Semantics-Driven Recognition of Collocations Using Word Embeddings

Sara Rodríguez-Fernández[1], **Luis Espinosa-Anke**[1], **Roberto Carlini**[1], and **Leo Wanner**[1,2]
[1]NLP Group, Department of Information and Communication Technologies, Pompeu Fabra University
C/ Roc Boronat, 138, 08018 Barcelona (Spain)
[2]Catalan Institute for Research and Advanced Studies (ICREA)
sara.rodriguez.fernandez|luis.espinosa|roberto.carlini|leo.wanner@upf.edu

Abstract

L2 learners often produce "ungrammatical" word combinations such as, e.g., **give a suggestion* or **make a walk*. This is because of the "collocationality" of one of their items (the *base*) that limits the acceptance of *collocates* to express a specific meaning ('perform' above). We propose an algorithm that delivers, for a given base and the intended meaning of a collocate, the actual collocate lexeme(s) (*make / take* above). The algorithm exploits the linear mapping between bases and collocates from examples and generates a collocation transformation matrix which is then applied to novel unseen cases. The evaluation shows a promising line of research in collocation discovery.

1 Introduction

Collocations of the kind *make* [*a*] *suggestion*, *attend* [*a*] *lecture*, *heavy rain*, *deep thought*, *strong tea*, etc., are restricted lexical co-occurrences of two syntactically bound lexical elements (Kilgarriff, 2006). The central role of collocations for second language (henceforth, L2) learning has been discussed in a series of theoretical and empirical studies (Hausmann, 1984; Bahns and Eldaw, 1993; Granger, 1998; Lewis and Conzett, 2000; Nesselhauf, 2005; Alonso Ramos et al., 2010) and is widely reflected in (especially English) learner dictionaries. In computational lexicography, several statistical measures have been used to retrieve collocations from corpora, among them, *mutual information* (Church and Hanks, 1989; Lin, 1999), *entropy* (Kilgarriff, 2006), *pointwise mutual information* (Bouma, 2010), and *weighted pointwise mutual information* (Carlini et al., 2014).[1] However, the needs of language learners go beyond mere lists of collocations: the cited studies reveal that language learners often build "miscollocations" (as, e.g., **give a suggestion* or **have the curiosity*) to express the intended meaning. In other words, they fail to observe, in Kilgarriff's terms, the "collocationality" restrictions of L2, which imply that in language production, one of the elements of a collocation (the *base*) is freely chosen, while the choice of the other (the *collocate*) depends on the base (Hausmann, 1989; Cowie, 1994). For instance, to express the *meaning* of 'do' or 'perform', the base *suggestion* prompts for the choice of *make* as collocate: *make* [*a*] *suggestion*, while *advice* prompts for *give*: *give* [*an*] *advice*; to express the meaning of 'participate in', *lecture* prompts for *attend*: *attend* [*a*] *lecture*, while *operation* prompts for *assist*: *assist* [*an*] *operation*; to express the meaning of 'intense' in connection with *rain*, the right collocate is *heavy*, while 'intense wind' is *strong wind*. And so on. The idiosyncrasy of collocations makes them also language-specific. Thus, in English, you *take* [*a*] *walk*, in Spanish you 'give' it (*dar* [*un*] *paseo*), and in German and French you 'make' it ([*einen*] *Spaziergang machen*, *faire* [*une*] *promenade*); in English, *rain* is *heavy*, while in Spanish and German it is 'strong' (*fuerte lluvia/starker Regen*).

In order to effectively support L2 learners, techniques are thus needed that are able not only to retrieve collocations, but also provide for a given base (or headword) and a given semantic gloss of a collocate meaning, the actual collocate lexeme. In what follows, we present such a technique, which is grounded in Mikolov et al. (2013c)'s word embeddings, and which leverages the fact that semantically related words in two different

[1] See (Pecina, 2008) for a detailed survey of such measures.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 499–505,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

vector representations are related by linear transformation (Mikolov et al., 2013b). This property has been exploited for word-based translation Mikolov et al. (2013b), learning semantic hierarchies (hyponym-hypernym relations) in Chinese (Fu et al., 2014), and modeling linguistic similarities between standard (Wikipedia) and non-standard language (Twitter) (Tan et al., 2015). In our task, we learn a *transition matrix* over a small number of collocation examples, where collocates share the same semantic gloss, to apply then this matrix to discover new collocates for any previously unseen collocation base. We discuss the outcome of the experiments with ten different collocate glosses (including 'do' / 'perform', 'increase', 'decrease', etc.), and show that for most glosses, an approach that combines a stage of the application of a gloss-specific transition matrix with a pruning stage that is based on statistical evidence outperforms approaches that exploit only one of these stages as well as a baseline that is based on collocation retrieval exploiting the embeddings property for drawing analogies, such as, e.g., $x \sim$ *applause* $\equiv$ *heavy* $\sim$ *rain* (implying x=*thunderous*) (Rodríguez-Fernández et al., 2016).

2 Theoretical model

The semantic glosses of collocates across collocations can be generalized into a generic semantic typology modeled, e.g., by Mel'čuk (1996)'s *Lexical Functions*. For instance, *absolute*, *deep*, *strong*, *heavy* in *absolute certainty*, *deep thought*, *strong wind*, and *heavy storm* can all be glossed as 'intense'; *make*, *take*, *give*, *carry out* in *make* [*a*] *proposal*, *take* [*a*] *step*, *give* [*a*] *hint*, *carry out* [*an*] *operation* can be glossed as 'do'/'perform'; etc. Our goal is to capture the relation that holds between the training bases and the collocates with the same gloss, such that given a new base and a gloss, we can retrieve its corresponding collocate(s) with this gloss. Thus, given *absolute certainty*, *deep thought*, and *strong wind* as training examples, *storm* as input base and 'intense' as gloss, we aim at retrieving the collocate *heavy*. As already mentioned above, our approach is based on Mikolov et al. (2013b)'s linear transformation model, which associates word vector representations between two analogous spaces. In Mikolov et al.'s original work, one space captures words in language L_1 and the other space words in language L_2, such that the found relations are between translation equivalents. In our case, we define a base space $\mathcal{B}$ and a collocate space $\mathcal{C}$ in order to relate bases with their collocates that have the same meaning, in the same language. To obtain the word vector representations in $\mathcal{B}$ and $\mathcal{C}$, we use Mikolov et al. (2013c)'s *word2vec*.[2]

The linear transformation model is constructed as follows. Let $\mathbf{T}$ be a set of collocations whose collocates share the semantic gloss τ, and let b_{t_i} and c_{t_i} be the collocate respectively base of the collocation $t_i \in \mathbf{T}$. The base matrix $B_\tau = [b_{t_1}, b_{t_2} \dots b_{t_n}]$ and the collocate matrix $C_\tau = [c_{t_1}, c_{t_2} \dots c_{t_n}]$ are given by their corresponding vector representations. Together, they constitute a set of training examples Φ_τ, composed by vector pairs $\{b_{t_i}, c_{t_i}\}_{i=1}^{n}$.

Φ_τ is used to learn a linear transformation matrix $\Psi_\tau \in \mathbb{R}^{\mathcal{B} \times \mathcal{C}}$. Following the notation in (Tan et al., 2015), this transformation can be depicted as:

$$B_\tau \Psi_\tau = C_\tau$$

We follow Mikolov et al.'s original approach and compute Ψ_τ as follows:

$$\min_{\Psi_\tau} \sum_{i=1}^{|\Phi_\tau|} \|\Psi_\tau b_{t_i} - c_{t_i}\|^2$$

Hence, for any given novel base b_{j_τ}, we obtain a novel list of ranked collocates by applying $\Psi_\tau b_{j_\tau}$ and filtering the resulting candidates by part of speech and $NPMI$, an association measure that is based on the pointwise mutual information, but takes into account the asymmetry of the lexical dependencies between a base and its collocate (Carlini et al., 2014):

$$NPMI = \frac{PMI(collocate, base)}{-log(p(collocate))}$$

3 Experiments

3.1 Setup of the Experiments

We carried out experiments with 10 of the most frequent semantic collocate glosses (listed in the first column of Table 1). As is common in previous work on semantic collocation classification (Moreno et al., 2013; Wanner et al., 2016), our training set consists of a list of manually annotated correct collocations. For this purpose, we

[2] https://code.google.com/archive/p/word2vec/

Semantic gloss	Example	# instances
‘intense’	*absolute certainty*	586
‘weak’	*remote chance*	70
‘perform’	*give chase*	393
‘begin to perform’	*take up a chase*	79
‘stop performing’	*abandon a chase*	12
‘increase’	*improve concentration*	73
‘decrease’	*limit* [*a*] *choice*	73
‘create’, ‘cause’	*pose* [*a*] *challenge*	195
‘put an end’	*break* [*the*] *calm*	79
‘show’	*exhibit* [*a*] *characteristic*	49

Table 1: Semantic glosses and size of training set

randomly selected nouns from the Macmillan Dictionary and manually classified their corresponding collocates with respect to the glosses.[3] Note that there may be more than one collocate for each base. Since collocations with different collocate meanings are not evenly distributed in language (e.g., speakers use more often collocations conveying the idea of ‘intense’ and ‘perform’ than ‘stop performing’), the number of instances per gloss in our training data also varies significantly (see Table 1).

Due to the asymmetric nature of collocations, not all corpora may be equally suitable for the derivation of word embedding representations for both bases and collocates. Thus, we may hypothesize that for modeling (nominal) bases, which keep in collocations their literal meaning, a standard register corpus with a small percentage of figurative meanings will be more adequate, while for modeling collocates, a corpus which is potentially rich in collocations is likely to be more appropriate. In order to verify this hypothesis, we carried out two different experiments. In the first experiment, we used for both bases and collocates vectors pre-trained on the Google News corpus (*GoogleVecs*), which is available at *word2vec*'s website. In the second experiment, the bases were modeled by training their word vectors over a 2014 dump of the English Wikipedia, while for modeling collocates, again, *GoogleVecs* has been used. In other words, we assumed that Wikipedia is a standard register corpus and thus better for modeling $\mathcal{B}$, while *GoogleVecs* is more suitable for modeling $\mathcal{C}$. The figures in Section 3.2 below will give us a hint whether this assumption is correct.

For the calculation of $NPMI$ during post-processing, the British National Corpus (BNC) was used.[4]

3.2 Evaluation

The outcome of each experiment was assessed by verifying the correctness of each retrieved candidate from the top-10 candidates obtained for each test base. A total of 10 bases was evaluated for each gloss. The ground truth test set was created in a similar fashion as the training set: nouns from the Macmillan Dictionary were randomly chosen, and their collocates manually classified in terms of the different glosses, until a set of ten unseen base–collocate pairs was obtained for each gloss.

For the outcome of each experiment, we computed both *precision* (p) as the ratio of retrieved collocates that match the targeted glosses to the overall number of obtained collocates for each base, and *Mean Reciprocal Rank* (MRR), which rewards the position of the first correct result in a ranked list of outcomes:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

where Q is a sample of experiment runs and $rank_i$ refers to the rank position of the *first* relevant outcome for the *i*th run. MRR is commonly used in Information Retrieval and Question Answering, but has also shown to be well suited for collocation discovery; see, e.g., (Wu et al., 2010).

We evaluated four different configurations of our technique against two baselines. The first baseline (**S1**) is based on the regularities in word embeddings, with the $vec(\text{king}) - vec(\text{man}) + vec(\text{woman}) = vec(\text{queen})$ example as paramount case. In this context, we manually selected one representative example for each semantic gloss to discover collocates for novel bases following the same schema; cf., e.g., for the gloss ‘perform’ $vec(\text{take}) - vec(\text{walk}) + vec(\text{suggestion}) = vec(\text{make})$ (where *make* is the collocate to be discovered); see (Rodríguez-Fernández et al., 2016) for details. The second baseline (**S2**) is an extension of S1 in that its output

[3] At this stage of our work, we considered only collocations that involve single word tokens for both the base and the collocate. In other words, we did not take into account, e.g., phrasal verb collocates such as *stand up*, *give up* or *calm down*. We also left aside the problem of subcategorization in collocations; cf., e.g., *into* in *take* [*into*] *consideration*.

[4] As one of the reviewers pointed out, BNC might not be optimal as a collocation reference corpus. On the one hand, it does not capture collocations that might be idiosyncratic to American English, and, on the other hand, it might be outdated (and thus not contain more recent collocations). It is subject of future work to verify whether another representative corpus of English serves better.

	Precision (p)						Mean Reciprocal Rank (MRR)					
Semantic gloss	S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6
‘intense’	0.08	0.43	0.04	0.50	0.24	**0.72**	0.18	0.35	0.35	0.15	0.66	**0.82**
‘weak’	0.09	0.11	0.23	**0.45**	0.27	0.39	0.31	0.15	**0.69**	0.64	0.61	0.47
‘perform’	0.05	0.17	0.01	0.06	0.13	**0.40**	0.22	0.32	0.01	0.35	0.70	**0.79**
‘begin to perform’	0.03	0.08	0.24	0.30	0.22	**0.38**	0.17	0.05	0.61	0.64	0.70	**0.71**
‘stop performing’	0.00	0.00	0.11	0.15	0.12	**0.20**	0.01	0.00	**0.90**	0.66	0.71	0.65
‘increase’	0.16	**0.53**	0.31	0.43	0.35	**0.53**	0.47	0.72	0.78	0.86	0.86	**0.90**
‘decrease’	0.07	0.05	**0.28**	0.25	0.27	**0.28**	0.18	0.04	**0.57**	0.38	0.37	0.30
‘create’, ‘cause’	0.10	0.16	0.01	0.15	0.14	**0.53**	0.41	0.23	0.11	0.11	0.48	**0.58**
‘put an end’	0.05	0.09	0.15	0.20	0.08	**0.25**	0.28	0.10	**0.38**	0.36	0.33	**0.38**
‘show’	0.10	0.55	0.24	0.49	0.49	**0.70**	0.44	0.54	**0.87**	0.82	0.73	0.81

Table 2: Precision and MRR

Semantic gloss	Base	Retrieved candidates
‘intense’	*caution*	*extreme*
‘weak’	*change*	*slight, little, modest, minor, noticeable, minimal, sharp, definite, small, big*
‘perform’	*calculation*	*produce, carry*
‘begin to perform’	*cold*	*catch, get, run, keep*
‘stop performing’	*career*	*abandon, destroy, ruin, terminate, threaten, interrupt*
‘increase’	*capability*	*enhance, increase, strengthen, maintain, extend, develop, upgrade, build, provide*
‘decrease’	*congestion*	*reduce, relieve, cut, ease, combat*
‘create’, ‘cause’	*challenge*	*pose*
‘put an end’	*ceasefire*	*break*
‘show’	*complexity*	*demonstrate, reveal, illustrate, indicate, reflect, highlight, recognize, explain*

Table 3: Examples of retrieved collocations

is filtered with respect to the valid POS-patterns of targeted collocations and $NPMI$.[5]

The four configurations of our technique that we tested were: **S3**, which is based on the transition matrix for which *GoogleVecs* is used as reference vector space representation for both bases and collocates; **S4**, which applies POS-pattern and $NPMI$ filters to the output of S3; **S5**, which is equivalent to S3, but relies on a vector space representation derived from Wikipedia for learning bases projections and on a vector space representation from *GoogleVecs* for collocate projections; and, finally, **S6**, where the S5 output is, again, filtered by POS collocation patterns and $NPMI$.

4 Discussion

The results of the experiments are displayed in Table 2. In general, the configurations S3 – S6 largely outperform the baselines, with the exception of the gloss ‘increase’, for which S2 equals S6 as far as p is concerned. However, in this case too MRR is considerably higher for S6, which achieves the highest MMR scores for 6 and the highest precision scores for 7 out of 10 glosses (see the S6 columns in Table 2). In other words, the full pipeline promotes good collocate candidates to the first positions of the ranked result lists and is also best in terms of accuracy.

Comparing S1, S3, S5 to S2, S4, and S6 , we may conclude that the inclusion of a filtering module (and, in particular, of an $NPMI$ filtering module) contributes substantially to the overall precision in nearly all cases (‘decrease’ being the only exception). The comparison of the precision obtained for configurations S3 and S5 also reveals that for 7 glosses the strategy to model $\mathcal{C}$ and $\mathcal{B}$ on different corpora paid off. This is different as far as MRR is concerned. Further investigation is needed for the examination of this discrepancy.

We can observe that certain glosses seem to exhibit less linguistic variation, requiring a less populated transformation function from bases to collocates. Consider the case of ‘show’, which generates with only 49 training pairs the second best transition matrix, with p=0.70. It is also informative to contrast the performance on pairs of glosses with opposite meanings, such as e.g., ‘begin to perform’ vs. ‘stop performing’; ‘increase’ vs. ‘decrease’; ‘intense’ vs. ‘weak’; and finally ‘create, cause’ vs. ‘put an end’. Better performance is achieved consistently on the *positive* counterparts (e.g. ‘begin to perform’ over ‘stop performing’). A closer look at the output reveals that in these

[5] At the first glan ce, a state-of-the-art approach on correction of collocation errors by suggesting alternative co-occurrences, such, as, e.g., (Dahlmeier and Ng, 2011; Park et al., 2008; Futagi et al., 2008), might appear as a suitable baseline. We discarded this option given that none of them uses explicit fine-grained semantics.

Semantic gloss	S6
'intense'	0.82
'weak'	0.45
'perform'	0.40
'begin to perform'	0.42
'stop performing'	0.22
'increase'	0.55
'decrease'	0.37
'create', 'cause'	0.59
'put an end'	0.43
'show'	0.85

Table 4: Precision of the coarse-grained evaluation of the S6 configuration

cases positive glosses are persistently classified as negative. Further research is needed to first understand why this is the case and then to come up with an improvement of the technique in particular on the *negative* glosses.

The fact that for some of the glosses precision is rather low may be taken as a hint that the proposed technique is not suitable for the task of semantics-oriented recognition of collocations. However, it should be also stressed that our evaluation was very strict: a retrieved collocate candidate was considered as correct only if it formed a collocation with the base, and if it belonged to the target semantic gloss. In particular the first condition might be too rigorous, given that, in some cases, there is a margin of doubt whether a combination is a free co-occurrence or a collocation; cf., e.g., *huge challenge* or *reflect* [*a*] *concern*, which were rejected as collocations in our evaluation. Since for L2 learners such co-occurrences may be also useful, we carried out a second evaluation in which all the suggested collocate candidates that belonged to a target semantic gloss were considered as correct, even if they did not form a collocation.[6] Cf. Table 4 for the outcome of this evaluation for the S6 configuration. Only for 'perform' the precision remained the same as before. This is because collocates assigned to this gloss are support verbs (and thus void of own lexical semantic content).

5 Conclusions

As already pointed out in Section 1, a substantial amount of work has been carried out to automatically retrieve collocations from corpora (Choueka, 1988; Church and Hanks, 1989; Smadja, 1993; Lin, 1999; Kilgarriff, 2006; Evert, 2007; Pecina, 2008; Bouma, 2010; Futagi et al., 2008; Gao, 2013). Most of this work is based on statistical measures that indicate how likely the elements of a possible collocation are to co-occur, while ignoring the semantics of the collocations. Semantic classification of collocations has been addressed, for instance, in (Wanner et al., 2006; Gelbukh and Kolesnikova., 2012; Moreno et al., 2013; Wanner et al., 2016). However, to the best of our knowledge, our work is the first to automatically retrieve and typify collocations simultaneously. We have illustrated our approach with 10 semantic collocation glosses. We believe that this approach is also valid for the coverage of the remaining glosses (Mel'čuk (1996) lists in his typology 64 glosses in total).

Distributed vector representations (or word embeddings) (Mikolov et al., 2013c; Mikolov et al., 2013a), which we use, have proven useful in a plethora of NLP tasks, including semantic similarity and relatedness (Huang et al., 2012; Faruqui et al., 2015; Camacho-Collados et al., 2015; Iacobacci et al., 2015), dependency parsing (Duong et al., 2015), and Named Entity Recognition (Tang et al., 2014). We show that they also work for semantic retrieval of collocations. Only a small amount of collocations and big unannotated corpora have been necessary to perform the experiments. This makes our approach highly scalable and portable. Given the lack of semantically tagged collocation resources for most languages, our work has the potential to become influential in the context of second language learning. The datasets on which we performed the experiments as well as the details concerning the code and its use can be found at http://www.taln.upf.edu/content/resources/765.

6 Acknowledgements

The present work has been partially funded by the Spanish Ministry of Economy and Competitiveness (MINECO), through a predoctoral grant (BES-2012-057036) in the framework of the project HARenES (FFI2011-30219-C02-02), and by the European Commission under the grant number H2020–645012–RIA. We also acknowledge support from the Maria de Maeztu Excellence Program (MDM-2015-0502). Many thanks to the three anonymous reviewers for insightful comments and suggestions.

[6] Obviously, collocate candidates were considered as incorrect if they formed incorrect collocations with the base. Examples of such incorrect collocations are *stop* [*the*] *calm* and *develop* [*a*] *calculation*.

References

M. Alonso Ramos, L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira, and S. Prieto. 2010. Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 3209–3214, La Valetta, Malta.

J. Bahns and M. Eldaw. 1993. Should we Teach EFL Students Collocations? *System*, 21(1):101–114.

G. Bouma. 2010. Collocation Extraction beyond the Independence Assumption. In *Proceedings of the ACL 2010, Short paper track*, Uppsala.

J. Camacho-Collados, M.T. Pilehvar, and R. Navigli. 2015. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577.

R. Carlini, J. Codina-Filba, and L. Wanner. 2014. Improving Collocation Correction by Ranking Suggestions Using Linguistic Knowledge. In *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*, Uppsala, Sweden.

Y. Choueka. 1988. Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In *Proceedings of the RIAO*, pages 34–38.

K. Church and P. Hanks. 1989. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pages 76–83.

A. Cowie. 1994. Phraseology. In R.E. Asher and J.M.Y. Simpson, editors, *The Encyclopedia of Language and Linguistics, Vol. 6*, pages 3168–3171. Pergamon, Oxford.

D. Dahlmeier and H.T. Ng. 2011. Correcting Semantic Collocation Errors with L1-Induced Paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 107–117. Association for Computational Linguistics.

L. Duong, T. Cohn, S. Bird, and P. Cook. 2015. A Neural Network Model for Low-Resource Universal Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 339–348.

S. Evert. 2007. Corpora and Collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

M. Faruqui, J. Dodge, Jauhar. S.K., C. Dyer, E.H. Hovy, and N.A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1606–1615.

R. Fu, J. Guo, B. Qin, W. Che, H. Wang, and T. Liu. 2014. Learning Semantic Hierarchies via Word Embeddings. In *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics: Long Papers*, volume 1.

Y. Futagi, P. Deane, M. Chodorow, and J. Tetreault. 2008. A Computational Approach to Detecting Collocation Errors in the Writing of Non-Native Speakers of English. *Computer Assisted Language Learning*, 21(1):353–367.

Z.M. Gao. 2013. Automatic Identification of English Collocation Errors based on Dependency Relations. *Sponsors: National Science Council, Executive Yuan, ROC Institute of Linguistics, Academia Sinica NCCU Office of Research and Development*, page 550.

A. Gelbukh and O. Kolesnikova. 2012. *Semantic Analysis of Verbal Collocations with Lexical Functions*. Springer, Heidelberg.

S. Granger. 1998. Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In A. Cowie, editor, *Phraseology: Theory, Analysis and Applications*, pages 145–160. Oxford University Press, Oxford.

F.J. Hausmann. 1984. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(1):395–406.

F.J. Hausmann. 1989. Le dictionnaire de collocations. In F.J. Hausmann, O. Reichmann, H.E. Wiegand, and L. Zgusta, editors, *Wörterbücher, Dictionaries, Dictionnaires: An international Handbook of Lexicography*, pages 1010–1019. De Gruyter, Berlin/New-York.

E.H. Huang, R. Socher, C.D. Manning, and A.Y. Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

I. Iacobacci, M.T. Pilehvar, and R. Navigli. 2015. SENSEMBED: Enhancing Word Embeddings for Semantic Similarity and Relatedness. In *Proceedings of ACL*, Beijing, China, July. Association for Computational Linguistics.

A. Kilgarriff. 2006. Collocationality (and How to Measure it). In *Proceedings of the Euralex Conference*, pages 997–1004, Turin, Italy. Springer-Verlag.

M. Lewis and J. Conzett. 2000. *Teaching Collocation. Further Developments in the Lexical Approach*. LTP, London.

D. Lin. 1999. Automatic Identification of Non-Compositional Phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324. Association for Computational Linguistics.

I.A. Mel'čuk. 1996. Lexical functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamins Academic Publishers, Amsterdam.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

T. Mikolov, Q.V. Le, and I. Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. *arXiv preprint arXiv:1309.4168*.

T. Mikolov, W. Yih, and G. Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, pages 746–751.

P. Moreno, G. Ferraro, and L. Wanner. 2013. Can we Determine the Semantics of Collocations without using Semantics?. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, and M. Tuulik, editors, *Proceedings of the eLex 2013 conference*, Tallinn & Ljubljana. Trojina, Institute for Applied Slovene Studies & Eesti Keele Instituut.

N. Nesselhauf. 2005. *Collocations in a Learner Corpus*. Benjamins Academic Publishers, Amsterdam.

T. Park, E. Lank, P. Poupart, and M. Terry. 2008. Is the Sky Pure Today? awkchecker: an Assistive Tool for Detecting and Correcting Collocation Errors. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 121–130. ACM.

P. Pecina. 2008. A Machine Learning Approach to Multiword Expression Extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57, Marrakech.

S. Rodríguez-Fernández, R. Carlini, L. Espinosa-Anke, and L. Wanner. 2016. Example-based Acquisition of Fine-grained Collocation Resources. In *Proceedings of LREC*, Portorož, Slovenia.

F. Smadja. 1993. Retrieving Collocations from Text: X-Tract. *Computational Linguistics*, 19(1):143–177.

L. Tan, H. Zhang, C. Clarke, and M. Smucker. 2015. Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 657–661, Beijing, China, July. Association for Computational Linguistics.

B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu. 2014. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *BioMed research international*, 2014.

L. Wanner, B. Bohnet, and M. Giereth. 2006. Making Sense of Collocations. *Computer Speech and Language*, 20(4):609–624.

L. Wanner, G. Ferraro, and P. Moreno. 2016. Towards Distributional Semantics-based Classification of Collocations for Collocation Dictionaries. *International Journal of Lexicography, doi:10.1093/ijl/ecw002*.

J.C. Wu, Y.C. Chang, T. Mitamura, and J.S. Chang. 2010. Automatic Collocation Suggestion in Academic Writing. In *Proceedings of the ACL Conference, Short paper track*, Uppsala.