A Large Automatically-Acquired All-Words List of Multiword Expressions Scored for Compositionality

Will Roberts, Markus Egg
Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany
{will.roberts, markus.egg}@anglistik.hu-berlin.de

Abstract

We present and make available a large automatically-acquired all-words list of English multiword expressions scored for compositionality. Intrinsic evaluation against manually-produced gold standards demonstrates that our compositionality estimates are sound, and extrinsic evaluation via incorporation of our list into a machine translation system to better handle idiomatic expressions results in a statistically significant improvement to the system's BLEU scores. As the method used to produce the list is language-independent, we also make available lists in seven other European languages.

Keywords: multiiword expressions, compositionality, machine translation

1. Introduction

Multiword expressions (MWEs) are phraseological units, which consist of more than one lexeme and exhibit some kind of idiosyncrasy (Sag et al., 2002); such idiosyncrasy may be lexical (*ad hoc*), syntactic (*by and large*), semantic (*middle of the road*), pragmatic (*all aboard*), or statistical (*black and white* but not *white and black*; these are commonly known as *collocations*) (Baldwin and Kim, 2010).

In this paper, we present a new linguistic resource, in the form of a large automatically-acquired all-words list of MWEs, which aims to support future research into semantically idiosyncratic MWEs. Semantically idiosyncratic MWEs, or *idiomatic expressions*, are *non-compositional* in that their meanings cannot be predicted from their parts; these expressions are used frequently to make language more fluent (Jackendoff, 1997), and often contain word senses not found in other contexts. Thus, identifying non-compositional MWEs presents a clear challenge for fields such as automatic machine translation (MT), information retrieval, natural language understanding, natural language generation, question answering, text summarisation, and word sense disambiguation (McCarthy et al., 2007). In recent years, there has been considerable interest in the MWE community in automatically estimating compositionality (Biemann and Giesbrecht, 2011; Reddy et al., 2011; Schulte im Walde et al., 2013; Salehi et al., 2015); however, to the best of our knowledge, this work has hardly been applied to real-world NLP tasks. We set out to distribute a convenient resource representing the best practices gleaned from this work, by automatically scoring the expressions on our list for compositionality.

This paper is structured in the following way: Section 2. lists previous work in this area, while section 3. details our acquisition method. Our resource is then evaluated intrinsically against manually-produced gold standards in section 4., and extrinsically, inside a MT system in section 5..

2. Related work

While the resource introduced in this paper is an all-words list acquired automatically, most existing MWE resources are produced manually and focus on a single part of speech (e.g., noun-noun compounds, verb-noun constructions, verb particle constructions, adjective-noun constructions);[1] some examples of these are used in Section 4..

Other more general resources include machine-readable dictionaries that happen to list MWEs; examples include the TED-MWE bilingual dictionary (Monti et al., 2015), with 2,484 automatically-extracted aligned EN-IT MWEs, and BabelNet (Navigli and Ponzetto, 2012), some of whose 8.5 M entries in 271 languages are MWEs.

MWE research dealing with compositionality tends to focus on methodologies rather than producing resources. There are also MWE compositionality resources that are not targeted towards natural language processing, such as Martinez and Schmitt (2012), who produce a list of 505 non-compositional English phrases for teaching English as a second language.

In contrast to the monolingual method we make use of here, some methods to estimate compositionality do so by measuring the relative difficulty of translating an expression into another language; an example is Villada Moirón and Tiedemann (2006), who leveraged parallel corpora to extract Dutch MWEs. However, for languages such as Basque this approach is not feasible, because parallel corpora are very limited in size and number and restricted to few languages (Leturia et al., 2009; Leturia, 2012).

3. Acquisition of non-compositional MWEs

We collect lexical co-occurrence statistics on all words in the English Wikipedia, using the WikiExtractor tool[2] to retrieve plain text from the April 2015 dump (ca. 2.8B words), and using simple regular expressions to segment sentences and words, and remove URLs and punctuation. We perform no POS tagging, lemmatisation, case normalisation, or removal of numbers or symbols; MWE acquisition using unlemmatised text in this way may be useful for capturing the morphological or syntactic fixedness of some idiomatic

[1] Losnegaard et al. (2016) offers a recent survey and `http://multiword.sourceforge.net/`, a list of MWE resources.

[2] `https://github.com/bwbaugh/wikipedia-extractor`

MWEs (e.g., identifying *spill the beans* but not *spill the bean*[3]). We collect word frequency information with the SRILM language modelling toolkit (Stolcke, 2002)[4], counting n-grams ($n \leq 3$), treating MWEs as contiguous[5] bigrams and trigrams), and identify MWE candidates by computing the Poisson collocation measure (Quasthoff and Wolff, 2002)[6] for all bigrams and trigrams (ca. 23M n-grams). This method should be readily extensible to include longer n-grams.

The Poisson measure we use is chosen after an empirical evaluation of several commonly used association measures[7]:

chi χ^2: $\sum_{i,j} \frac{[f_{ij} - f'_{ij}]^2}{f'_{ij}}$

conf confidence (Omiecinski, 2003) $\max[\frac{P(AB)}{P(A)}, \frac{P(AB)}{P(B)}]$

mi mutual information: $\sum_{i,j} P_{ij} \log(\frac{P_{ij}}{P'_{ij}})$

pe permutation entropy (Zhang et al., 2006)

poisson Poisson collocation measure (Quasthoff and Wolff, 2002) $\frac{f'(xy) - f(xy) \log f'(xy) + \log[f(xy)!]}{\log N}$. This is identical up to a constant factor with the "log likelihood measure" introduced by Dunning (1993).

poissonT Poisson balanced for trigrams

ps Piatetsky-Shapiro (Piatetsky-Shapiro, 1991) $P(AB) - P(A)P(B)$

psT Piatetsky-Shapiro balanced for trigrams $P(ABC) - P(A)P(B)P(C)$

ttest t-test: $\frac{f(AB) - f'(AB)}{\sqrt{f(AB)[1 - f(AB)/N]}}$

ttestT t-test balanced for trigrams

We estimate the quality of these rankings by searching for known collocations and multiword expressions, and finding the ranks of these known expressions in the lists. We define a good association measure as one which tends to rank these known expressions highly (as operationalised by the Mean Reciprocal Rank). For this comparison, we use manually-constructed lists of multiwords intended as gold standards in MWE acquisition work:

English noun compound (NC) (Nakov, 2008)

English verb particle constructions (VPC) (Baldwin, 2005)

	NC	VPC
χ^2	7.7	2.8
conf	1.2	1.1
mi	55.0	64.5
pe	2.2	2.6
poisson	66.3	74.6
poissonT	70.0	78.5
ps	34.0	77.6
psT	32.3	74.7
ttest	38.7	79.0
ttestT	35.7	73.8

Table 1: Mean Reciprocal Rank ($\times 10^{-7}$) by association measure for two test corpora. Higher values are better.

Score	MWE	Cosine similarities		
0.005	*a front for*	—	0.005	—
0.012	*red tape*	−0.056	0.081	
0.191	*stops short of*	0.285	0.097	—

Table 2: Some compositionality-scored MWE candidates.

Table 1 lists the results for the association measures on these two corpora, demonstrating that the Poisson measure works best for our task.

We then automatically score the million most strongly associated n-grams (i.e., roughly the top 5% of the Poisson-ranked list) for compositionality. Compositionality scores are assigned using a method based on the work of Salehi et al. (2015), which represents the current state of the art. Using word2vec (Mikolov et al., 2013)[8] with the parameters[9] found to be most effective by Baroni et al. (2014), we build a word embedding vector for every simplex word in the vocabulary (ca. 1M types), as well as for each MWE candidate. We then compute the cosine similarity of the vector representation for a MWE candidate with the vectors of its constituent words, and take the arithmetic mean. In scoring the compositionality of a candidate, we do not measure the cosine similarity of the MWE with any stop words it may contain, as stop words may be assumed to be semantically uninformative[10]. Table 2 presents several extracted MWE candidates with their computed compositionality scores and shows that cosine similarity scores with determiners (*a*) and prepositions (*for* and *of*) are ignored.

The embedding vectors are trained on the extracted Wikipedia text, where each occurrence of a MWE candidate is greedily replaced with a single token representing the MWE as a word-with-spaces. The string rewriting is performed efficiently using the Aho-Corasick algorithm (Aho and Corasick, 1975). This greedy rewriting procedure cannot deterministically handle n-grams which overlap with

[3] For example, Villada Moirón and Tiedemann (2006) found lemmatisation to be unhelpful for identifying non-compositional MWEs, because of the tendency of idiomatic MWEs to display more morphosyntactic fixedness than literal text.

[4] http://www.speech.sri.com/projects/srilm/

[5] Note that, while many MWEs are contiguous (e.g., *in a nutshell*), some may be non-contiguous (e.g., *take a (long) bath*).

[6] This measure is almost identical to the log-likelihood ratio introduced by Dunning (1993).

[7] For a more complete list of association measures commonly used in the MWE acquisition literature, the reader is referred to (Pecina, 2008).

[8] https://code.google.com/p/word2vec/

[9] Continuous bag of words model with 400-dimensional vectors, window size 5, subsampling with $t = 10^{-5}$, negative sampling with 10 samples. We build vectors only for tokens observed 20 times or more in the corpus.

[10] Stop words are taken here to be the 50 most frequent words in the vocabulary.

	Found	Total	Spearman ρ	Pearson's r
F_ENC	631	1042	0.458	0.473
R_ENC	61	90	0.615	0.603
MC_VPC	48	117	0.432	0.379
D_ADJN	64	68	0.525	0.581
MC_VN	132	638	0.392	0.395

Table 3: Correlation of our compositionality-ranked list against manually-constructed gold standards.

other n-grams, so we sort the MWE candidates into 10 disjoint batches such that, for any two candidates e_1, e_2 in the same batch, e_1 is neither a substring, nor a superstring of e_2, and there is no prefix of e_1 which is a suffix of e_2. This sorting is performed greedily, by processing candidates in order of decreasing Poisson score, and assigning each candidate to the first batch for which this property obtains; candidates which cannot be assigned to a batch (ca. 6.8%) are discarded. Each batch thus results in a word embedding model for all single words in the vocabulary, and some subset of the MWE candidates; after computing the compositionality scores, we recombine the candidates from all batches to produce a single list that is sorted in order of increasing compositionality, containing 917,647 expressions.

4. Intrinsic Evaluation

We conducted an in-vitro evaluation of the compositionality scores by measuring correlations against several gold standard datasets from the MWE compositionality literature, which contain human judgements of how predictable the meaning of a MWE is from its constituent words. The datasets are:

F_ENC (Farahmand et al., 2015) 1,042 noun compounds (e.g., "cat fight", "chicken breast", "crash course", etc.) annotated by five judges, with some filtering, resulting in a 5-point Likert scale. Inter-annotator agreement by Fleiss' κ was 0.62. Yazdani et al. (2015) report $\rho = 0.410$ on this dataset.

R_ENC (Reddy et al., 2011) 90 noun compounds (e.g., "snail mail", "guilt trip", etc.) annotated over Amazon Mechanical Turk using a 6-point Likert scale. Inter-annotator agreement by averaged Spearman correlation between rankings was $\rho = 0.686$. Salehi et al. (2015) reported achieving $r = 0.796$.

MC_VPC (McCarthy et al., 2003) 117 verb-particle pairs (e.g., "rule out", "clamp down", etc.) annotated by 3 judges, with averaged scores on a 11-point Likert scale. Inter-annotator agreement with Kendall's Coefficient of Concordance is reported to be $W = 0.594$. The original paper reports $\rho = 0.49$ using a method based on measuring the size of overlap in synonyms of the phrasal verb and in those of the bare ("simplex") verb, using an automatically constructed thesaurus.

D_ADJN (Biemann and Giesbrecht, 2011) $58 + 10 = 68$ compounds (Adj-NN compounds only) from the training and validation sets of the Disco 2011 Shared Task (e.g., "mental health", "soft drink", "small group", etc.). Annotated over Amazon Mechanical Turk using a 11-point Likert scale, with scores averaged over judges. No inter-annotator agreement figures are available. Krčmář et al. (2013) achieved $\rho = 0.54$ using a LSA-based model.

MC_VN (McCarthy et al., 2007) This subset of the resource constructed by Venkatapathy and Joshi (2005) contains 638 verb-object pairs (e.g., "lend money", "turn back", "watch television", etc.) annotated by two judges using a 6-point Likert scale. This list also contains some non-contiguous items (e.g., "lose temper", "beg question", etc.) not found in our list. Inter-annotator agreement by Kendall's $\tau = 0.61$; Spearman rank correlation between annotators: $\rho = 0.71$. Kiela and Clark (2013) reported $\rho = 0.461$.

Table 3 shows the correlation of our compositionality scores against these gold standards. The table lists the size of each gold standard dataset, and its overlap with our resource.

The compositionality ranking accords well with human judgements, with correlation scores not far from the state of the art, and 10–30 percentage points below the human inter-annotator agreement.. In the case of the largest resource, F_ENC, we are not aware of a better correlation than the one we report here. The list is positively correlated with all gold standard judgements, representing a variety of parts of speech, and all correlations are statistically significant. This demonstrates the validity of our compositionality scoring.

The n-gram statistics we collect contain 1,562 of a total 1,931 items from the gold standards; we can take this number to be the count of these compounds which are attested in the English Wikipedia. Our compositionality-ranked list contains only 912 items from the gold standards. Part of this decrease represents the MWE candidates which are discarded due to low association measure scores, and part likely results from MWE candidates lost because they could not be assigned to a batch for compositionality computation. Note that the largest number of missing compounds come from MC_VN, which, as noted, contains many discontinuous (non-n-gram) compounds.

5. Extrinsic Evaluation: MT

To evaluate the utility of our resource for NLP applications, we conduct an extrinsic evaluation by incorporating MWE knowledge into an automatic English-Spanish translation system.

TectoMT (Žabokrtský et al., 2008) is a linguistically sophisticated hybrid MT system which uses a combination of statistical and rule-based components in a modular pipeline model to analyse source language up to a highly abstract (*tectogrammatical*) level of representation; this so-called t-tree is a dependency tree structure containing only nodes for autosemantic words. The morphosyntactic properties of the nodes (t-nodes) in this t-tree are represented by *formemes*, which encode grammatical roles and complements (e.g., *n:subj* for a noun in subject position, or *n:for+X* for a noun preceded by the preposition *for*).

In the *transfer* stage, translation is performed by first **copying** the source language t-tree structure into the target lan-

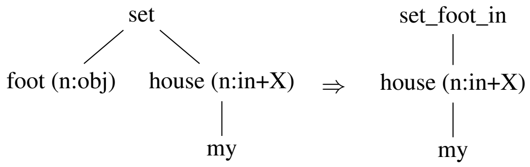

Figure 1: Tectogrammatical reduction of multiple t-nodes (representing the non-compositional MWE *set foot in*) into a single composite t-node.

guage; the formemes and lemmas on each tree node are then translated into the target language using a maximum entropy translation model. The copying done in the first step means that the translation produced by the system is (on the tectogrammatical level) structurally identical to the input; thus, the system operates using the strong assumption that translation can be performed *isomorphically*. The translation model is learnt by analysing parallel corpora using the TectoMT pipeline, and then inducing maxent models for lemma and formeme translation from Giza++ alignments; in the experiments we report here, we train our models on Europarl (Koehn, 2005). Following transfer, further pipeline components generate successively concrete representations in the target language, until the system can produce a linearised string of words as its output translation.

The isomorphicity assumption built into TectoMT generally works well, but is problematic when the source language contains non-compositional MWEs. Therefore, we integrate lexical information about MWEs in the form of our compositionality-ranked list into the TectoMT pipeline, by collapsing multiple t-nodes in a t-tree which represent a single MWE into a single composite t-node. Note that this paradigm will only work for MWEs which can be translated into single lexical nodes in the target language; MWEs which are translated by other multiwords will result in translation failures (i.e., insertion or deletion errors). However, we expect that such failures will happen relatively infrequently[11].

For this work, we perform some semi-automatic filtering of our MWE list, removing several of the more common errors that we observed, using a simple pattern-based filter (e.g., discarding those candidates which begin or end with a conjunction or some form of the copula). We also discard some MWE candidates which are superstrings or substrings of another MWE candidate with a lower compositionality score, when the two candidates have very similar word embedding vectors. This results in the removal of around 11% of the candidates from the list, leaving us with 817,592 MWE candidates.

Immediately prior to the transfer stage, we identify MWEs in the source language greedily by searching on word forms in the input and finding their corresponding t-nodes in the t-tree; we match only sets of nodes in the t-tree that are fully connected to each other by dependency relations (i.e., which are *treelets*). In this search, MWEs with lower compositionality scores are preferred; ties are broken arbitrarily by taking the leftmost match. Successfully matched MWE instances have their lemma altered to a word-with-spaces representation, and are collapsed by deleting dependent MWE nodes and rearranging arguments so that these depend on the new composite node. Figure 1 shows the reduction performed in the analysis of a successfully matched MWE instance[12].

Threshold	Types	Tokens
$\theta \leq 0.1$	1,093	32,956
$\theta \leq 0.2$	5,020	174,015
$\theta \leq 0.5$	90,133	2,808,015

Table 4: Counts of MWEs observed during TectoMT training on Europarl with varying compositionality thresholds.

Performing this analysis during training of the TectoMT system allows the translation model to learn how to translate English MWEs observed in the training corpus into Spanish. We record all MWEs seen during training, and use only this list for analysis during testing, to ensure that no MWEs in the test corpus are reduced for which the trained translation model has not learnt any translations (which would create new out-of-vocabulary items). This has the effect of filtering our MWE candidate list, so that, at test time, only those expressions found in the translation training corpus are used to analyse the test data. We manipulate the compositionality value θ as an independent variable, using a threshold to control the number and compositionality of MWEs that are analysed in the source text. For example, with $\theta \leq 0.1$ we restrict the MWE candidate list to contain only those items whose compositionality score is less than or equal to 0.1. Table 4 shows the number of MWEs found in the English section of Europarl for different values of the threshold.

We train four English-Spanish models on Europarl: a baseline model, which does not analyse MWEs, and three MWE-enabled models, using threshold values of $\theta \leq 0.1$, $\theta \leq 0.2$, and $\theta \leq 0.5$. We test these models on the ACL 2008 shared translation task (Callison-Burch et al., 2008), containing 2,000 sentences (ca. 55 K words) from Europarl. We also build a MWE-rich test corpus by filtering the test split of Europarl (Oct.–Dec. 2000), retaining only sentences that contain one or more highly non-compositional ($\theta \leq 0.1$) MWEs from our list. This produces a small English-Spanish test corpus of 518 sentences (ca. 18K words).

Case-insensitive BLEU scores (Papineni et al., 2002) summarising our results are presented in Table 5, which also shows the counts of MWEs observed during testing. On both test sets, we observe a similar pattern: Analysing MWEs improves translation over the baseline model, but only when using low values of the compositionality threshold; performance falls below the baseline as this threshold is increased. This effect is expected, because it is likely that composite t-nodes representing compositional English MWEs cannot be adequately translated by single lexemes in Spanish.

On the ACL 2008 test set, we observe an absolute improvement over the baseline of $+0.18$ BLEU points (1% relative)

[11] Cf. Uresova et al. (2013), who found in the Parallel Czech-English Dependency Treebank that most verbal MWEs are not translated by other MWEs.

[12] In this example, the preposition *in* has been encoded in the formeme of the t-node under it (*house*) by the TectoMT system, but our analysis will still find this treelet because it can find *set* and *foot*.

Experiment	MWE Counts Types	Tokens	BLEU
ACL 2008 shared task			
Baseline			12.55
$\theta \leq 0.1$	7	17	**12.73** $*$
$\theta \leq 0.2$	39	74	12.66
$\theta \leq 0.5$	715	1,175	11.99
MWE-rich test set			
Baseline			11.59
$\theta \leq 0.1$	20	71	11.39
$\theta \leq 0.2$	37	99	**11.83**
$\theta \leq 0.5$	299	449	11.28

Significance relative to the baseline: $*$: $p < 0.01$

Table 5: TectoMT experimental results: BLEU scores of different MWE-enabled models on two test corpora.

when using the lowest value of the compositionality threshold; this effect is statistically significant at the $p < 0.01$ level[13]. Increasing the threshold to $\theta \leq 0.2$, the improvement is smaller but still positive; the effect is not significant. On the MWE-rich test set, the $\theta \leq 0.2$ model obtains an absolute improvement over the baseline of +0.24 BLEU (2% relative); due to the small test corpus size, this effect is not significant ($p = 0.066$). The $\theta \leq 0.1$ model, by contrast, performs more poorly than the baseline. Error analysis does not conclusively explain this, but we have observed the model making mistakes due to instances of non-compositional MWEs, such as *came into force*, which happen to have literal translations in Spanish (*entró en vigor*). The $\theta \leq 0.2$ model appears to contain helpful MWEs, such as *(on) the one hand*, which help to offset these errors.

It is interesting to note that the improvement to BLEU scores is out of proportion to the number of MWEs analysed at test time; for instance, the best improvement seen on the ACL 2008 test set occurs when TectoMT finds only 17 instances of MWEs in the test corpus. We have observed this phenomenon while training models on other parallel corpora, and while using other test sets—sometimes this results in better-than-baseline performance on test sets containing no MWEs at all. We surmise that treating non-compositional MWEs while training TectoMT allows the translation model to learn to ignore spurious translations of polysemous verbs (e.g., *come, enter, set*) and nouns (e.g., *point, term*) which enter into idiomatic expressions; that is, when learning to translate a particular lexeme, the model is not distracted by the translations of MWEs which include that lexeme. E.g., suppose that the analysis of the parallel corpora couples *come to terms* with its Spanish translation *llegar a un acuerdo*. If we identify the English expression as a MWE, we make sure that there is no spurious analysis of *terms* as the English equivalent of *acuerdo* 'agreement' regardless of whether or not *come to terms* shows up in the material to be translated automatically.

6. Conclusion

We have introduced a new automatically-acquired all-words list of MWEs, automatically ranked for compositionality. Evaluation against manually-created gold standards validates our compositionality scores, and incorporating our list into a MT system to detect idiomatic language gave a statistically significant improvement to the system's BLEU scores.

We used the same language-independent method to build compositionality-ranked lists for other languages (Bulgarian, Czech, German, Spanish, Basque, Dutch, and Portuguese); we make these lists available here without evaluation.

7. Acknowledgements

This work was supported by the EC's FP7 (FP7/2007- 2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches".

8. Bibliographical References

Aho, A. V. and Corasick, M. J. (1975). Efficient string matching: An aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In Nitin Indurkhya et al., editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Boca Raton, USA, second edition.

Baldwin, T. (2005). Deep lexical acquisition of verb–particle constructions. *Computer Speech & Language*, 19(4):398–414.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore. Association for Computational Linguistics.

Biemann, C. and Giesbrecht, E. (2011). Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 21–28. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106. Association for Computational Linguistics.

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

Farahmand, M., Smith, A., and Nivre, J. (2015). A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, Denver, CO. Association for Computational Linguistics.

Jackendoff, R. (1997). *The architecture of the language faculty*. MIT Press, Cambridge, MA.

Kiela, D. and Clark, S. (2013). Detecting compositionality of multi-word expressions using nearest neighbours

[13] In this paper, significance tests use bootstrap resampling, and one-tailed p values are reported (Koehn, 2004). We use the MT-ComparEval software (Klejch et al., 2015), `https://github.com/choko/MT-ComparEval`.

in vector space models. In *Proceedings of the Short Papers of the Conference on Empirical Methods in Natural Language Processing (EMNLP-13)*, pages 1427–1432.

Klejch, O., Avramidis, E., Burchardt, A., and Popel, M. (2015). MT-ComparEval: Graphical evaluation interface for machine translation development. *Prague Bulletin of Mathematical Linguistics*, 104(1):63–74.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Dekang Lin et al., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*, pages 79–86.

Krčmář, L., Ježek, K., and Pecina, P. (2013). Determining compositionality of word expressions using word space models. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 42–50, Atlanta, GA. Association for Computational Linguistics.

Leturia, I., San Vicente, I., and Saralegi, X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *Proceedings of the 5th International Web as Corpus Workshop (WAC5)*, pages 53–61.

Leturia, I. (2012). Evaluating different methods for automatically collecting large general corpora for Basque from the web. In *COLING 2012*, pages 1553–1570.

Losnegaard, G. S. G., Sangati, F., Escartín, C. P., Savary, A., Bargmann, S., and Monti, J. (2016). PARSEME survey on MWE resources. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Martinez, R. and Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, 33(3):299–320.

McCarthy, D., Keller, B., and Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80. Association for Computational Linguistics.

McCarthy, D., Venkatapathy, S., and Joshi, A. K. (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. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Monti, J., Sangati, F., and Arcan, M. (2015). TED-MWE: A bilingual parallel corpus with MWE annotation. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, pages 193–197, Trento, Italy.

Nakov, P. (2008). Paraphrasing verbs for noun compound interpretation. In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 46–49.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Omiecinski, E. R. (2003). Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In Gregory Piatetsky-Shapiro et al., editors, *Knowledge Discovery in Databases*, chapter 13, pages 229–238. MIT Press, Cambridge, MA.

Quasthoff, U. and Wolff, C. (2002). The Poisson collocation measure and its applications. In *Proceedings of the 2nd International Workshop on Computational Approaches to Collocations*, Vienna, Austria.

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai. Asian Federation of Natural Language Processing.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico.

Salehi, B., Cook, P., and Baldwin, T. (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 977–983. Association for Computational Linguistics.

Schulte im Walde, S., Müller, S., and Roller, S. (2013). Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 255–265.

Stolcke, A. (2002). SRILM: An extensible language modeling toolkit. In John H. L. Hansen et al., editors, *Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002*, Denver, CO. ISCA.

Uresova, Z., Sindlerova, J., Fucikova, E., and Hajic, J. (2013). An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 58–63, Atlanta. Association for Computational Linguistics.

Venkatapathy, S. and Joshi, A. K. (2005). Relative compositionality of multi-word expressions: A study of verb-noun (VN) collocations. In *Natural Language Processing–IJCNLP 2005*, pages 553–564. Springer, Berlin, Heidelberg.

Villada Moirón, B. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006): Workshop on multiword expressions in a multilingual context*, pages 33–40, Trento.

Yazdani, M., Farahmand, M., and Henderson, J. (2015). Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon. Association for Computational Linguistics.

Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170. Association for Computational Linguistics.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties (MWE 06)*, pages 36–44, Sydney. Association for Computational Linguistics.