Using Linguistic Data for English and Spanish Verb-Noun Combination Identification

Uxoa Iñurrieta, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola
IXA NLP group, University of the Basque Country
usoa.inurrieta|a.diazdeilarraza|gorka.labaka|kepa.sarasola@ehu.eus

Itziar Aduriz
Department of Linguistics, University of Barcelona
itziar.aduriz@ub.edu

John Carroll
Department of Informatics, University of Sussex
j.a.carroll@sussex.ac.uk

Abstract

We present a linguistic analysis of a set of English and Spanish verb+noun combinations (VNCs), and a method to use this information to improve VNC identification. Firstly, a sample of frequent VNCs are analysed in-depth and tagged along lexico-semantic and morphosyntactic dimensions, obtaining satisfactory inter-annotator agreement scores. Then, a VNC identification experiment is undertaken, where the analysed linguistic data is combined with chunking information and syntactic dependencies. A comparison between the results of the experiment and the results obtained by a basic detection method shows that VNC identification can be greatly improved by using linguistic information, as a large number of additional occurrences are detected with high precision.

1 Introduction

Multiword Expressions (MWEs) are recurrent combinations of two or more words expressing a single unit of meaning, this meaning not always derivable directly from the meanings of the component words (Sag et al., 2002). Therefore, Natural Language Processing (NLP) tasks that need to be sensitive to lexical meaning should treat MWEs as single units. However, this is a challenging problem since many MWEs can have multimple morphosyntactic variants, which makes them difficult to recognise or generate. Examples (1)-(3) below contain *take steps*; correct translation of this MWE into another language, for instance, requires it to be recognised as a single unit[1].

(1) The Government will *take* all the necessary *steps* to prepare.

(2) They set out five important *steps* the Minister needs to *take*.

(3) What were the *steps* that should have been *taken*?

Although the most straightforward method for recognising MWEs is to attempt to match word sequences against entries in a lexicon, this method does not work for combinations that can have multiple variants. This is often the situation for verb+noun combinations (VNCs), since this kind of MWE is usually morphosyntactically flexible.

In the case of Machine Translation (MT), there are two challenges that need to be addressed concerning VNCs: (1) the detection of a given combination in the source language, and (2) its translation into the target language. If the first part fails, the words that constitute the MWE will be translated separately,

[1]Google Translate English-French and English-Spanish https://translate.google.co.uk apparently detects *take steps* as an MWE in (1) but not in (2) or (3).

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 857–867, Osaka, Japan, December 11-17 2016.

which will usually result in an incorrect translation. Then, for the second part, it is vital to have the necessary information to know what translation should be given to each VNC. A further problem arises here, since the morphosyntax of this kind of MWE varies a great deal from one language to another, meaning that it is not necessarily translated by another VNC into the target language. This problem is especially acute when the source and target languages are typologically different, as with English, Spanish and Basque[2]. This is what happens in example (4).

(4) English (EN): *get married* (V+V)
Spanish (ES): *contraer matrimonio* (V+N)
'contract marriage'
Basque (EU): *ezkondu* (V)
'(to) marry'

In this paper, we present a linguistic analysis undertaken with the aim of improving the detection of VNCs in *Matxin* (Mayor et al., 2011), a rule-based MT system which translates English and Spanish into Basque. Although we ground our study in this particular MT system, our methodology, analysis and conclusions are relevant to any kind of NLP task that needs to be sensitive to lexical meaning.

The paper is structured as follows. After discussing related work (Section 2), we present our linguistic analysis (Section 3) including: our procedure for VNC tagging, how we classify the combinations, and levels of inter-annotator agreement. In Section 4 we present a VNC detection experiment, and give the results obtained by combining linguistic information with chunking and dependency parsing. Finally, in Section 5, we draw conclusions and propose directions for future work.

2 Related Work

It is widely acknowledged that good MWE processing strategies are necessary for NLP systems to work effectively (Sag et al., 2002), since these kinds of word combinations are very frequent in both text and speech. It is estimated that the number of MWEs in an English speaker's vocabulary is of the same order of magnitude as that of single words (Jackendoff, 1997), and that at least one MWE is used per sentence on average (Sinclair, 1991).

Various classifications of MWEs have been proposed, employing different criteria to match the requirements of a particular kind of target application. Some researchers propose a binary categorisation of literal and non-literal word combinations (Birke and Sarkar, 2006; Cook et al., 2008), whereas others propose a grading containing several MWE types based on semantic idiomaticity, considered as a continuum (Wulff, 2008). Within the Meaning-Text Theory, collocations are sorted according to the notion of lexical functions (Mel'ćuk, 1998), that is, taking into account how the component words are semantically related. Furthermore, some experiments have investigated automatic methods—such as distributional similarity or word embeddings—for the task of classification, leading to fairly good results (Baldwin et al., 2003; McCarthy et al., 2003; Fazly et al., 2007; Rodríguez-Fernández et al., 2016).

In addition to MWE classification, a great deal of work has been undertaken over the last two decades on MWE acquisition (Ramisch, 2015) and identification (Li et al., 2003; Seretan and Wehrli, 2009; Sporleder and Li, 2009). Precise and detailed syntactic information is crucial for both tasks, and, at the same time, MWE identification can also help parsers obtain better results (Seretan, 2013). Moreover, accurate MWE detection is crucial for MT, since MWEs vary greatly from one language to another, and are not usually translated word for word. In the context of MT systems, Wehrli (2014) states "the non-identification of collocations dramatically affects the quality of the output".

3 Linguistic Analysis

The linguistic analysis we present here aims at improving MWE processing in MT. More specifically, we base our study on Matxin (Mayor et al., 2011), a rule-based MT system for English-Basque and

[2]Whereas English (Germanic) and Spanish (Romance) are Indo-European languages, Basque is a non-Indo-European language which moreover belongs to no known language family.

Spanish-Basque translation. One of the problems Matxin has concerning MWEs is that it currently fails to detect many instances of morphologically flexible word combinations, since it only searches for word sequences against entries in a lexicon.

As mentioned in Section 1, our study foccusses on one particular kind of MWE: verb+noun combinations (VNCs). As well as the principal constituents of a verb and a noun, we also allow for combinations containing a preposition and/or a determiner in between. Candidate combinations were first gathered from machine-readable dictionaries and were then searched for in corpora, the most frequent combinations being selected for detailed analysis.

More details about the procedure for selecting the combinations are given in the following subsections, as well as explanations of a manual tagging process, the criteria used to classify the combinations, and the overall results and conclusions drawn from this analysis. How this information is used for VNC identification is explained in Section 4.

3.1 Selection of Verb+Noun Combinations

The Spanish combinations for this study were extracted from the Elhuyar Spanish-Basque dictionary[3], and the corpus used to obtain frequency information was made up of 491,853 sentences taken from a Spanish-Basque parallel corpus containing a range of text genres. A total of 150 distinct VNCs were selected, each of which occurred more than five times as a word sequence in the corpus.

For English, our original intention was to extract combinations from the Elhuyar *English*-Basque dictionary, in part because the Basque translations would be useful for the translation process in the MT system. However, the dictionary contained too few combinations for this study, so instead we decided to use the Oxford Collocations Dictionary (Deuter, 2008). After extracting the combinations matching our grammatical pattern, we searched for them in the British National Corpus (Burnard, 2007). If the verb and the noun (and the preposition, when necessary) were found as main elements in adjacent chunks more than 500 times, the combination was selected. The final set consisted of 173 combinations in all.

3.2 Tagging Process

The combinations were tagged manually and classified along lexico-semantic and morphosyntactic dimensions, as discussed in the next sections. Although annotators looked at corpora to take decisions, the tagging was not done on instances in a corpus but on combinations out of sentential context. Therefore, each annotator gave each combination a single tag per task.

The lexico-semantic classification was done for two reasons: to determine which combinations were worth detecting and which ones should not be treated as MWEs, and because making groups depending on the combinations' idiomaticity was considered relevant for the later translation process. The morphosyntactic data, on the other hand, was analysed to be used for VNC detection (Section 4).

The tagging was performed by five linguists, all of whom are Spanish native speakers and fluent in English. Firstly, a 'super-annotator' tagged all the data, comprising a total of 323 distinct combinations in Spanish and English. Then, the data were split in four parts, and a further four annotators each tagged one of these parts, following the guidelines created for this purpose.

3.3 Lexico-Semantic Classification

The tags assigned by the annotators separated the combinations into four lexico-semantic groups, from less to more idiomatic: (1) free expressions, (2) collocations and light verb constructions, (3) metaphoric expressions, and (4) idioms. This was not an easy task, as the boundaries between one group and another are not always clearly defined. Idiomaticity is rather understood as a continuum (Wulff, 2008), and some combinations are very difficult to classify (we return to this point in Section 3.5).

Idioms (also called **opaque expressions**) are combinations in which the whole meaning cannot be understood by looking at the meanings of the words separately. Two clear examples of these would be the sentences in examples (5) and (6), which are impossible to interpret correctly without knowledge of

[3] http://hiztegiak.elhuyar.eus/

the figurative meaning of the expressions in italics.

(5) Do not believe her, she is just *pulling your leg*.
= Do not believe her, she is just *joking*.

(6) Ese chico *no se corta un pelo*, es un descarado.
'That boy *does not cut a hair*, he is shameless.'
= That boy *is never intimidated*, he is shameless.

Metaphoric expressions are not used in their literal sense either, but it is possible to understand their meaning in terms of a metaphor, as in examples (7) and (8).

(7) He did not come to the meeting and the boss *had a word* with him.
= He did not come to the meeting and the boss *spoke* with him.

(8) Las experiencias de ese tipo *dejan huella*.
'These kinds of experiences *leave (a) mark*.'
= These kinds of experiences have a very significant effect (on people's life).

Unlike the combinations in examples (5)–(8), those in examples (9) and (10) are easily understandable on the basis of their component words; they belong to the group of **collocations and light verb constructions**. Collocations are defined as lexically constrained and recurrent combinations of words which are in a given syntactic relation (Evert, 2008; Bartsch, 2004). When they are VNCs, the verb is often a very common word which is semantically bleached—meaning that it loses its usual sense to a certain extent (Butt, 2010). These kinds of combinations are called light verb constructions (LVCs). Examples (9) and (10) would be classified in this group.

(9) Volunteers *gave support* to disadvantaged children.

(10) La educación *tiene* vital *importancia* para los niños desaventajados.
'Education *has* vital *importance* for disadvantaged children.'

Finally, **free expressions** are groups of words that can be combined freely, that is, following the standard lexical and grammatical rules of a given language. These kinds of expressions are not idiomatic, and are thus not considered MWEs, as in examples (11) and (12). Therefore, the combinations sorted in this group by the annotators were excluded for the later detection experiment (Section 4).

(11) They *are using a* new *technique* now.

(12) Este año *iremos a un lugar* diferente.
'This year *we will go to a* different *place*.'

As mentioned in Section 3.2, we consider that classifying the VNCs is relevant for translation. Our hypothesis is that the kind of translation a VNC should be given is often dependent on its lexico-semantic class. For instance, the combinations we have analysed so far suggest that, although idioms are usually translated by other (morphosyntactically equivalent or non-equivalent) idioms into the target language, they are unlikely to receive a word-for-word translation (see example (13)). On the other hand, in collocations, the noun is very likely to receive a direct translation, whereas the verb is often given a translation other than the one expected when it is not part of the collocation (see example (14)).

(13) EN: *pull* (somebody)*'s leg*
ES: *tomar el pelo* (a alguien)
'take (somebody)'s hair'
EU: (norbaiti) *adarra jo*

'play (somebody) the horn'

(14) EN: *take steps*
ES: *dar pasos*
'give steps'
EU: *pausoak eman*
'give steps'

We will not focus on the correlation between VNC classes and their translation in this paper. However, we do consider it an interesting topic for future investigation.

3.4 Morphosyntactic Classification

As well as the lexico-semantic tagging described above, we examined morphosyntactic features of combinations to classify them into three groups: (1) fixed combinations, (2) semi-fixed combinations, and (3) morphosyntactically free combinations. The annotators had to consider five questions to determine how fixed the combinations were:

- Does the noun phrase (NP) have a determiner? (always/never/optional)
- Is the NP singular or plural? (singular/plural/optional)
- Can there be a modifier (i.e. an adjective) inside the NP? (yes/no)
- Can the verb and the NP be separated by other words? (yes/no)
- Can the order of the elements be altered? (yes/no)

A given VNC needed to be classified as completely free when: the determiner and the number of the NP were marked as optional; there could be a modifier inside the NP; the verb and the NP could be separated by other words; and the order of the elements in the expression was judged to be alterable. When some of the answers were different to these, the combination had to be marked as semi-fixed, and as completely fixed if all the answers were different (that is, when the syntactic variability of the VNC was completely restricted).

None of the combinations was tagged as **fixed** by both the super-annotator and the second annotator, but this was not surprising, as VNCs which do not accept any kind of morphosyntactic variation are extremely rare. Usually, they can undergo some alterations (**semi-fixed expressions** as in examples (15) and (16)), or they can even be completely flexible (**morphosyntactically free expressions** as in examples (17) and (18)).

(15) be in love; be always in love; *be in the love; *be in loves.

(16) dar paso (a algo); dar siempre paso (a algo); *dar pasos (a algo)
'give way (to sth); always give way (to sth); *give ways (to sth)'

(17) cause a problem; cause two important problems; the problem was caused

(18) hacer un favor; hacer un gran favor; hacer dos favores; el favor que se hizo
'do a favour; do a big favour; do two favours; the favour that was done'

As these features have a direct impact on the detection of the combinations, the answers to the above-mentioned questions were also specified by the super-annotator one by one, so that this information could later be used to improve detection (see Section 4).

	Lexico-semantics	Morphosyntax
Agreement	70.52%	84.39%
κ	0.55	0.55

Table 1: IAA for English VN combinations.

	Lexico-semantics	Morphosyntax
Agreement	76.00%	81.34%
κ	0.63	0.61

Table 2: IAA for Spanish VN combinations.

3.5 Inter-Annotator Agreement

Inter-annotator agreement (IAA) was measured in two ways: the percentage of combinations in which the annotators agreed, and Cohen's Kappa, κ (Cohen, 1960).

As shown in Tables 1 and 2, annotator agreement was 70% to 84% for all tagging tasks and for both languages. With κ scores between 0.55 and 0.63, we conclude that the task is coherent and that the tagging results are usable for further investigation. The lexico-semantic IAA for English is similar to the IAA obtained in previous related work (Fazly et al., 2007; Vincze, 2012), and for Spanish it is appreciably higher.

Consistent with previous work (Seretan, 2013), we found that in our selection of 323 of the most frequently occurring VNCs in Spanish and English, collocations and LVCs are the most common type of combination, and that opaque expressions (idioms) are very scarce.

We also found that the combinations that led to disagreements among annotators were not classified in random groups, but were almost always in classes lexico-semantically (and morphosyntactically) close to each other (see Tables 3 and 4). Indeed, only a few combinations were classified in two groups that were not directly adjacent on the idiomaticity continuum. This provides further evidence that MWEs form a continuum of idiomaticity with no clear boundaries between MWE types (McCarthy et al., 2003).

		Other annotators			
		Idiom	**Metaphoric**	**Colloc/LVC**	**Free**
Super-annotator	**Idiom**	**0**	0	0	0
	Metaphoric	1	**24**	0	1
	Colloc/LVC	0	12	**73**	22
	Free	0	2	13	**25**

Table 3: Confusion matrix for English showing lexico-semantic tag agreement between the annotators.

		Other annotators			
		Idiom	**Metaphoric**	**Colloc/LVC**	**Free**
Super-annotator	**Idiom**	**1**	0	1	0
	Metaphoric	0	**20**	2	1
	Colloc/LVC	0	8	**69**	15
	Free	0	1	8	**24**

Table 4: Confusion matrix for Spanish showing lexico-semantic tag agreement between the annotators.

4 Identification Experiment

To test whether the analysed morphosyntactic data (see Section 3.3) could improve MWE detection, we undertook an experiment where three **identification methods** were combined and compared: (A)

the old one, used by Matxin, which searched only for word sequences; (B) a second one, based on the analysed linguistic data and automatically-produced chunking information; and (C) a third one, based on the analysed linguistic data and automatically-produced syntactic dependencies. Depending on how morphosyntactically fixed a given combination was, more or less linguistic restrictions were applied to identify them.

The **experimental set** was made of the combinations presented in Section 3, excluding the ones tagged as completely free by the super-annotator (Section 3.3). The final set consisted of 117 combinations in Spanish and 133 in English.

4.1 Results of the English Experiment

The corpus used for the experiment on English VNCs was the British National Corpus (Burnard, 2007), and chunking and dependency information was computed by the Stanford parser (Manning et al., 2014). A total of 152,051 occurrences of the 133 VNCs were identified by combining all three methods, 78.92% of which were not detected by method A, currently used for English-Basque translation in Matxin. Figure 1 shows the percentages of all the instances detected by each of the methods.

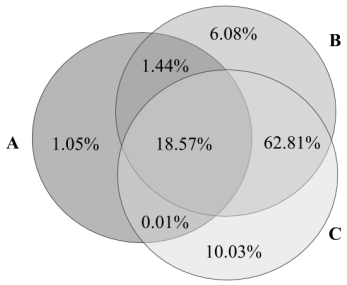

Figure 1: Percentages of English VNC occurrences identified by each method. (For clarity, areas are not drawn in scale with percentages)

We cannot calculate recall since our evaluation dataset contains only the occurrences identified collectively by the three methods, and it is almost certain that some occurrences of the VNCs under investigation were not detected. For future work, we would need to use MWE-tagged corpora to calculate recall, such as the Prague Czech-English Dependency Treebank (Uresova et al., 2013). In any case, the results obtained clearly show that the number of identified occurrences is increased considerably by using linguistic data specific to VNCs, as well as confirming that VNCs are commonly used in multiple morphosyntactic variations, as only 21.08% of the instances could be identified by searching for word sequences against entries in a lexicon.

To estimate the precision of VNC detection, we considered a representative sample of the full set, and evaluation was carried out manually by linguists. The precisions of methods B and C were not as good as that of method A. However, the evaluation on instances identified by both B and C methods reveals that detection quality is still very high when linguistic data specific to VNCs is combined with parsing (the second row of scores in Table 5).

	Additional VNCs %	**Precision**
Method A (in all)	21.08%	99%
Method B+C but not A	62.81%	96%
Method B only	6.08%	70%
Method C only	10.03%	79%

Table 5: Identification precision for the additional VNC occurrences detected in English

The least satisfactory results were those obtained by method B. When verifying the results, we noticed that the vast majority of false instances detected were light verb constructions (LVCs) containing verbs

that could also work as auxiliaries. In example (19), for instance, *have influences* is erroneously detected since *influence* is mis-analysed as the object of *have* rather than the subject of *have been likened*.

(19) These *influences have* also been likened to the forces effected by a millenarian journey to a new faith...

The overall improvement we obtained was substantial, as expected from previous work. Li et al. (2003) report an F-score improvement of 9 percentage points (86.9% to 95.6%) when using parsers and hand-crafted lexical patterns to identify phrasal verbs in English, as well as a precision improvement of 8 percentage points (90% to 98%). In our case, precision falls from 99% to 93% when combining all three methods, but the number of new instances detected suggests an appreciable increase in recall. As we already mentioned, MWE-annotated corpora would be needed to calculate recall and F-score and compare our results to those reported by other authors.

4.2 Results of the Spanish Experiment

For the experiment on Spanish, VNCs were searched in 15,182,385 sentences taken from the parallel English-Spanish corpus made public for the shared task in the ACL 2013 workshop on statistical MT[4], and the parser used was Freeling (Padró and Stanilovsky, 2012). A total of 433,092 occurrences were identified, 27.80% of which were not detected by method A (the percentages of the combinations identified by each method are shown in Figure 2). Consistent with the results obtained for English, this further reveals that the morphosyntactic data we took into account (Section 3.4) is very relevant for VNC identification.

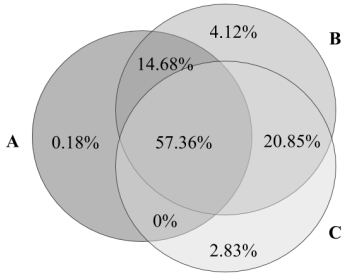

Figure 2: Percentages of Spanish VNC ocurrences identified by each method

Furthermore, as well as the quantity improving considerably, the manual evaluation reveals that the quality of our method is also very satisfactory. As is shown in Table 6, methods B and C, although not as precise as method A, got very good precision scores.

	Additional VNCs %	**Precision**
Method A (in all)	72.20%	99%
Method B+C but not A	20.85%	97%
Method B only	4.12%	93%
Method C only	2.83%	83%

Table 6: Identification precision for the additional VNCs detected in Spanish

As the corpora and parsers we used were different for English and Spanish, the experiments in both languages are not really comparable. However, it is evident that the improvement obtained for English was considerably higher than the one obtained for Spanish. Taking into account that the Freeling and Stanford parsers work in similar ways and that the manual tagging of the VNCs was done following the same criteria, this difference could suggest that syntactic variations of VNCs other than the canonical form are more common in English than in Spanish. One of the possible reasons for this could be the

[4] http://www.statmt.org/wmt13/translation-task.html

different word order inside NPs in both languages. In Spanish, adjectives can either precede or follow the head noun, whereas in English adjectives are almost never placed after the noun: *importantes pasos* or *pasos importantes* vs. *important steps* but not **steps important*. An exhaustive analysis would be needed to verify this hypothesis or identify other possible reasons.

5 Conclusions and Future Work

Morphosyntactically flexible MWEs constitute a problem for NLP systems, which often fail to process these kinds of word combinations correctly. In this paper, we presented a linguistic analysis undertaken with the aim of improving the identification of VNCs, as well as an experiment which shows how linguistic data can improve identification results greatly.

Firstly, we classified a selection of frequent VNCs in English and Spanish, following both lexico-semantic and morphosyntactic criteria. A total of 323 distinct combinations (173 in English and 150 in Spanish) were tagged by several annotators, with very reasonable inter-annotator agreement scores (κ 0.55 to 0.63). We noted moreover that the combinations that led to disagreements among annotators were always tagged in groups that were lexico-semantically and morphosyntactically close to each other, which gives further evidence that idiomaticity should be viewed as a continuum. More detailed morphosyntactic information was also specified for each combination, and this information was then used to improve VNC identification.

Our experiment confirmed that specific linguistic data about VNCs is useful for the identification of this kind of word combination, as it allows for the recognition of occurrences that do not match a combination's canonical form. Indeed, a large number of instances that were not identified by searching for fixed word sequences could be identified by combining linguistic data with chunking information and syntactic dependencies, with fairly good precision scores (79% to 97%).

Building on the satisfactory results obtained, we will test our methods in the context of MT, and we will keep analysing more VNCs. The next step will be to explore what kind of data is needed for an adequate translation of VNC combinations within MT systems. In addition, we intend to investigate how semantic information can be used within the translation process.

Acknowledgements

Uxoa Iñurrieta's doctoral research is funded by the Spanish Ministry of Economy and Competitiveness (BES-2013-066372). The work was carried out in the context of the SKATeR (TIN2012-38584-C06-02) and TADEEP (TIN2015-70214-P) projects. We thank Diana McCarthy for helpful advice, as well as Begoña Altuna, Nora Aranberri, Ainara Estarrona and Larraitz Uria for helping us with the tagging work.

References

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

Sabine Bartsch. 2004. *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Gunter Narr Verlag, Tübingen.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 329–336.

Lou Burnard (ed.) 2007. *Reference Guide for the British National Corpus (XML Edition)*. Oxford University Computing Services.

Miriam Butt. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker, and Mark Harvey (eds.), *Complex Predicates: Cross-linguistic Perspectives on Event Structure*. Cambridge University Press, 48–78.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, 19–22.

Margaret Deuter (ed.) 2008. *Oxford Collocations Dictionary for Students of English*. Oxford University Press.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD dissertation, IMS, University of Stuttgart.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*. Mouton de Gruyter, Berlin, 1212–1248.

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

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Wei Li, Xiuhong Zhang, Cheng Niu, Yuankai Jiang, and Rohini Srihari. 2003. An expert lexicon approach to identifying English phrasal verbs. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics: Long Papers*, 513–520.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, 25(1): 53–82.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 73–80.

Igor A. Mel'ćuk. 1998. Collocations and lexical functions. In Anthony P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*. Oxford University Press, 23–53.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2473–2479.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer International Publishing, Switzerland.

Sara Rodríguez-Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers*, 499–505.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 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'02)*. Springer Berlin Heidelberg, 1–15.

Violeta Seretan and Eric Wehrli. 2009. Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation*, 43(1): 71–85.

Violeta Seretan. 2013. On collocations and their interaction with parsing and translation. *Informatics*, 1(1): 11–33.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Caroline Sporleder, and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 754–762.

Zdenka Uresova, Jana Sindlerova, Eva Fucikova, and Jan Hajic. 2013. An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. In *Proceedings of the Ninth Workshop on Multiword Expressions (MWE 2013)*, 58–63.

Veronika Vincze. 2012. Light verb constructions in the SzegedParalellFX English-Hungarian parallel corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2381–2388.

Eric Wehrli. 2014. The relevance of collocations for parsing. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014)*, 26–32.

Stefanie Wulff. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. Continuum, London, New York.