Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics

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

Universidade da Coruña, CITIC, Grupo LyS, Dpto de Letras,
Fac. de Filoloxía. 15071, A Coruña
E-mail: {marcos.garcia.gonzalez,marcos.garcias,margarita.alonso}@udc.gal

Abstract

This paper presents the method used to create a multilingual online dictionary of collocations of English, Portuguese, and Spanish. This resource is built automatically and contains three types of collocations: verb–object (e.g., "[to] issue [an] invoice"), adjective–noun ("deep shame"), and nominal compounds ("cigarette packet"). We take advantage of dependency parsing and statistical association measures to compile collocations of each language, and then we align them with their equivalents in the other languages by means of compositional methods which use cross-lingual models of distributional semantics. Collocations are extracted from large and assorted corpora, and the cross-lingual models are mapped using unsupervised approaches. For each collocation in a given language, the system shows different equivalents in the other languages, ranked by a confidence value. Besides the multilingual perspective, the resulting dictionary can also serve as a monolingual resource to retrieve the collocates of a given base, thus being a useful application to both native speakers and language learners. The dictionary will be published as an online tool, and all the resources generated in this research will be freely available.

Keywords: collocations; distributional semantics; dictionary; multilinguality

1. Introduction

One of the main characteristics of collocations is that the selection of one of its elements is unpredictable. In this regard, when learning English, one should know that a horse *gallops* but a dog *scampers*, even if both verbs convey basically the same meaning. In a multilingual scenario, this unpredictability is even more important, because a collocation equivalent in a target language is often non-congruent, i.e., it is not the direct translation of both lexical units of the source combination (Nesselhauf, 2003). For instance, while in English an *invoice* is *issued*, in Portuguese a *factura* ('invoice') is *emitida* (literally 'emitted'). Thus, mastering the use of collocations and other formulaic sequences presents advantages for processing and improves the production performance of language learners (Millar, 2010).

Dictionaries with collocational information are becoming more frequent, allowing both native speakers and language learners to produce idiomatic combinations in different

domains (Benson et al., 1986; Crowther et al., 2009; Bosque, 2006; Alonso-Ramos et al., 2010). However, multilingual resources of collocations and other multiword expressions, such as idioms, are scarce, but they are very useful to command such structures in other languages (Alonso-Ramos, 2015). In this respect, building multilingual dictionaries of collocations is a hard task which requires a huge effort from expert lexicographers in different languages (Orenha-Ottaiano, 2017).

Taking the above into account, this paper presents the steps to automatically create a multilingual dictionary of collocations of English, Portuguese, and Spanish. The dictionary includes three types of collocational patterns: (i) verb–object (*obj*) such as the "[to] issue [an] invoice"; (ii) adjective–noun (*amod*), e.g., "deep shame", and (iii) nominal compounds (*nmod*) such as "cigarette packet" (or "packet of cigarettes").

Broadly speaking, the method consists of the following steps: first, we compile large corpora in each of the three languages, and analyse them using natural language processing (NLP) tools to obtain morphosyntactic and syntactic information (Gamallo et al., 2018; Straka & Straková, 2017). Then, we apply different statistical association measures (AMs) to automatically select collocation candidates from the corpora (Evert et al., 2017; Garcia et al., 2019). After that, we use cross-lingual models of distributional semantics to apply compositional strategies able to identify equivalents of a given collocation in other languages (Garcia et al., 2017; Gamallo & Garcia, 2019). Instead of using parallel corpora, the cross-lingual models can be generated with monolingual resources, thus avoiding the need of obtaining large parallel texts for each language pair (Artetxe et al., 2018). The resulting dictionary provides, for each collocation in a source language, a set of equivalents in the target ones, ranked by a confidence value which represents the translation probabilities. The dictionary will be published as an online tool, and all the resources generated in this research will be freely available.

The rest of this paper is organized as follows. Section 2 briefly presents some previous work concerning different methods to extract collocations from corpora. Then, the approaches to both identify monolingual and multilingual collocations are introduced in Section 3, which also discusses some shortcomings and further lines of research. Finally, Section 4 summarizes the main properties of the online dictionary, while Section 5 contains the conclusions of our study.

2. Methods to extract collocations with a lexicographic aim

In order to create the lexicographic resources to release a multilingual dictionary, our work takes advantage of different NLP methods aimed at identifying monolingual collocations from corpora as well as at finding their equivalents in other languages.

The first approaches to extract collocations from corpora consisted of applying AMs to short sequences of ngrams (Smadja, 1993). Using similar approaches, other studies defined patterns of part-of-speech tags to identify specific constructions (Krenn & Evert,

2001), while the use of syntactic dependencies was evaluated in articles such as Lin (1999) or Seretan and Wehrli (2006). Besides classical association measures such as pointwise mutual information or t-score, several authors have proposed directional AMs to capture the asymmetry of collocations (Gries, 2013; Carlini et al., 2014).

With a view to comparing the performance of different AMs, studies such as Pecina (2010), Evert et al. (2017), or Garcia et al. (2019) performed different evaluations of various measures to extract collocations in several languages. The results, however, differ with respect to the collocation type as well as to the interpretation of collocations, which involves divergent annotations on each gold-standard data.

With regard to the multilingual identification, the first studies exploited parallel corpora to find bilingual translations of collocations and other multiword expressions (Smadja, 1992; Kupiec, 1993; Haruno et al., 1996). More recently, the use of syntactic analysis was also proposed to restrict the search to predefined patterns (Wu & Chang, 2003; Seretan & Wehrli, 2007).

Other studies tackled this problem using comparable and unrelated corpora in two languages, by performing word-to-word translations of each component of the collocations — and other similar constructions — (Grefenstette, 1999; Baldwin & Tanaka, 2004; Delpech et al., 2012). Similar approaches, which improve the word-to-word translation by taking advantage of distributional models were presented in Morin and Daille (2012) and Garcia (2018). Finally, recent articles investigate the use of contextualized compositional models as well as weighted additive vectors to improve the identification of equivalents of multiword expressions in different languages (Gamallo & Garcia, 2019; Garcia et al., 2019).

Concerning dictionaries with collocational information, the majority of the publications are monolingual resources mostly focused on language learners. In this respect, English is the most represented language among the three targets (Benson et al., 1986; Crowther et al., 2009; Rundell, 2011), but there also exist dictionaries for Spanish, oriented to both native speakers and language learners (Alonso-Ramos, 2004; Bosque, 2004, 2006). For Portuguese, the COMBINA-PT project has generated a database of different multiword expressions, including collocations (Mendes et al., 2006), while *Syntax Deep Explorer* provides an online tool to retrieve co-occurrence information from large corpora (Correia et al., 2016). Moreover, the work presented in Larens (2016) describes the creation of a collocational database of Brazilian Portuguese.

From a multilingual perspective, some dictionaries with collocational information have been published for various language pairs, such as English–Russian (Benson & Benson, 1993), German–French (Ilgenfritz et al., 1989), or Italian–German (Konecny & Autelli, 2014). Several articles and projects have also carried out research aimed at creating multilingual dictionaries of collocations (Grefenstette et al., 1996; Nerima et al., 2003; Konecny & Autelli, 2014; Alonso-Ramos, 2015; Garcia et al., 2017; Orenha-Ottaiano, 2017). Concerning the three languages of our study, Alegro et al. (2010) presents a

bilingual dictionary of adjectival collocations in English and Portuguese. However, to the best of our knowledge there is no freely available multilingual resource of collocations for English, Portuguese, and Spanish, so our research aims at contributing to this area with an online tool and free resources in the three target languages. It is worth mentioning, however, that online dictionaries such as *Linguee*[1] contain not only monolexical entries, but also some multiword expressions (including several collocations). In this regard, the main difference with respect to our work is that we extract the equivalents from comparable and unrelated corpora instead of parallel data.

3. Automatic extraction of collocations

This section presents the different steps to automatically generate equivalents of collocations in various languages. First, we explain the processes used to obtain collocation candidates in one language, and then we introduce the approach to obtain their equivalents in other languages. Finally, we briefly discuss some features and shortcomings of the proposed strategies.

3.1 Monolingual extraction

We understand collocations as phraseological combinations of two lexical units (LUs) which are directly linked by a syntactic relation (Hausmann, 1989; Mel'čuk, 1995). The internal structure of these expressions is not symmetrical, since while one of the LUs is freely selected due to its meaning (the base), the selection of the other component (the collocate) is restricted by the former (Mel'čuk, 1996). Thus, a base such as *shame* may select the collocates *deep* or *intense* (but not *strong* or *heavy*) in order to convey the meaning 'intense'.

Aimed at identifying the syntactic relation between two lexical units we employ dependency parsing, which establishes binary relations between the different words of a sentence (Tesnière, 1959; Kübler et al., 2009). To capture the collocability of two syntactically related words we use various association measures which assign numerical values that allow us to rank the *attraction* or *repulsion* of the word pairs (Evert, 2008).

With this in mind, our method to extract monolingual collocation candidates is as follows: First, we obtain large amounts of corpora for each language (in our case, English, Portuguese, and Spanish). So far we have been working with texts from different sources, such as the Wikipedia, the Europarl (Koehn, 2005), OpenSubtitles (Lison & Tiedemann, 2016), as well as text from other genres such as essays, literature, and web pages. These corpora are first processed using LinguaKit to identify sentence boundaries and to provide tokenization, lemmatization and PoS-tagging (Gamallo et al., 2018). Then, the corpora are enriched with syntactic information using UDPipe

[1] https://www.linguee.com/

models (Straka & Straková, 2017), which are based on *Universal Dependencies* annotation (Nivre, 2015).[2] It is important to note that the use of dependency parsing also allows us to identify long distance dependencies which are not captured in a short span of text.

Once we have the processed data for each language, we select as candidate collocations those pairs of lemmas that belong to the following dependency relations, structured as base–collocate tuples: *obj* (*invoice,issue*), *amod* (*shame,deep*), *nmod/compound* (*cigarette,packet*). We use lemmas instead of tokens (i.e., we represent the different inflected forms of a word by a single entry) to reduce the data sparseness.[3]

Over these candidates, we apply different association measures (e.g., *t-score*, *log-likelihood*, *Dice*) to rank each list of pairs. From the results of previous studies, we use different AMs for each dependency relation (Garcia et al., 2019). Moreover, and since most frequent candidates tend to be phraseological, these ranks are combined with frequency data to select the top$-n$ combinations (Krenn & Evert, 2001). At the end of this process we have, for each language, large sets of collocation candidates for the three mentioned patterns.

3.2 Bilingual equivalents

In order to obtain equivalents in various languages of a given collocation in a source we use compositional semantics strategies by means of cross-lingual distributional models.

3.2.1 Cross-lingual distributional models

Monolingual models of distributional semantics (also known as *word embeddings*) use contextual information to represent words as $n-$dimensional vectors, so words occurring in similar contexts tend to have similar vectors (Landauer & Dumais, 1997). Likewise, cross-lingual models represent the words of different languages in the same vector space, thus allowing for the computation of distributional similarities between those different languages (Rapp, 1999; Ruder et al., 2019).

To build our collocational database we have used two different approaches to obtain cross-lingual models of distributional semantics. On the one hand, we have used MultiVec (Bérard et al., 2016) to train bilingual models using parallel data from the Europarl and OpenSubtitles corpora (Koehn, 2005; Lison & Tiedemann, 2016). On the other hand, and taking into account that large amounts of parallel data from different domains are scarce, we have also trained monolingual models using *word2vec* (Mikolov

[2] https://universaldependencies.org/

[3] Note that, for instance, a single verb in several Romance languages (including Portuguese and Spanish) may have more than 50 different inflected forms.

et al., 2013), and then mapped into a shared vector space with *vecmap* (Artetxe et al., 2018). The latter approach obtains high-quality cross-lingual models by means of unrelated corpora, so it allows us to use a large variety of texts from different genres which in turn generate better word embeddings.

We train the distributional models converting the original tokens of each corpus into *lemma_PoSTag* entries. This strategy alleviates both the sparseness produced by morphological variation as well as the potential ambiguity of words with different morphosyntactic categories which have the same lemma (e.g., plane_NOUN, plane_VERB, plane_ADJ). Besides, using these linguistically-enriched models allows us to select only those base and collocate candidates which belong to a specific part-of-speech.

In sum, cross-lingual models of distributional semantics allow us to obtain distributionally similar words in a target language for a given input in a source language. For instance, if we search for the most similar nouns (in English) to *adversário* (in Portuguese), we may get words such *adversary*, *foe*, or *opponent*.

3.2.2 Compositional semantics methods

A collocation encodes semantic information from both the base and the collocate, so that they are semantically compositional expressions. Nevertheless, it is worth noting that a collocate may convey a particular meaning in each specific combination (Mel'čuk, 1995). For instance, different adjectives such as *heavy* and *strong* convey basically the same meaning in collocations such as "heavy rain" and "strong coffee", while the verb *[to] pay* has a different meaning in "pay attention" and "pay the bills". The bases, however, have a stable meaning across the different combinations in which they appear. With that in mind, the semantic properties of collocations should be taken into account when searching for equivalents in other languages.

The approach that we use to find multilingual equivalents has been evaluated in various languages and relations with high precision results (Garcia et al., 2017; Garcia, 2018). On the one hand, we rely on the previously extracted monolingual collocations to select candidates which have some degree of collocability (or are at least frequent combinations) in each language. On the other hand, we select as candidate translations those collocations with a high degree of similarity between the input and target bases. The procedure is as follows: given an input collocation in a source language (e.g., *lío tremendo*, 'huge mess' in Spanish), we select its base (*lío*) and retrieve the n most similar words with the same part-of-speech in the target language: e.g., "trouble", "mess", etc. in English (where $n = 5$ and the similarity is computed by their cosine distance). Then, we select those collocations in the target language with the candidate bases (e.g., "little trouble", "deep trouble", "huge mess", "fine mess", etc.). After that, we compute the similarity between the source collocate and the target ones (e.g., "tremendo" *versus* "little", "deep", "huge", and "fine"). If the cosine distances between

both the source and target bases and collocates are higher than a given threshold, they are selected as potential equivalents, and the average similarity between both components is set as the translation confidence value (e.g., "lío tremendo–huge mess": 0.72).

This strategy follows the base–collocate structure of collocations by selecting in the target language only candidates with very similar bases. Also, it allows us to identify not only word-to-word translations between the collocates, since we use distributional similarity to compute the distance between the different candidates (Morin & Daille, 2012). Finally, using previously extracted collocations (instead of artificially generating new instances) avoids the creation of unconventional combinations in the target languages.

3.3 Discussion

Even though the proposed approaches effectively obtain equivalents in various languages with high precision (about 90%, depending on the scenario), it is worth noting that the results and error analyses carried out in different studies have pointed to some issues that could be improved in further research (Garcia et al., 2017, 2018).

On the one hand, using statistical data (frequency and various association measures) to rank the monolingual collocation candidates may result in non-phraseological expressions such as free combinations (e.g., "buy [a] beer") or quasi-idioms ("big deal") (Mel'čuk, 1995). In this regard, we do not consider this circumstance a serious problem as long as the equivalents in the other languages (if any) are valid. However, and with a view to refine the monolingual identification, several strategies can be implemented to improve the ranking of the candidates and to automatically identify non-compositional expressions (Pecina, 2010; Cordeiro et al., 2019).

On the other hand, and even if distributional semantics models are able to identify some non-congruent bilingual equivalents, several collocates convey a very different meaning (with respect to their most frequent one) in some specific collocations. In these cases, finding appropriate candidates without using parallel corpora may be a difficult task: for instance, both the Portuguese and Spanish translations of the English verb "[to] pay" will probably belong to the economic field (*pagar* '[to] pay', *cobrar* '[to] earn', etc.), so our approach may not identify *prestar atenção/atención* (literally '[to] lend attention') as equivalents of "[to] pay attention". There is, however, recent research which could improve the extraction of these cases: as mentioned in Section 2, Garcia et al. (2019) propose a compositional strategy to find bilingual collocation equivalents using weighted additive vectors. Besides, in Gamallo and Garcia (2019) the authors use contextualized word embeddings based on syntactic dependencies to represent the meaning of composite expressions. In this regard, combining both approaches could be an interesting line of research for further work.

Finally, the performance of our current approaches is also influenced by one of the main shortcomings of standard distributional methods, which represent in the same vector different senses of the same word. To overcome this issue (apart from the mentioned strategy of Gamallo and Garcia (2019)), studies such as Iacobacci et al. (2015) have implemented sense-based distributional models, while recent research in NLP obtains pre-trained contextual representations, where the vector of a given word is based on the other words which occur in the same sentence (Weir et al., 2016; Devlin et al., 2019).

4. Towards a multilingual dictionary of collocations

This section illustrates how we leverage the multilingual resources generated by the methods presented above to create an online tool with monolingual and multilingual collocational information. This tool is not a finished multilingual dictionary of collocations, but rather an instrument to help language users by exploiting our database. In this regard, it is worth remembering that this database is automatically constructed and freely available, and that it can be updated both with new information obtained from corpora as well as with manual annotation from lexicographers.

The query interface is based on a *source–target* structure, so that the user should first select the desired translation direction (e.g., English→Portuguese, Portuguese→Spanish, etc.). As in other resources, the basic units of the dictionary are nouns (Lea & Runcie, 2002). In our case, however, the selection of nouns as the main unit derives from the fact that they are the bases of the three considered patterns. Nevertheless, the same strategy can be applied to other collocational patterns such as verb–adverb (e.g., "really want", where the verb is the base), or adjective–adverb (e.g., "extremely powerful", in which the adjective is the base), among others.

Thus, after selecting a source and target language, the user introduces a noun (by its lemma) in the search box (e.g., "wine" in English→Portuguese). As the input query is performed, the dictionary will show, in three columns, the highest ranked combinations in the source language with the given base. In the previous example, the verb–object column may include "drink wine", "produce wine", or "export wine"; adjective–noun collocates such as "red wine", "white wine", or "varietal wine", and "bottle [of] wine", "glass [of] wine", or "wine grape" as nominal compounds. The user can expand one specific column to search for other collocations in the desired pattern. Besides, the tool allows for clicking in a particular collocation to see a few usage examples extracted from corpora. At this point, the dictionary can be also seen as a database of collocations in a specific language.

Continuing with the multilingual tool, the user can select a collocation in the source language to search for equivalents in the target one (e.g., adjective–noun: "red wine"). Then, the dictionary will show the collocation equivalents in the target language, sorted by the confidence value obtained using the compositional strategies presented in Section

3.2. Following the previous example, the Portuguese equivalents (and their confidence values) of the English collocation "red wine" may be *vinho tinto*: 0.95 ('red wine'), *vinho rosé*: 0.86 ('rosé wine'), or *vermute tinto*: 0.83 ('red vermouth'), among others. Again, the tool allows the user to expand the number of bilingual equivalents as well as to see real examples in corpora, this time in the target language.

It is worth mentioning that the database is automatically enlarged with new entries built by transitivity. Therefore, in those cases where it has a specific collocation translated in two language directions it infers the third one if it has not been extracted. As an example, let us say that we obtained the English→Portuguese and English→Spanish equivalents of "red vermouth" (*vermute tinto* and *vermú rojo*, respectively), but not the Portuguese→Spanish translation. Thus, the tool will infer that *vermú rojo* may be a suitable translation of *vermute tinto*. In these cases, the inferred equivalents are presented using a slightly different colour to inform the user of this fact.

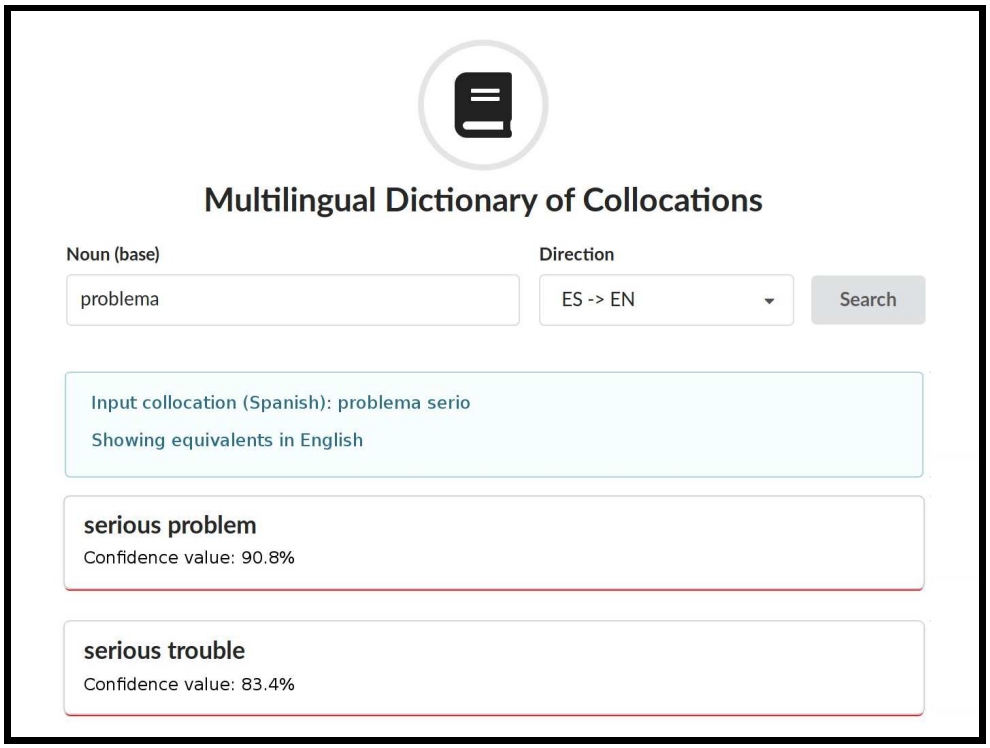

Figure 1: Example of the online interface of the dictionary.

Figure 1 shows an example of the online interface. The inserted noun (top row) is *problema* ('problem') and the translation direction Spanish→English. The figure includes the second visualization of the tool, after selecting the input collocation *problema serio* (adjective– noun). It displays the top two translations together with their confidence values, and allows the user to click on any of them to see real examples.

The current version of the online tool (together with the multilingual database) presents two issues that can be addressed in future research. First, as our approach relies on

lemmatized instances of syntactic dependencies, we do not pay particular attention to the surface structures allowed by each collocation. Thus, the dictionary provides the users with base–collocate data, but it does not explicitly inform, for instance, whether a noun requires a determiner or not (e.g., *"take the advantage" *versus* "have a look"). The second peculiarity concerns the order in which both LUs are shown to the users. In each pattern, collocations are presented in their canonical structure in the three languages (e.g., adjective–noun pairs are shown as noun–adjective in Portuguese and Spanish), but while some of them are mostly used only in a particular pattern (e.g., "football manager" *versus* *"manager of football"), others may appear in both ways (e.g., "energy consumption – consumption of energy").[4] Both problems are partially addressed with the usage examples of each collocation, but further work could also focus on these issues in order to improve the representation of each combination.

5. Conclusions and further work

This paper has presented a set of methods to automatically create a database of collocation equivalents in English, Portuguese, and Spanish. This database is used to supply an online dictionary which aids language users in the selection of both monolingual and multilingual combinations of a given noun.

To extract candidate collocations we employ dependency parsing and statistical association measures applied to large monolingual corpora. We have focused on the following three collocational patterns: verb–object, adjective–noun, and nominal compounds. To identify bilingual equivalents of a given collocation in a source language, we use compositional distributional methods which rely on pre-extracted collocations in the target languages. The cross-lingual distributional models can be directly learned using parallel corpora, or mapped after monolingual training with unrelated resources.

Apart from presenting the different strategies to build the collocation database, this study also discusses some shortcomings of the proposed approaches, aimed at improving both the monolingual extraction and the multilingual alignment in further work.

Finally, the paper presents the main structure and functionalities of the online tool, which can be useful for language users in a monolingual scenario (searching for collocates in a particular language) and in a multilingual one (to find equivalents in other languages). It is worth noting that all the resources created in this research will be freely available.

[4] A different issue occurs in some constructions which may have a different meaning with respect to their structure, such as "coffee cup" and "cup of coffee".

6. Acknowledgements

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). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

7. References

Alegro, A., Mobaid, R. & Brezolin, A. (2010). *Happy Couples. Dicionário de Colocações Lexicais Adjetivas*. Disal.

Alonso-Ramos, M. (2004). DiCE: Diccionario de Colocaciones del Español. Universidade da Coruña. http://dicesp.com.

Alonso-Ramos, M. (2015). Discovering hidden collocations in a bilingual Spanish–English dictionary. In I. Kosem, M. Jakubiček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 170–185.

Alonso-Ramos, M., Nishikawa, A. & Vincze, O. (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. Presses univ. de Louvain, pp. 369–374.

Artetxe, M., Labaka, G. & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 789–798.

Baldwin, T. & Tanaka, T. (2004). Translation by Machine of Complex Nominals: Getting it Right. In *Second ACL Workshop on Multiword Expressions: Integrating Processing*. Association for Computational Linguistics, pp. 24–31.

Benson, M. & Benson, E. (1993). *Russian-English dictionary of verbal collocations*. John Benjamins Publishing.

Benson, M., Benson, E. & Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. John Benjamins Publishing.

Bosque, I. (2004). *Redes. Diccionario combinatorio del español contemporáneo*. SM.

Bosque, I. (2006). *Diccionario combinatorio práctico del español contemporáneo. Las palabras en su contexto*. SM.

Bérard, A., Servan, C., Pietquin, O. & Besacier, L. (2016). MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In N.C.C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo,

A. Moreno, J. Odijk & S. Piperidis (eds.) *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*. Paris, France: European Language Resources Association, pp. 4188–4192.

Carlini, R., Codina-Filba, J. & Wanner, L. (2014). Improving collocation correction by ranking suggestions using linguistic knowledge. In *Proceedings of the third workshop on NLP for computer-assisted language learning*. Uppsala: LiU Electronic Press, pp. 1–12.

Cordeiro, S., Villavicencio, A., Idiart, M. & Ramisch, C. (2019). Unsupervised Compositionality Prediction of Nominal Compounds. *American Journal of Computational Linguistics*, 45(1), pp. 1–57.

Correia, J., Baptista, J. & Mamede, N. (2016). Syntax Deep Explorer. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami & A. Branco (eds.) *Computational Processing of the Portuguese Language*, volume 9727 of *Lecture Notes in Computer Science*. Springer, pp. 189–201.

Crowther, J., Dignen, S. & Lea, D. (eds.) (2009). *Oxford Collocations Dictionary for student's of English*. Oxford University Press.

Delpech, E., Daille, B., Morin, E. & Lemaire, C. (2012). Extraction of Domain-Specific Bilingual Lexicon from Comparable Corpora: Compositional Translation and Ranking. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, pp. 745–762.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics. An international handbook*, volume 2. Berlin: Mouton de Gruyter, pp. 1212– 1248.

Evert, S., Uhrig, P., Bartsch, S. & Proisl, T. (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*. pp. 531–549.

Gamallo, P. & Garcia, M. (2019). Unsupervised Compositional Translation of Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019),* at the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics, pp. 40–48.

Gamallo, P., Garcia, M., Pineiro, C., Martinez-Castaño, R. & Pichel, J. C. (2018). LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, pp. 239–244.

Garcia, M. (2018). Comparing bilingual word embeddings to translation dictionaries

for extracting multilingual collocation equivalents. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (eds.) *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, volume 3 of *Phraseology and Multiword Expressions*, chapter 12. Language Science Press, pp. 319–342.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2017). Using bilingual word-embeddings for multilingual collocation extraction. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, pp. 21–30.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2018). Discovering bilingual collocations in parallel corpora: A first attempt at using distributional semantics. In I. Doval & M.T. Sánchez Nieto (eds.) *Parallel corpora for contrastive and translation studies: New resources and applications*, volume 90 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company, pp. 267–279.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019),* at the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics, pp. 49–59.

Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, volume 21.

Grefenstette, G., Heid, U., Schulze, B., Fontenelle, T. & Gerardy, C. (1996). The DECIDE project: Multilingual collocation extraction. In *Euralex 96 Proceedings*. Göteborg, pp. 93–108.

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

Haruno, M., Ikehara, S. & Yamazaki, T. (1996). Learning bilingual collocations by word level sorting. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 1 of *COLING 1996*. Association for Computational Linguistics, pp. 525–530.

Hausmann, F. J. (1989). Le dictionnaire de collocations. *Wörterbücher, Dictionaries, Dictionnaires*, 1, pp. 1010–1019.

Iacobacci, I., Pilehvar, M. T. & Navigli, R. (2015). SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 95–105.

Ilgenfritz, P., Schneider, G. & Stephan-Gabinel, N. (1989). *Langenscheidts Kontextwörterbuch Französisch-Deutsch*. Langenscheidt.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT, AAMT, pp. 79–86.

Konecny, C. & Autelli, E. (2014). *Kollokationen Italienisch - Deutsch*. Helmut Buske.

Krenn, B. & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*. Association for Computational Linguistics, pp. 39–46.

Kupiec, J. (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. Association for Computational Linguistics, pp. 17–22.

Kübler, S., McDonald, R. & Nivre, J. (2009). *Dependency Parsing*. Morgan and Claypool Publishers.

Landauer, T. & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), pp. 211–240.

Larens, J. (2016). Colocações *do Português Brasileiro: Tipologia, Categorização, e Construção de uma Base de Dados*. Ph.D. thesis, Universidade Federal do Ceará.

Lea, D. & Runcie, M. (2002). Blunt Instruments and Fine Distinctions: a Collocations dictionary for students of English. In *Proceedings of the Tenth EURALEX International Congress. Copenhagen: Center for Sprogteknologi*, volume 2. pp. 819–829.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL 1999. Association for Computational Linguistics, pp. 317–324.

Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In N.C.C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association, pp. 923–929.

Mel'čuk, I. (1995). Phrasemes in language and phraseology in linguistics. In M. Everaert, E. J. van der Linden, A. Schenk & R. Schreu (eds.) *Idioms: Structural and psychological perspectives*. Hillsdale: Lawrence Erlbaum Associates, pp. 167–232.

Mel'čuk, I. (1996). Lexical Functions: a Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Corpus Linguistics*. John Benjamins Publishing, pp. 37–102.

Mendes, A., Antunes, S., Nascimento, M.F.B.d., Casteleiro, J. M., Pereira, L. & Sá, T. (2006). COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. European Language Resources Association, pp. 1900–1905.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (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.

Millar, N. (2010). The processing of malformed formulaic language. *Applied Linguistics*, 32(2), pp. 129–148.

Morin, E. & Daille, B. (2012). Revising the Compositional Method for Terminology Acquisition from Comparable Corpora. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, pp. 1797–1810.

Nerima, L., Seretan, V. & Wehrli, E. (2003). Creating a multilingual collocations dictionary from large text corpora. In *10th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 131–134.

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied linguistics*, 24(2), pp. 223–242.

Nivre, J. (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*. Springer, pp. 3–16.

Orenha-Ottaiano, A. (2017). The compilation of an online Corpus-Based bilingual Collocations Dictionary: motivations, obstacles and achievements. In *Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch*. Lexical Computing, pp. 458–473.

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

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, pp. 519–526.

Ruder, S., Vulić, I. & Søgaard, A. (2019). A Survey of Cross-Lingual Word Embedding Models. *Journal of Artificial Intelligence Research*. arXiv preprint arXiv:1706.04902.

Rundell, M. (ed.) (2011). *Macmillan Collocations Dictionary*. Macmillan.

Seretan, V. & Wehrli, E. (2006). Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. pp. 953–960.

Seretan, V. & Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conference sur le traitement automatique des langues naturelles*, TALN 2007. IRIT Press, pp. 401–410.

Smadja, F. (1992). How to compile a bilingual collocational lexicon automatically. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*. AAAI Press, pp. 57–63.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational*

Linguistics, 19(1), pp. 143–177.

Straka, M. & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, pp. 88–99.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Librairie C. Klincksieck.

Weir, D., Weeds, J., Reffin, J. & Kober, T. (2016). Aligning Packed Dependency Trees: A Theory of Composition for Distributional Semantics. *Computational Linguistics*, 42(4), pp. 727–761.

Wu, C. C. & Chang, J. S. (2003). Bilingual collocation extraction based on syntactic and statistical analyses. In *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing*. Association for Computational Linguistics and Chinese Language Processing, pp. 1–20.