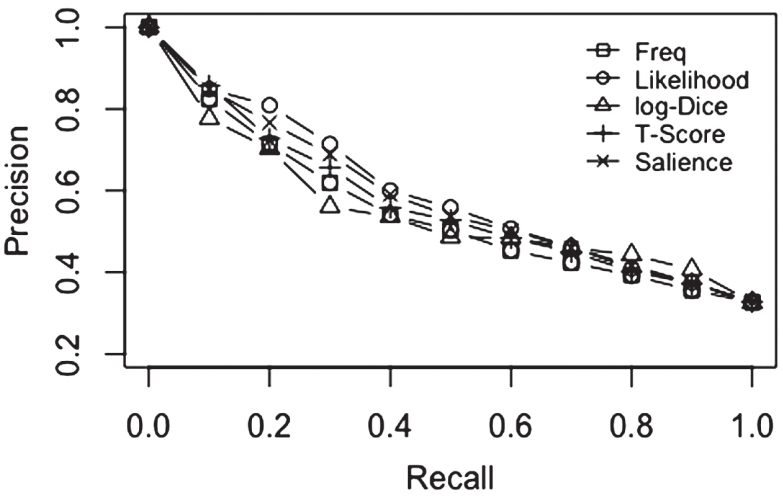

Figure 2: 11p-IAP evaluation of different ranking measures for candidates in Spanish

The results in Figure 2 show that the performance of the different association measures when ranking candidates in Spanish, does not necessarily follow the trend established for ranking candidates in English. In particular, Salience does not fare very well for Spanish expressions but Frequency and Likelihood do reasonably in ranking expressions to be proposed as MWEs.

The classification accuracy of the above statistical measures was also computed by setting the average of all the values for each target measure as a threshold. Specifically, for each measure the average of all its values for all candidate expressions was computed, and set as a threshold. Then each expression with a value higher than a threshold is considered as an MWE and each with a value lower than the threshold is classed as non-MWE according to the target association measure. The accuracies are computed and compared to the annotation.

Table 2 displays the accuracies for setting the average as a threshold for different measures.

Table 2: Comparing the classification performance of different ranking methods of English and Spanish candidate expressions (average as threshold)

	Accuracy	
	English	**Spanish**
Frequency	0.63	0.71
Log-Likelihood	0.65	0.73
Salience	0.70	0.67
T-Score	0.69	0.69
Log-Dice	0.66	0.65

As can be seen from Table 2, Salience work best for classifying English candidate expressions. For Spanish expressions, the best results are observed for Likelihood which outperforms Frequency and which in turn does better than Salience. The different trends between English and Spanish can by justified by the evaluation results reported by Evert and Krenn (2005) where they concluded that one evaluation measure (likelihood in their case) may not always be the best to extract collocations. They showed that the practical usefulness of individual association measures depends on different issues such as the size of the corpora, the tools used for syntactic preprocessing and candidate extraction, and the amount of low-frequency data excluded by setting a frequency threshold.

6 Translation of multiword expresions: methodology and evaluation

The second stage of the automatic translation of MWEs is a matching procedure for the extracted MWEs in each language which proposes the translation equivalents. The methodology pursued in this study employs distributional similarity across bilingual corpora. The terms 'equivalent expressions' or 'equivalents' are used here to refer to expressions which are translations of each other across languages. A fundamental premise is that equivalent expressions are expected to appear in the same or similar contexts across languages.

Corpus-based distributional similarity has already offered promising results in the discovery of translationally equivalent words/terms in a bilingual scenario.

Most of the work in this line (Rapp 1999; Fung and McKeown 1997; Bouamor et al. 2012)), including our own work (Pekar et al. 2006), covers single words and not multiword expressions. According to the distributional similarity premise, translation equivalents share common words in their contexts and this applies also to multiword expressions. For example, among the most frequent words in the context around the Spanish MWE *tener lugar* are *terremoto*, *disturbio*, *periodo*, *seminario*, *elecciones*, *debate*, *sesiones públicas*, *evaluación*, *actividad* whose translations in English are *earthquake*, *riot*, *period*, *seminar*, *elections*, *debate*, *public meetings*, *evaluation*, *activity*, respectively. These English words are among the most frequents words in the context of the English MWE *take place*. In Table 3 every value represents the pointwise mutual information (PMI) of the occurrences of the target term around the corresponding context word. On the basis of these values, the cosine vector similarity between the vectors for *tener lugar* and *take place* is computed as 0.71.

Table 3: Very simplified context vectors for the candidates *tener lugar* in Spanish and *take place* in English

	Riot/ Disturbio	Earthquake/ Terremoto	Justice/ Justicia	Period/ Periodo	Problem/ Problema	...	Reality/ Realidad
Tener lugar	0.97	0.74	0.20	0.51	0.45	...	−0.35
Take place	0.92	0.62	0.19	0.51	0.44	...	−0.36

The translation matching premise is that if MWE_{EN} is a candidate translation of MWE_{ES} then the words with which MWE_{EN} co-occurs in a specific window are the translations of the words with which MWE_{ES} co-occurs within a window of the same size. The implemented approach represents candidates with bilingual contexts as vectors (see below) and matches translation equivalents between two languages (in this case English and Spanish).

The Vector Space Model (VSM) is traditionally made use of to represent words with their co-occurrences and to measure similarity. The vector representation for any word is constructed from the statistics of the occurrences of that word with other specific/context words in a corpus of texts. More specifically in this study, the word2vec method (Mikolov et al. 2013) was employed. Mikolov et al.'s method utilises patterns of word co-occurrences within a small window to predict similarities among words. The underlying idea is to represent words as dense vectors (also referred to as word embeddings) which are learned by neural networks (Levy and Goldberg 2014). This new word embedding approach uses a neural network to learn low-dimensional word vectors from raw (monolingual) text. The standard

implementation of Word2Vec constructs bag-of-words contexts for all single-word terms which appear in a training corpus. This increasingly popular model was adapted to this study by (i) regarding sequences of words as single units and (ii) defining bilingual contexts as set of translation pairs which are obtained by an automatically learned Machine Translation system (see below). While the generalised version of word2vec was originally used to extract dependency-based word embeddings (Levy and Goldberg 2014), in this study it was adapted in a straightforward manner to the task of vector construction for multi-word collocations using the bilingual contexts.

In order to construct vectors for the candidate English and Spanish expressions, a seed list of paired context words (bilingual context pairs) was derived from the word alignments obtained after the application of GIZA++ on the English-Spanish Europarl parallel corpus (Koehn 2005). Only pairs of frequent nouns with more than 50 occurrences in Europarl and with alignment probability higher than 0.2, were considered. This resulted in 4700 bilingual contexts.

Both the English and Spanish components of Europarl and the comparable corpora compiled for this study, were used to train modes in order to learn vectors. All English and Spanish verb combinations (unigrams, bigrams and trigrams) were indexed according to their occurrences along with the context word pairs. For every candidate expression, words that appear in bilingual context pairs within a window of 10 words were extracted. Next the *word2vecf* software was used to train the vectors on the indexed data; translations were mined for in both directions: Spanish to English and English to Spanish.

Recall that translations for expressions Verb + Noun were sought. The premise was that for most such expressions, the translation is either a verb (unigram), verb + noun (bigram) or verb + noun with an intervening word such as determiner or an adjective. Therefore, as candidate translations all unigram verbs, bigram verb + noun and trigrams verb + noun with an intervening word were considered. For every expression from the source language the goal was to identify the most similar expression in the target language.

A native speaker was asked to examine and rate the top-ranked translations identified by each of the methods for each expression. The annotator was instructed to give a score of 1 if there was at least one correct translation in the top-ranked list, and a score of 0, otherwise.

A simple distributional similarity approach based on the Jaccard similarity coefficient was implemented as a baseline. Given two expressions from two different languages, their similarity was computed by comparing their corresponding sets of bilingual context pairs within a window of 10 words from the comparable corpora.

As aforementioned, similarity measures were used to rank the candidate translations for each candidate expression. Experiments with different similarity thresholds were conducted which as expected, reflected the trade-off between coverage (recall) and accuracy (precision). The results for each method associated with three different threshold values are reported below. These thresholds have been set to account for three different coverage levels: 20%, 50% and 80%.

Experiments were conducted with threshold values for the cross-lingual similarities between English and Spanish documents which served as the basis for pairing documents. These experiments sought to establish the impact of different thresholds which leads to different sizes of paired documents (comparable data). In this particular study, the performance of finding translation equivalents by applying the methodology to these four different sizes of corpora by setting the threshold to four different similarity values 0.6, 0.45, 0.3 and 0.2 (referred to as *CC limit 0.6*, *CC limit 0.45*, *CC limit 0.3*, *CC limit 0.2*, respectively in Tables 4 and 5), was evaluated.

A rationale behind this approach was to establish what plays a more influential role in this study: whether it is the size of the data or the quality of the data. Table 4 shows the performance figures for accuracy and coverage for the task of identifying translations of Spanish expressions; Table 5 displays the same results for obtaining translations of English expressions. The results are reported for both baseline and the word2vec method on the basis of different sizes of comparable corpora of paired documents.

As expected, at low recall the precision is high and vice-versa. Note that when recall is low, many expressions do not have translation equivalents. However, for those expressions which do have translation equivalents, the precision of identifying them is quite good – this is also the case for a simple baseline method. For all cases of higher coverage however, the word2vec method fares better and is more consistent at translating into any of the languages.

Table 4: Accuracies of finding translation equivalents for Spanish expressions

Coverage		20%	50%	80%
Baseline	CC limit 0.6	14%	–	–
	CC limit 0.45	38%	17%	9%
	CC limit 0.3	86%	28%	13%
	CC limit 0.2	79%	31%	13%
Word2vec	CC limit 0.6	32%	–	–
	CC limit 0.45	48%	38%	25%
	CC limit 0.3	49%	46%	34%
	CC limit 0.2	54%	48%	41%

Table 5: Accuracies of finding translation equivalents for English expressions

Coverage		20%	50%	80%
Baseline	CC limit 0.6	15%	10%	–
	CC limit 0.45	46%	20%	12%
	CC limit 0.3	65%	32%	16%
	CC limit 0.2	68%	33%	14%
Word2vec	CC limit 0.6	18%	13%	–
	CC limit 0.45	30%	33%	25%
	CC limit 0.3	44%	37%	36%
	CC limit 0.2	41%	31%	33%

A high comparability threshold for the paired documents (0.6) results in many of the texts not being paired to any document; therefore, the size of the comparable corpora decreases substantially and the coverage of the methods to find translations is very low. In this case, there are many expressions which do not occur in the smaller comparable corpora (CC with limit of 0.6). As a consequence, there is no coverage of 50% or higher for finding translation equivalents of Spanish expressions; equally, there is no coverage of higher than 50% in terms of finding translation equivalents for English expressions (see the missing values in the tables).

It can be seen that setting the comparability threshold at lower values, which in turn guarantees corpora of larger size, results in better performance for both the baseline approach and the word2vec approach. This trend applies to both English and Spanish expressions and these results indicate that the size of the corpora indeed matters. However, at least for English and for word2vec (and for the baseline at 80% coverage), the accuracy for CC limit of 0.2 is lower than for CC limit of 0.3. This indicates that quality of the corpora (in terms of the similarity score of comparable documents) also matters and that the size of the corpora appears to be decisive as long as the comparable corpora are of 'minimal quality'. A worthwhile future experiment would be to establish the optimal threshold for pairing documents which guarantees the best performance for translating MWEs.

Finally, an observation worth reporting is that at higher coverage (50% and 80%) the performance of the word2vec approach on smaller corpora (*CC limit of 0.45*) outperforms that of the baseline approach on bigger corpora (CC limit of 0.3). This finding can be interpreted to imply that the word2vec approach is more stable in terms of the size of corpora and is more stable with regard to different levels of coverage than the baseline approach.

7 Conclusion

The main conclusion from this proof-of-concept study is that it is feasible to extract MWEs and find their translations from comparable corpora without any translation resources such as dictionaries, parallel corpora, translation memories or bilingual grammars. While this study reports results for English and Spanish, the described methodology can be applied to any pair of languages. In addition, while the performance results are not very high, an important finding of this study is that as long as the comparable corpora used are of minimal similarity, then the size of the comparable corpora appears to be a more important factor for achieving higher accuracy than the similarity between them. An interesting future experiment would be to establish the optimal threshold for pairing documents which guarantees the best performance for translating MWEs.

Acknowledgments: The author would like to express his unreserved gratitude to Shiva Taslimipoor for implementing the above methodology and for assisting with the above experiments.

References

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*, Second Edition, 267–292. CRC Press.

Baldwin, Timothy & Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning* (CONLL'2002), 98–104.

Bouamor, Dhouha, Nasredine Semmar & Pierre Zweigenbaum. 2012. *Identifying bilingual multi-word expressions for statistical machine translation. In Proceedings of the eight international conference on language resources and evaluation* (LREC'2012).

Church, Kenneth & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Dice, Lee R. 1945. Measures of the amount of ecologic association between species. Ecology, 26, 297–302.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Evert, Stefan. 2005. *The statistics of word cooccurrences: word pairs and collocations* (Doctoral dissertation, Universität Stuttgart).

Fung, Pascale & Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. *Proceedings of the 5th Annual Workshop on Very Large Corpora*, 192–202.

Jackendoff, Ray. 2007. *Language, Consciousness, Culture: Essays on Mental Structure.*

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz & David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology*, 105–116.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*. Phuket, Thailand, 79–86.

Krenn, Brigitte & Stefan Evert. 2001. Can we do better than frequency? A case study on extracting pp-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, 39–46.

Levy, Omer. & Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, 302–308.

Manning, Christopher D, Prabhakar Raghavan & Hinrich Schutze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Mendoza Rivera, Oscar, Ruslan Mitkov & Gloria Corpas Pastor. 2013. A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In *Proceedings of the Workshop on Multiword Units in Machine Translation and Translation Technology* (MUMTTT).

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

Pecina, Pavel. 2005. *An extensive empirical study of collocation extraction methods*. In Proceedings of the ACL student research workshop. Stroudsburg, PA, USA, 13–18.

Pekar, Viktor, Ruslan Mitkov, Dimitar Blagoev & Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4), 247–266.

Pinnis, Marcus, Radu Ion, Dan Ştefănescu, Fangzhong Su, Inguna Skadiņa, Andrejs Vasiļjevs & Bogdan Babych. 2012. "ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. *Proceedings of the ACL 2012 System Demonstrations*, 91–96.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* 519–526.

Rapp, Reinhard & Serge Sharoff. 2014. Extracting multiword translations from aligned comparable documents. In *Proceedings of the EACL 3rd Workshop on hybrid approaches to translation*. Gothenburg, Sweden, 83–91.

Schone, Patrick & Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP'2001), 100–108.

Skadiņa, Inguna, Ahmet Aker, Nikos Mastropavlos, Fangzhong Su, Dan Tufiș, Mateja Verlic, Andrejs Vasiļjevs, Bogdan Babych, Paul Clough, Robert Gaizauskas, Nikos Glaros, Monica Lestari Paramita & Marcis Pinnis. 2012. Collecting and Using Comparable Corpora for Statistical Machine Translation. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'12), 438–445.

Su, Fangzhong & Bogdan Babych. 2012a. Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents. *Proceedings of the EACL'12 Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation* (ESIRMT) *and Hybrid Approaches to Machine Translation* (HyTra), 10–19.

Su, Fangzhong & Bogdan Babych. 2012b. Development and Application of a Cross-language Document Comparability Metric. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), 3956–3962.