Discovering Bilingual Collocations in Parallel Corpora: A First Attempt at Using Distributional Semantics

Marcos Garcia, Marcos García-Salido and Margarita Alonso-Ramos

NOTICE: this is the final peer-reviewed manuscript that was accepted for publication in *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*. Changes resulting from the publishing process, such as editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version will be published in *Parallel Corpora for Contrastive and Translation Studies: New resources and applications* (John Benjamins Publishing), edited by Irene Doval and M. Teresa Sánchez Nieto: https://doi.org/10.1075/scl.90.16gon

Introduction

When learning a new language, non-native speakers usually face difficulties regarding the use of conventionalized lexical combinations in each linguistic variety (Altenberg and Granger, 2001), it being common for language learners to produce false combinations, usually influenced by their mother tongue (Nesselhauf, 2004; Alonso-Ramos *et al.*, 2010).

One such type of expressions is *collocations*, which can be defined as unpredictable lexically restricted combinations of two lexical units (Mel'čuk, 1998). As an example, a native speaker of Portuguese might use **give a walk* instead of *take a walk* in English, guided by the Portuguese collocation *dar um passeio,* since the primary meaning of the verb *dar* corresponds to *to give*.

The manual creation of large structured lists of bilingual collocation equivalents involves a great deal of effort from expert lexicographers. Nevertheless, these resources could be very useful for a variety of activities such as second language learning or tasks such as machine translation (Orliac and Dillinger, 2003).

In this respect, parallel corpora are interesting resources for the extraction of interlinguistic equivalents and have been widely used for identifying both monolexical equivalents and plurilexical entries (named entities, multiword expressions, etc.) (Fung, 1998; Smadja, 1993, Şulea *et al.*, 2016).

In this paper we explore the use of parallel corpora not only to discover bilingual collocation equivalents, but also to create a bilingual distributional semantics model. This model is used to extract, with a high degree of precision, bilingual collocation equivalents from corpora.

In the proposed approach, we follow the phraseological criterion that defines a collocation as a restricted pair of lexical units (LUs), where one of them (the *base)* is freely selected by the speaker (e.g., *walk* in *take a walk*), while the other one (the *collocate*; e.g., *take* in *take a walk*) is restricted by the base (Mel'čuk, 1998). Thus, our strategy for discovering bilingual collocation synonyms consists of searching for semantically similar equivalents of both the base (in a first step) and the collocate of each collocation. As the model only makes use of distributional information, it is worth noting that, ideally, it could be applied to other non-parallel corpora for extracting additional examples of collocation equivalents not present in the original bilingual resource.

In order to test the proposed strategy, we carried out an experiment to automatically discover bilingual *verb-object* collocations in a Spanish-Portuguese corpus. The results of this test show that this method enables high-quality bilingual collocation equivalents to be extracted with no manual effort whatsoever.

Previous Research on Bilingual Collocation Extraction

Since the 1990s, a number of works have exploited parallel corpora to extract bilingual collocation equivalents. Smadja (1992) and Smadja *et al.* (1996) use a parallel English-French corpus first to identify monolingual collocations in a source language (English), and then to discover the French equivalents of the source collocations by applying different similarity measures such as *Mutual Information* or *Dice*.

Kupiec (1993) also uses the same parallel corpus to find English-French noun phrase equivalents. This method applies an expectation maximization algorithm on monolingual collocations that had previously been extracted.

Wu and Chang (2003) also take advantage of parallel corpora to extract Chinese and English *n-grams* from aligned sentences, by computing their *log-likelihood* ratio. Then, they apply a competitive linking algorithm to decide if the bilingual pairs are real Chinese-English translations.

Using syntactic analysis, Seretan and Wehrli (2007) extract bilingual collocations from parallel corpora. First, the authors obtain monolingual collocations using *log-likelihood*, and then they search for equivalents of each base using bilingual dictionaries. Although this is not always the case,

this method assumes that bilingual collocation equivalents occur in the same syntactic relation in both the source and the target languages.

More recently, Rivera *et al.* (2013) present different methods for bilingual collocation extraction which can be applied in both parallel and comparable corpora. This work relies on *n-grams* to extract monolingual collocations, and on bilingual dictionaries (or WordNet) to identify the bilingual equivalents.

A different approach was adopted by Lü and Zhou (2004), who use non-related monolingual corpora for finding bilingual collocations. The work they present first applies dependency parsing and the *log-likelihood* ratio for discovering English and Chinese monolingual collocations. They subsequently search for bilingual collocations using translation equivalents – with the expectation maximization algorithm and bilingual dictionaries– of words that occur in the same dependency relation in both languages.

Most of these papers use contextual information (e.g., the position of the collocations in the corpus) to find the bilingual equivalents, together with bilingual dictionaries. In the present paper, we introduce a different approach which eliminates the need for bilingual dictionaries and does not make use of explicit contextual information, making it easy to extend to other languages and applicable to different corpora. In Garcia *et al.* (2017) we presented the results of a similar research project using both parallel and comparable corpora in Portuguese, Spanish, and English.

Proposed Strategy

This section presents a method for discovering bilingual collocations within corpora. The strategy consists of the following phases: First, we extract monolingual collocation candidates from corpora, using dependency parsing to identify syntactically related words together with the application of statistical measures for ranking the extracted candidates. Then, we train a bilingual distributional semantics model using a word level automatic alignment of the parallel corpora. Finally, this model is applied on the monolingual collocations to identify bilingual equivalents automatically. Both the monolingual collocations and the distributional model are trained using lemmas instead of tokens, so that the system groups together different occurrences of the same collocation into its canonical form. For example, expressions such as *took several pictures*, *take one more picture*, *pictures were taken*, etc. are clustered in the collocation "$picture_{Base}$, $take_{Collocate}$".

Extracting Monolingual Collocation Candidates

As mentioned, we use dependency parsing to extract monolingual candidates of collocations, enabling us to avoid the incorrect identification of non-related words that frequently co-occur in corpora (Evert, 2008). Moreover, syntactic analysis also makes it possible to discover word combinations that appear in a relatively large span of text, which are not usually obtained using extraction methods based on *n-grams* (Smadja, 1993).

Before extracting the collocation candidates, we enrich the corpus with linguistic information by means of automatic tokenization, lemmatization and PoS-tagging. Then, we apply a dependency parser to syntactically

annotate the text using *Universal Dependencies* (Nivre *et al.*, 2016). Universal Dependencies is a recent initiative aimed at providing treebanks in many languages labeled with a uniform annotation. The use of the same criteria for labeling the syntactic information facilitates a multilingual work such as the one presented in this paper. Table 1 shows a representation of the sentence "She hates black money". In this representation, the first column represents the position of each token in the sentence, followed by its form, lemma and PoS-tags (columns 2 to 4). The last two columns indicate, respectively, the syntactic head of each token (where a 0 means that this is the root of the sentence) and the dependency relation both tokens bear (the dependent and the head) in this context.

id	*token*	*lemma*	*PoS-tag*	*head*	*dep*
1	She	she	PRON	2	nsubj
2	hates	hate	VERB	0	root
3	black	black	ADJ	4	amod
4	money	money	NOUN	2	dobj

Table 1: Example of a labeled sentence.

After labeling the corpus, we extract candidates of different types of collocations: *verb-object* ("$statement_B$, $make_C$"), *adjective-noun* ("$money_B$, $black_C$"), and *noun-(preposition)-noun* ("$cigarette_B$, $packet_C$").[1] These candidates are then ranked using standard association measures such as *Mutual Information* (MI, which promotes low-frequency candidates and works well in large corpora (Pecina, 2010)), *t-score* (which assigns high scores to frequent word combinations (Krenn and Evert, 2001)), or others

1 Note that each collocation type has both syntactic and morphosyntactic restrictions: *verb-object*, for instance, would require a noun (as the base) and a verb (as the collocate) occurring in a *dobj* dependency relation.

such as *z-score* or *log-likelihood*. Table 2 shows some examples of candidate collocations in Portuguese ranked by their mutual information values.

base	*collocate*	*MI*
pastilha	mascar	11.6
formulário	preencher	11.3
homenagem	render	10.9
susto	pregar	10.6
isco	morder	10.1

Table 2: Example of Portuguese *verb-object* candidates ranked by MI.

Thus, we obtain several lists (per language, per collocation type, and per association measure), which are then merged into a unique file per language and collocation type from a given threshold.

Bilingual Distributional Semantics Model

A (monolingual) model of distributional semantics maps each word of a vocabulary into a vector of real numbers which represents the distributional properties of the word. Recently, distributional semantics has become very popular in the natural language processing (NLP) community due to the publication of *word2vec* (Mikolov *et al.*, 2013), which uses neural network approaches to reduce the dimensionality of the vectors (*word-embeddings*). In a monolingual scenario, we can compute the cosine distance between the vectors of two words, obtaining a similarity measure between them.

In our case, the hypothesis behind the use of a bilingual model of distributional semantics is that it can effectively learn words with a similar distribution (in a target language) semantically related to those in the source one. Thus, as an independent task to the extraction of monolingual

candidates, we train a bilingual model of distributional semantics using a naïve approach that takes advantage of an automatic alignment, at word level, of parallel corpora.

In order to rapidly build a bilingual model, we use the following strategy: As a prior step to training the bilingual model, we replace each token of the corpus by its lemma with a view to reducing data sparseness and making the corpus compatible with the monolingual collocations (which had been extracted using the lemmas). We also incorporate a language suffix, in order to better identify the language of each lemma (Figure 1).

tirou várias fotografias	sacó varias fotografías	took several pictures
tirar_{PT} vário_{PT} fotografia_{PT}	sacar_{ES} vario_{ES} fotografía_{ES}	take_{EN} several_{EN} picture_{EN}

Figure 1: Example of original (top) and lemmatized and suffixed sentences (bottom) in Portuguese (left), Spanish (center), and English (right).

Then, we apply an automatic aligner to identify the correspondence of each word of the source language in the target one (Dyer *et al.*, 2013). We use the information provided by the word alignment to create a *mixed bilingual* corpus by concatenating each equivalent lemma in the source and target languages:

"tirar_{PT} sacar_{ES} vário_{PT} vario_{ES} fotografia_{PT} fotografía_{ES}"

Thus, we obtain a mixed corpus with bilingual word equivalents occurring together (or close to each other), in order to train a naïve bilingual distributional model. To create the model, we apply *word2vec* in this final corpus using the *skip-gram* architecture, which assigns a higher weight to words nearby in the context than to more distant ones. We defined the vector dimension as 300 and used a context window of 20 tokens (≈10 in each

language). Furthermore, we created distributional vectors only for those words with at least 5 occurrences in the corpus.

The resulting model contains, for each word in the corpus (in Portuguese and Spanish), its distributional vector which encodes information about the linguistic contexts in which the word occurs. Thus, given two words, the distance between their vectors (computed by their cosine distance) will ideally reflect their semantic similarity.

Using this bilingual model, we can search for the most similar words (in terms of their distribution) in a target language given a certain input from a source language. The plot in Figure 2 shows a simple 2D visualization example of a bilingual Spanish-Portuguese model.

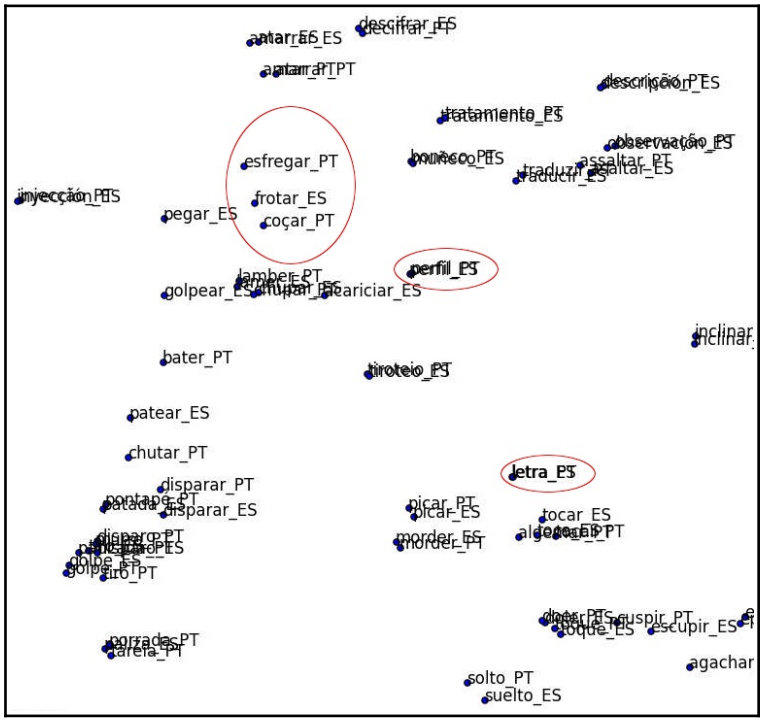

Figure 2: Visualization example of a two dimensional Portuguese–Spanish model. Circles were manually included to facilitate the interpretation.

Each point represents a word vector (whose size was automatically reduced to two dimensions using t-SNE (Maaten and Hinton, 2008)), and the distance between two points corresponds to the semantic distance between the words they represent. In this example, we can see that the circled words such as *perfil* or *letra* (*profile* and *character* in Spanish and Portuguese, respectively) occupy (almost) the same vector space in both Spanish and Portuguese. However, other words have different equivalents in the two languages: *frotar* (*to rub*, in Spanish) is very similar to both *coçar* (*to scratch*) and *esfregar* (*to scrub*) in Portuguese. Bilingual models of this kind will be used as probabilistic dictionaries to find equivalents of both the base and the collocate of the monolingual collocations.

Bilingual Alignment of Monolingual Collocations

As pointed out, the distributional model is used to search for bilingual equivalents in the previously extracted monolingual collocations of the source language (*lang_A*) and in the target one (*lang_B*). We apply the following strategy:

- First, the collocations are traversed starting from those with a higher association score in *lang_A*.
- For each one (e.g., "*autobús$_B$, coger$_C$*" *–to take the bus* in Spanish), we select its base and search for equivalents in *lang_B* using the distributional model (e.g., *autocarro*, *camioneta*, *comboio*, etc., in Portuguese).
- For words in *lang_B* with a similarity higher than a given threshold (*thres_base*), we verify whether collocations with these words as base are present in the list of *lang_B* (e.g., *"comboio$_B$, apanhar$_C$", "autocarro$_B$, apanhar$_C$", "comboio$_B$, perder$_C$"*, etc.).
- If this is the case, we calculate the cosine distance between the collocates of the bilingual equivalents (*coger* versus *apanhar* and *coger* versus *perder*).
- Finally, we select a collocation equivalent if the similarity between the collocates is also higher than a predefined threshold (in this case, *thres_coll*). In the previous example, the Spanish collocation "*autobús$_B$, coger$_C$*" will be aligned with the Portuguese *"autocarro$_B$, apanhar$_C$"*.

Note that *"comboio$_B$, apanhar$_C$"* (*to take the train*, in Portuguese) could also be aligned to "*autobús$_B$, coger$_C$*" using the proposed method. In order to

decide the best candidate among those in one of such sets, we compute the average of the distances between both bases and collocates (in *lang_A* and *lang_B*) and use the resulting number as a confidence value for each pair of aligned collocations. Thus, *"autocarro$_B$, apanhar$_C$"* will have a confidence value of 0.85 regarding the source collocation, while *"comboio$_B$, apanhar$_C$"* will have 0.78.

Evaluation

To test the method proposed in the previous section, we performed a preliminary evaluation extracting *verb-object* bilingual collocations from a Spanish (*es*)–Portuguese (*pt*) corpus.

Data

We used a part of the Spanish-Portuguese parallel corpus of OpenSubtitles2016 (Lison and Tiedemann, 2016) for extracting the monolingual (*es* and *pt*) collocations, and to learn the bilingual word-embeddings. This corpus contains sentences from movies and TV series subtitles, the alignment of which is determined by the time they occur in the movie, so it is roughly similar to a sentence level alignment.

For extracting the monolingual collocations we used about 2 million sentences, obtaining more than 150,000 verb-*dobj*-noun pairs in Spanish, and almost 215,000 in Portuguese.

For training the distributional model we used the proposed strategy in the first 11 million sentences of the corpus.

The data were processed using LinguaKit for tokenizing, lemmatizing and PoS-tagging (Garcia and Gamallo, 2015), and MaltParser for performing dependency parsing (Nivre *et al.*, 2007). The MaltParser models were previously trained on the Portuguese and Spanish Universal Dependencies treebanks (version 1.3).

Monolingual Extraction and Bilingual Alignment

The verb-*dobj*-noun extractions were ranked using mutual information. We then defined a threshold of *MI=>3* (and a frequency of *f=>6*), obtaining >1,000 collocations with highest MI values (1,024 in Spanish and 1,059 in Portuguese).

From these two sets of ≈1,000 collocations we applied our method to identify bilingual equivalents, empirically defining *thres_base* and *thres_coll* as 0.65. In those cases where the system extracted more than one equivalent, we selected the most reliable one. The proposed strategy obtained 483 collocation equivalents out of ≈1,000 input collocations.

Results

We performed an initial evaluation of both the monolingual extraction and the bilingual alignment processes. To do so, we randomly selected 50 pairs of collocation equivalents and evaluated them regarding their (i) *collocability* (whether the combination could be actually classified as a phraseological collocation), and their (ii) bilingual equivalence (whether they could be used as translations in a real scenario). The evaluation was performed by two of

the authors, while the third checked the dubious cases marked by the other reviewers.

The results of these *verb-object* collocation extractions achieved an average of 74% collocability, and 86% bilingual equivalence. In this regard, it is worth noting that some of the bilingual equivalents extracted by the system were marked as incorrect even though they appeared in real translations in the original corpus (e.g., "es:*dar un beso*=pt:*dar um abraço*" *–to kiss* in Spanish, and *to hug* in Portuguese). These cases may behave as translations in some contexts, but they were not labeled as semantically equivalent by the reviewers. If we had labeled these equivalents as correct, this process would have attained 92% precision.

Error Analysis

Although the evaluation of our strategy is at an initial stage, we have carried out a brief error analysis aimed at discovering the main sources of errors in the bilingual alignment of monolingual collocations. The errors were classified in four different groups:

1. NLP tools: Containing errors produced by the NLP tools used for analyzing the corpus. E.g., *loco* in *volver loco* (*to turn mad*, in Spanish) was incorrectly analyzed as a *dobj.*
2. Distributional semantics: A frequent issue in using distributional semantics approaches for finding synonyms (both monolexical and plurilexical) is that antonyms often occur in very similar contexts. In this regard, our strategy aligned (as bilingual equivalents) some collocations containing antonyms of the bases or of the collocates,

e.g., *“es:*esposas*$_B$,poner$_C$*=pt:*algemas*$_B$,tirar$_C$*” (*to put the handcufffs on* in Spanish, and *to remove the handcuffs*, in Portuguese).

3. Association measures: The association measures extracted several combinations that are not phraseological collocations. E.g., “*roupa$_B$, comprar$_C$*” (*to buy clothes*, in Portuguese), even though the method identified bilingual equivalents correctly (e.g., “*ropa$_B$, comprar$_C$*” in Spanish).
4. Corpora: Even if OpenSubtitles2016 is a useful resource for bilingual research, some translations are not particularly appropriate, yielding to the extraction of undesired equivalents. Also, several subtitles belonging to a specific variety appear in another variety of the same languages (for instance, many collocations extracted from the European Portuguese data were actually Brazilian combinations).

Conclusions

This paper presents a method for automatically extracting bilingual collocation equivalents from parallel corpora. First, we extract monolingual candidates of collocations using dependency parsing. Then, we train a naïve bilingual model of distributional semantics. Finally, this bilingual model is used to find equivalents of both the base and the collocate of the monolingual collocations.

Preliminary results of *verb-object* collocation extraction in Spanish and Portuguese show that this strategy can effectively discover bilingual collocation equivalents with 86% precision. Thus, using the proposed method it is feasible to automatically obtain large lists of bilingual

collocation equivalents from corpora, which in turn can be useful resources for different tasks such as machine translation as well as the compilation of teaching material for foreign languages.

Our method does not require any information about the position of the collocation in the original corpus, this being the reason why it could be applied in different resources such as comparable corpora or even non-related texts. In this respect, further work will focus on testing this strategy in comparable corpora (such as those used in Gamallo (2018)) and with different collocation patterns (such as *noun-noun, adjective-noun* or *subject-verb* collocations), extending the work presented in Garcia *et al.* (2017). Moreover, it will be interesting to apply this method to less closely related languages.

Finally, the error analysis performed allowed us to detect some of the main sources of errors that should also be taken into account in further research.

Acknowledgments

This work has been supported by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through projects FFI2016-78299-P and FFI2014-51978-C2-1-R, by a *Juan de la Cierva formación* grant (FJCI-2014-22853), and by a postdoctoral fellowship endowed by the Galician Government (POS-A/2013/191)

References

Alonso-Ramos, Margarita, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez and Sabela

Prieto González. 2010. “Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora”. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 3209–3214. Paris: European Language Resources Association (ELRA).

Altenberg, Bengt and Sylviane Granger. 2001. “The grammatical and lexical patterning of MAKE in native and non-native student writing.” *Applied Linguistics* 22: 173–195.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. “A Simple, Fast, and Effective Reparameterization of IBM Model 2”. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, 644–649. Atlanta: Association for Computational Linguistics.

Gamallo, Pablo. 2018. “Strategies to Build High Quality Bilingual Lexicons from Comparable Corpora”.

Garcia, Marcos and Pablo Gamallo. 2015. “Yet Another Suite of Multilingual NLP Tools”. In *Languages, Applications and Technologies. Communications in Computer and Information Science*, 563, ed. by José-Luis Sierra-Rodríguez, José Paulo Leal and Alberto Simões, 65–75. Switzerland: Springer. Revised Selected Papers of the Symposium on Languages, Applications and Technologies (SLATE 2015), Madrid.

Garcia, Marcos, Marcos García-Salido and Margarita Alonso-Ramos. 2017. "Using bilingual word-embeddings for multilingual collocation extraction". In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* at the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), Valencia: 21-30.

Evert, Stefan. 2008. "Corpora and collocations". In *Corpus Linguistics. An International Handbook*, volume 2, ed. by Anke Lüdeling and Merja Kytö, 1212–1248. Berlin: Mouton de Gruyter.

Fung, Pascale. 1998. "A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora". In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas. Machine Translation and the Information Soup (AMTA 1998)*, 1–17, Langhorne, Pennsylvania: Association for Machine Translation in the Americas.

Krenn, Brigitte and 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. Toulouse: Association for Computational Linguistics.

Kupiec, Julian. 1993. "An algorithm for finding noun phrase correspondences in bilingual corpora". In *Proceedings of the 31st Annual*

Meeting on Association for Computational Linguistics (ACL 1993), 17–22, Columbus, Ohio: Association for Computational Linguistics.

Lison, Pierre and Jörg Tiedemann. 2016. "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 923–929. Paris: European Language Resources Association (ELRA).

Lü, Yajuan and Ming Zhou. 2004. "Collocation translation acquisition using monolingual corpora". In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, 167–174. Barcelona: Association for Computational Linguistics.

Maaten, Laurens van der and Geoffrey Hinton. 2008. "Visualizing data using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605.

Mel'čuk, Igor. 1998. "Collocations and Lexical Functions". In *Phraseology. Theory, Analysis and Applications*, ed. by Anthony Paul Cowie, 23–53. Oxford: Clarendon Press.

Rivera, Oscar Mendoza, Ruslan Mitkov and Gloria Corpas Pastor. 2013. "A Flexible Framework for Collocation Retrieval and Translation from Parallel and Comparable Corpora". In *Proceedings of the Workshop on Multi-word*

Units in Machine Translation and Translation Technology, ed. By J. Monti, R. Mitkov, G. Corpas Pastor and V. Seretan, 18–25. Nice.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space". In *Workshop Proceedings of the International Conference on Learning Representations (ICLR 2013).* Scottsdale, Arizona. ArXiv preprint arXiv:1301.3781.

Nesselhauf, Nadja. 2004. *Collocations in a Learner Corpus*. Amsterdam/Philadelphia: John Benjamins.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty and Daniel Zeman. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1659–1666. Paris: European Language Resources Association (ELRA).

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov and Erwin Marsi. 2007. "MaltParser: A language-independent system for data-driven dependency parsing". *Natural Language Engineering* 13 (02): 95–135.

Orliac, Brigitte and Mike Dillinger. 2003. “Collocation extraction for machine translation”. In *Proceedings of Ninth Machine Translation Summit (MT Summit IX)*, 292–298, New Orleans, Louisiana.

Pecina, Pavel. 2010. “Lexical association measures and collocation extraction”. *Language Resources and Evaluation* 44 (1-2): 137–158.

Seretan, Violeta and Eric Wehrli. 2007. “Collocation translation based on sentence alignment and parsing”. In *Actes de la 14e conference sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, 401–410, Toulouse.

Smadja, Frank, Kathleen R McKeown and Vasileios Hatzivassiloglou. 1996. “Translating collocations for bilingual lexicons: A statistical approach”. *Computational linguistics* 22 (1): 1–38.

Smadja, Frank. 1992. “How to compile a bilingual collocational lexicon automatically”. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, 57–63, San Jose, CA.

Smadja, Frank. 1993. “Retrieving Collocations from Text: Xtract”. *Computational linguistics* 19 (1): 143–177.

Şulea, Octavia-Maria, Sergiu Nisioi and Liviu P. Dinu. 2016. “Using Word Embeddings to Translate Named Entities”, In *Proceedings of the Tenth*

International Conference on Language Resources and Evaluation (LREC 2016), 3362–3366. Paris: European Language Resources Association (ELRA).

Wu, Chien-Cheng and Jason S Chang. 2003. “Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses”. In *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing (ROCLING 2003)*, 1–20, Taiwan: Association for Computational Linguistics and Chinese Language Processing.