Pay attention when you pay the bills. A multilingual corpus with dependency-based and semantic annotation of collocations.

Marcos Garcia
Universidade da Coruña
Grupo LyS, Dpto. de Letras
Campus da Zapateira, Coruña

Marcos García-Salido
Universidade da Coruña
Grupo LyS, Dpto. de Letras
Campus da Zapateira, Coruña

Susana Sotelo
Universidade da Coruña
Grupo LyS, Dpto. de Letras
Campus da Zapateira, Coruña

Estela Mosqueira
Universidade da Coruña
Grupo LyS, Dpto. de Letras
Campus da Zapateira, Coruña

Margarita Alonso-Ramos
Universidade da Coruña
Grupo LyS, Dpto. de Letras
Campus da Zapateira, Coruña

{marcos.garcia.gonzalez,marcos.garcias
susana.sotelo,estela.mosqueira,margarita.alonso}udc.gal

Abstract

This paper presents a multilingual corpus with semantic annotation of collocations in English, Portuguese, and Spanish. The whole resource contains $155k$ tokens and $1,526$ collocations labeled in context. The annotated examples belong to three syntactic structures (*adjective-noun*, *verb-object*, and *nominal compounds*), and represent 60 lexical functions in the Meaning-Text Theory (e.g., *Oper*, *Magn*, *Bon*, etc.). Each collocation was annotated by three linguists and the final resource was revised by a team of experts. The resulting corpora, which are freely available, can serve as a basis to evaluate different approaches for collocation identification and classification in three languages, which in turn can be useful for different NLP tasks such as natural language generation or understanding.

1 Introduction

The automatic identification of collocations, as well as other multiword expressions (MWEs), is crucial for many natural language processing (NLP) tasks, since their linguistic behaviour differs from other combinations of words (Mel'čuk, 1995; Sag et al., 2002; Ramisch and Villavicencio, 2018). On the one hand, approaches to natural language generation may take advantage of collocational information to produce natural utterances with the desired meanings (Wanner et al., 2010; Lareau et al., 2011). In this regard, while different adjectives such as *heavy* and *strong* can convey basically the same meaning (e.g., 'intensification' in *heavy load* and in *strong fragrance*), *great* has different senses in *great loss* and in *great time* (with 'intensification' and 'positive' meanings, respectively). On the other hand, to interpret the meaning of a sentence, a system should take into account the properties of these expressions: for instance, the meaning of the verb *[to] take* in the collocation *take [a] cab* is different from the same verb in a free combination such as *take [a] pencil*, so natural language understanding or abstract meaning representation systems could benefit from the correct identification of collocations (Bonial et al., 2014; O'Gorman et al., 2018). It is worth mentioning that collocations are pervasive and frequent in all domains and text typologies, so their correct interpretation should be critical to progress in the automatic processing of natural languages.

The concept of collocation was formalized in the Meaning-Text Theory as a combination of two lexical units (LUs) where one of them (the BASE, e.g., *attention* in the collocation *pay attention*) is freely selected due to its meaning, while the selection of the other one (the COLLOCATE, e.g., *[to] pay*) is restricted by the former (Mel'čuk, 1995). Under this theory, lexical functions (LF) represent a relation between a LU (the base) and a set of expressions (the potential collocates) (Mel'čuk, 1996, 1998; Wanner, 1996). For instance, the LF $Oper$ means 'to carry out', so we could define $Oper$(*picture*)=*[to] take* to formalize the collocation *take a picture*. Similarly, the adjective–noun collocation *loud screech* can be represented as $Magn$(*screech*)=*loud*, where the lexical function $Magn$ denotes 'intensification'.

The automatic identification of collocations has deserved a substantial number of works of different researchers from NLP and computational linguistics as well as from lexicography and corpus linguistics (Evert, 2008; Pecina, 2010; Gries, 2013). Most approaches rely on statistical association measures (AMs), both symmetrical and directional, and recent works incorporate distributional

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4012–4019
Florence, Italy, July 28 - August 2, 2019.

semantics to automatically identify the collocate of a given base and LF, or to classify the compositionality of MWEs including collocations (Wanner et al., 2006; Carlini et al., 2014; Rodríguez-Fernández et al., 2016; Cordeiro et al., 2019). To evaluate the extraction, some researchers use manual selection of true collocations from ranked lists, while others take advantage of examples extracted from collocation dictionaries. However, most of these approaches are carried out only in one language, and they do not always permit to obtain precise recall values. Moreover, they usually do not include semantic information.

Taking the above into account, this paper attempts to fill this gap by releasing a freely available multilingual corpus of English, Portuguese, and Spanish with manual annotation of collocations and their lexical functions. The whole resource, annotated by five experts, has more than $155k$ tokens and $1,526$ collocations classified into 60 lexical functions. For each language, we provide both the labeled data of each annotator as well as the gold-standard data.[1]

2 Related Work

Different statistical methods have been applied to automatically identify and classify collocations from corpora. Studies such as Wanner et al. (2006), Wanner et al. (2016), or Gelbukh and Kolesnikova (2010) train statistical models using Spanish data (from EuroWordNet, from the DiCE dictionary, and using a Spanish corpus, respectively). For French, Fonseca et al. (2017) explore the combination of dependency parsing with a lexical network based on lexical functions.

The semantic classification of base-collocate pairs allowed for implementing multilingual natural language generation systems which take advantage of lexical functions to select the most appropriate combinations for each context (Wanner et al., 2010). In this regard, Lareau et al. (2011) propose a methodology to use lexical functions in Lexical Functional Grammar.

With respect to the extraction process, there have been a large number of studies focusing on the automatic identification of collocations in corpora. In this regard, most approaches have relied on statistical association measures applied both to windows of n-grams (Church and Hanks, 1990; Smadja, 1993; Pecina, 2010), and to syntax-based dependency triples (Seretan, 2011; Carlini et al., 2014; Garcia et al., 2017; Uhrig et al., 2018). In Rodríguez-Fernández et al. (2016) it is presented a method to retrieve potential collocates of a given LF and a base. Other studies address the identification process as a classification problem. Karan et al. (2012) take advantage of a set of true positive and true negative collocations to evaluate machine learning algorithms which use, among others, features based on association values.

To evaluate such methods, some authors carry out a manual review of the *n* best combinations of candidate collocations lists, ranked by a given AM (Seretan and Wehrli, 2006; Garcia, 2018). A different approach consists of collecting an inventory of true collocations (usually from existing dictionaries), which is then used to compare the performance of various AMs (Evert and Krenn, 2001; Pearce, 2002; Pecina, 2010; Kilgarriff et al., 2014; Evert et al., 2017). Concerning the available data with collocational information, it is worth noting that a vast majority of the resources are dictionaries and lexicons often targeted at language learners (Benson et al., 1986; Alonso-Ramos et al., 2010). From a different perspective, initiatives such as PropBank and abstract meaning representation also provide corpora with semantic annotation of MWEs, some of which may be considered collocations (Banarescu et al., 2013; Bonial et al., 2014; O'Gorman et al., 2018).

The approach to evaluate the process of collocation extraction proposed here consists of using a gold-standard corpus with manual annotation of such expressions. On the one hand, this allows for accurate precision and recall values to be obtained, also taking into account ambiguous combinations which may be collocations or not depending on the context. On the other hand, a gold-standard enables the research community to evaluate different strategies in a more comparable way. In this regard, the 2017 and 2018 PARSEME Shared Tasks released multilingual corpora with annotation of verbal MWEs (Savary et al., 2017; Ramisch et al., 2018). Even if the initial objectives of these shared tasks differ from ours (they annotate idioms, verb-particle constructions and other non-collocation MWEs), some of the units actually intersect with the expressions we want to identify. Thus, we rely on these corpora to initiate the construction of a multilingual corpus with

[1]The corpora are freely available at the following url: http://www.grupolys.org/~marcos/pub/collocations.zip

dependency-based and semantic annotation of collocations.

3 Source Data and Annotation

This section describes both the source data used to build our multilingual corpora as well as the annotation guidelines and procedure.

3.1 Corpora

We decided to take advantage of three subcorpora of the edition 1.1 of the PARSEME Shared Task, which include annotation of different verbal multiword expressions in 20 languages (Ramisch et al., 2018). Since we understand collocations as lexically restricted combinations of words, some of the MWEs annotated in the PARSEME corpora are also useful for our objectives (see Section 3.2).

Our main purpose is to provide datasets to evaluate unsupervised strategies for extracting collocations, so we selected the *test* datasets for Portuguese ($58k$ tokens) and Spanish ($39k$ tokens), and the *train* corpus for English ($53k$ tokens), because the test data for this language are fewer. These corpora are annotated with Universal Dependencies (Nivre, 2015) and released in *.cupt* format[2] (an extension of *.conllu*).[3]

3.2 Annotation Guidelines

In general, our annotation follows Mel'čuk (1996) with specific guidelines for each collocation type. Also, we tried to be compatible with the PARSEME principles with a view to combine both annotations. As our strategy relies on dependency analysis to obtain candidate combinations (which are subsequently revised), we defined annotation guidelines for three syntactic patterns of collocations (for each pattern, a set of tests for identifying collocations was included). In the following we present some examples of the most productive lexical functions in each pattern (see Appendix A for the whole list of LFs):

Adjective-noun (*amod*): collocations where the adjective has a function of intensification and attenuation ($Magn$: *high priority*, or $AntiMagn$: *weak resource*), expresses a positive or negative consideration from the speaker (Bon: *great event*, $AntiBon$: *unfortunate mistake*), or conveys a specific sense ($NonStandard$) in combination with the noun (e.g., *dark sorcerer*) (Mel'čuk, 1996).

[2] http://multiword.sourceforge.net/cupt-format/

[3] https://universaldependencies.org/format.html

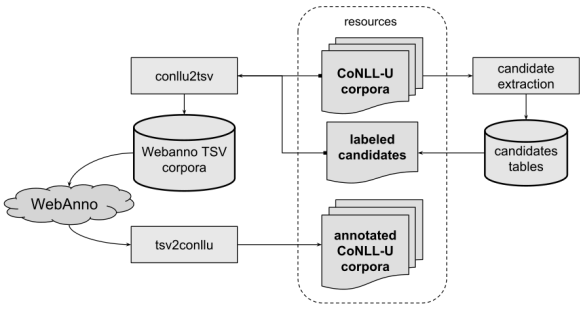

Figure 1: Annotation workflow.

Nominal compounds (*nmod* and *compound*): nominal compounds where the head of the relation may express the notion of 'head of a collective' (Cap: *police chief*), 'a part of' ($Sing$: *member [of the] government*), or of a 'group' or the 'totality' of the dependent ($Mult$: *wolf pack*).

Verb-object (*obj*): verb-object collocations occur between a predicative noun (Polguère, 2011) depending of a verb which either do not contribute to the meaning of the combination ($Oper$: *[to] have breakfast*), or express causation ($CausOper$, *[to] cause damage*) or a specific meaning in combination with the base ($NonStandard$: *[to] shake hands*). As some of these types were covered by PARSEME (labeled as light verb constructions), we revised each case and added their LFs. Some structures such as *verb-obj* candidates occurring in passive voice as subjects (e.g., *[the] damage was caused*) or relative constructions (which do not have a direct dependency between the lexical base and the collocate) were not extracted.

A simplified version of the guidelines can be accessed at the following url: http://www.grupolys.org/~marcos/collocations/guidelines.html.

3.3 Annotation Procedure

In order to facilitate the labeling by each annotator as well as to speed up the whole process we defined the following procedure (see Figure 1):

We start by extracting the instances of the desired relations (*amod, nmod, compound* and *obj*) from the referred corpora, and arrange them into triples (*base;collocate;relation*). Despite the fact that most collocation extraction approaches set up a frequency threshold to avoid noisy and less frequent combinations, we revised every single instance of each dependency relation.

Then, for each language and collocation pattern,

base	collocate	deprel	label	freq	Examples	LF
contract	global	amod	n	2	Examples	
time	great	amod	y	4	Examples	Bon
cure	quick	amod	d	1	Examples	

time ~ great (amod)

id	text
2416	You 'll have a **great** **time** , so no worries there .
2810	Looks like the kids had a **great** **time** !

Figure 2: Example of an annotation sheet (top) and HTML view (bottom) of the examples.

we created a sheet including the triples, a link to an automatically generated HTML site with examples from the corpora, and annotation fields (see Figure 2). Each collocation candidate was revised by three experts (native or near-native speakers of the target languages) which classified them as *collocation*, *non-collocation*, or *doubt*.[4] Doubts include (i) combinations which are collocations in some examples but not in others, (ii) collocations which include internal MWEs (e.g., verb-particle constructions), and consequently they need a specific annotation, and (iii) cases in which the annotator is not sure about the collocational status of the candidate. After that, we created a final sheet including the most frequent classification for each candidate. Those cases in which there is no agreement (i.e., each annotator selected a different label, or there is more than one *doubt*) were also marked as *doubt*. The dubious cases in these final sheets were revised in common by the whole team of language experts, who decided on the LF of each collocation.

Finally, we automatically transferred the annotation to a new version of the initial corpora, and convert it to the *.tsv* format required by WebAnno (Eckart de Castilho et al., 2016). Using this tool, we corrected those special cases (MWEs bases and collocates, combinations which are collocations only in some contexts, etc.) and performed a general revision of the corpus. At the end of this process, we generated the final corpus in *.conllu* format using the original resources and the *.tsv* files.

It is worth mentioning that we did not perform a systematic evaluation of the syntactic analysis of each corpus. In this respect, we could miss some true collocations incorrectly labeled with a wrong dependency relation. However, the annotated cases were manually checked, and therefore they have a correct syntactic analysis (except for human errors).

[4]Light verb constructions already annotated in the original corpus were initially marked as *doubt*, so each annotator also revised again these cases.

id	*token*	*h*	*dep*	*collocational information*
1	He	2	nsubj	_
2	took	0	root	1
3	a	5	det	_
4	deep	5	amod	2
5	breath	2	obj	1:obj_Oper1;2:amod_Magn

Figure 3: Example of two annotated collocations (in the last column) with the base *breath* in a simplified *.conllu* format (where *h* refers to the *id* of the syntactic head, and *dep* to the dependency relation): $Oper1$(*breath*)=*[to] take*, $Magn$(*breath*)=*deep*.

This resulting corpus contains the collocational annotation in the last column of the *.conllu* file (see an example in Figure 3). On the one hand, the base of each collocation has a numerical *id* followed by the syntactic pattern (e.g., *obj, amod*) and by its lexical function. On the other hand, the collocate is labeled with the same *id* as the base it depends on. In blended collocations (as in the example), the base contains information about both combinations separated by a semicolon.

4 Final Resources and Results

The final multilingual corpus has $155,794$ tokens and $1,526$ annotated collocations ($1,394$ unique) Table 1 includes the number of revised candidates and annotated collocations for each language and dependency relation. As expected, adjective-noun and verb-object collocations were the most pro-

Pattern	English		Portuguese		Spanish		Total	
	Cand.	Colloc.	Cand.	Colloc.	Cand.	Colloc.	Cand.	Colloc.
amod	1,841	272	1,540	199	1,557	149	4,938	620
nmod	813	28	1,529	91	1,235	76	3,577	195
obj	1,495	184	1,572	250	914	145	3,981	579
Total	4,149	484	4,641	540	3,706	370	12,496	1,394

Table 1: Collocation candidates and unique annotated combinations per language and dependency pattern.

Pattern	English	Port.	Spa.	Total
amod	0.548	0.482	0.537	0.526
nmod	0.400	0.388	0.330	0.370
obj	0.541	0.706	0.630	0.632
Total	0.532	0.566	0.571	0.541

Table 2: Multi-k inter-annotator agreement per dependency pattern and language.

ductive ones, and nominal compounds combinations were less frequent.

We computed multi-k inter-annotator agreement (Davies and Fleiss, 1982; Artstein and Poesio, 2008) for each language and relation (Table 2), with values between $k = 0.370$ and $k = 0.706$. The higher agreement occurs in verb-object collocations, while in nominal compounds it was lower.

Once the final sheets (for each relation and language) were created, a total of 447 combinations (3.6%) were labeled as *doubt* (there was no agreement between the annotators). Out of these, 260 (58.2%) were finally considered true collocations by the team of experts. Among the most frequent disagreements we found adjective-noun pairs for which the annotators doubted whether they were technical terms (e.g., *light cluster*), nominal compounds in which one of the nouns seems lexically selected by the other (e.g., *golf tournament*), and verb-object combinations in which the noun could be predicative and the verb has scarce lexical content, but lacks a single-word verb equivalent (e.g., *tener velocidad* 'have speed' in Spanish). In the latter group, we harmonized their annotation in the three languages.

The final resource includes a total of 60 lexical functions, some of them complex (e.g., $Magn + AntiBon$), and not all of them in every language (i.e., less frequent LFs appear only in one or two corpora). The most frequent ones are $Oper1$, $Magn$, Bon, and $NonStandard$ (see Appendix A for the full list of LFs per language).

5 Conclusions and Further Work

This paper presented a multilingual corpus with manual annotation of collocations and their lexical functions in English, Portuguese, and Spanish. The resource contains $155k$ tokens and $1,526$ collocations classified into 60 lexical functions. Each collocation candidate was revised by three language experts, and those which were dubious were corrected by the whole team of annotators.

We release both the final corpus of each annotator as well as the gold-standard resource in *.conllu* format. This dataset can serve as a basis to evaluate systems designed to automatically extract collocations and identify their lexical functions, which in turn may be useful for different NLP and corpus linguistics tasks. As we provide resources for three languages (and with different dependency relations), the corpora can be also useful to verify whether some methods behave similarly or not in each language and syntactic pattern.

In further work we plan to carry out a multilingual alignment of the collocations in each language. This process, also enlarged with other multilingual equivalents, will generate a new dataset for evaluating the automatic translation of this type of multiword expressions.

Acknowledgments

This research was supported by a 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation), by Ministerio de Economía, Industria y Competitividad (project with reference FFI2016-78299-P), and by the Galician Government (Xunta de Galicia grant ED431B-2017/01). Marcos Garcia has been funded by a Juan de la Cierva-incorporación grant (IJCI-2016-29598), and Marcos García-Salido by a post-doctoral grant from Xunta de Galicia (ED481D-2017-009).

References

Margarita Alonso-Ramos, Alfonso Nishikawa, and Orsolya Vincze. 2010. DiCE in the web: An online Spanish collocation dictionary. In *ELexicography in the 21st Century: New Challenges, New Applications: Proceedings of ELex 2009*, volume 7, pages 369–374. Presses univ. de Louvain.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4).

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186. Association for Computational Linguistics.

Morton Benson, Evelyn Benson, and Robert Ilson. 1986. *The BBI combinatory dictionary of English: A guide to word combinations*. John Benjamins Publishing.

Claire Bonial, Meredith Green, Jenette Preciado, and Martha Palmer. 2014. An approach to take multiword expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 94–98. Association for Computational Linguistics.

Roberto Carlini, Joan Codina-Filba, and Leo Wanner. 2014. Improving collocation correction by ranking suggestions using linguistic knowledge. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 1–12, Uppsala. LiU Electronic Press.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84. The COLING 2016 Organizing Committee.

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

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Mark Davies and Joseph L Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An international handbook*, volume 2, pages 1212–1248. Mouton de Gruyter, Berlin.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France. Association for Computational Linguistics.

Stefan Evert, Peter Uhrig, Sabine Bartsch, and Thomas Proisl. 2017. E-VIEW-affilation–A large-scale evaluation study of association measures for collocation identification. In *Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch*, pages 531–549.

Alexsandro Fonseca, Fatiha Sadat, and François Lareau. 2017. Combining dependency parsing and a lexical network based on lexical functions for the identification of collocations. In *International Conference on Computational and Corpus-Based Phraseology*, pages 447–461. Springer.

Marcos Garcia. 2018. Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 319–342. Language Science Press, Berlin.

Marcos Garcia, 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)*, pages 21–30, Valencia, Spain. Association for Computational Linguistics.

Alexander Gelbukh and Olga Kolesnikova. 2010. Supervised learning for semantic classification of spanish collocations. In *Mexican Conference on Pattern Recognition*, volume 6256 of *Lecture Notes in Computer Science*, pages 362–371. Springer-Verlag.

Stefan Th. Gries. 2013. 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18(1):137–165.

Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Evaluation of classification algorithms and features for collocation extraction in croatian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 657–662, Istanbul, Turkey. European Language Resources Association (ELRA).

Adam Kilgarriff, Pavel Rychlý, Milos Jakubicek, Vojtěch Kovář, Vit Baisa, and Lucia Kocincová. 2014. Extrinsic corpus evaluation with a collocation dictionary task. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 545–552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Francois Lareau, Mark Dras, Benjamin Borschinger, and Robert Dale. 2011. Collocations in multilingual natural language generation: Lexical functions meet lexical functional grammar. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 95–104.

Igor Mel'čuk. 1995. Phrasemes in language and phraseology in linguistics. In Martin Everaert, Erik-Jan van der Linden, André Schenk, and Rob Schreu, editors, *Idioms: Structural and psychological perspectives*, pages 167–232. Hillsdale: Lawrence Erlbaum Associates.

Igor Mel'čuk. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Corpus Linguistics*, pages 37–102. John Benjamins Publishing.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony Paul Cowie, editor, *Phraseology. Theory, analysis and applications*, pages 23–53. Clarendon Press, Oxford.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, volume 9041 of *Lecture Notes in Computer Science*, pages 3–16. Springer.

Tim O'Gorman, Sameer Pradhan, Martha Palmer, Julia Bonn, Kathryn Conger, and James Gung. 2018. The new propbank: Aligning propbank with amr through pos unification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

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

Alain Polguère. 2011. Propriétés sémantiques et combinatoires des quasi-prédicats sémantiques. *Scolia*, 26:131–152.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. Association for Computational Linguistics.

Carlos Ramisch and Aline Villavicencio. 2018. Computational treatment of multiword expressions. In Ruslan Mitkov, editor, *Oxford Handbook on Computational Linguistics*, 2nd edition. Oxford University Press.

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 (Volume 2: Short Papers)*, pages 499–505. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2276/2010 of *CICLing '02*, pages 1–15, London, UK. Springer-Verlag.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47. Association for Computational Linguistics.

Violeta Seretan. 2011. *Syntax-based collocation extraction*, volume 44 of *Text, Speech and Language Technology*. Springer Science & Business Media.

Violeta Seretan and Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 953–960, Sydney, Australia. Association for Computational Linguistics.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Peter Uhrig, Stefan Evert, and Thomas Proisl. 2018. Collocation Candidate Extraction from Dependency-Annotated Corpora: Exploring Differences across Parsers and Dependency Annotation Schemes. In *Lexical Collocation Analysis*, Quantitative Methods in the Humanities and Social Sciences, pages 111–140. Springer.

Leo Wanner. 1996. *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Corpus Linguistics*. John Benjamins Publishing.

Leo Wanner, Bernd Bohnet, Nadjet Bouayad-Agha, François Lareau, and Daniel Nicklaß. 2010. MARQUIS: Generation of user-tailored multilingual air quality bulletins. *Applied Artificial Intelligence*, 24(10):914–952.

Leo Wanner, Bernd Bohnet, and Mark Giereth. 2006. Making sense of collocations. *Computer Speech & Language*, 20(4):609–624.

Leo Wanner, Gabriela Ferraro, and Pol Moreno. 2016. Towards Distributional Semantics-Based Classification of Collocations for Collocation Dictionaries. *International Journal of Lexicography*, 30(2):167–186.

A Appendices

Lexical Function	Eng.	Pt.	Sp.	Total	Lexical Function	Eng.	Pt.	Sp.	Total
$Oper1$	151	181	106	438	$FinOper1$	–	1	2	3
$Magn$	89	81	69	239	$Centr$	–	–	3	3
Bon	58	49	14	121	$Caus1Manif$	–	3	–	3
$A\_NonStandard$	55	28	30	113	$AntiMagn\_temp$	2	–	1	3
$S\_NonStandard$	8	31	16	55	$SLoc$	1	1	–	2
$AntiMagn$	18	14	16	48	$SIncepPredPlus$	2	–	–	2
$CausOper1$	16	12	16	44	$S2$	–	2	–	2
$Sing$	8	20	11	39	$AntiVer$	–	2	–	2
$Mult$	4	9	19	32	$AntiMagn\_quant$	–	1	1	2
$AntiBon$	20	9	1	30	$S\_SingCaus1Manif$	–	1	–	1
Cap	2	13	13	28	$SOper1$	–	1	–	1
$V\_NonStandard$	4	11	6	21	$SOper$	–	–	1	1
$CausFunc$	9	5	5	19	$SManif$	–	1	–	1
$Oper2$	1	14	2	17	$SLiqu1Func$	–	–	1	1
$Magn\_quant$	9	2	6	17	$SIncepFunc$	–	–	1	1
$Magn\_temp$	8	3	3	14	$SFinOper1$	–	1	–	1
Ver	3	5	1	9	$SFinOper$	–	–	1	1
$Germ$	3	5	1	9	$SFinFunc$	–	1	–	1
$Magn + AntiBon$	4	1	3	8	$SContOper1$	–	1	–	1
$CausFunc1$	–	7	–	7	$SCausOper1$	–	–	1	1
$Real1$	–	6	–	6	$SCausFunc$	–	–	1	1
$Magn + Bon$	3	1	2	6	$Real$	1	–	–	1
$Liqu1Func$	1	2	2	5	Loc	–	–	1	1
$Gener$	1	2	2	5	$Culm$	–	–	1	1
$SReal$	–	1	3	4	$ContOper1$	–	1	–	1
Pos	1	2	1	4	$CausPredPlus$	–	1	–	1
$NonStandard\_Oper1$	–	3	1	4	$CausManif$	–	1	–	1
$LiquFunc$	–	–	3	3	$AntiPos$	–	1	–	1
$IncepOper1$	1	1	1	3	$A\_NonStd./Magn$	1	–	–	1
$Func$	–	2	1	3	$A1IncepPredPlus$	–	–	1	1

Table 3: Number of unique collocations per lexical function and language in the final corpus.