Example-based Acquisition of Fine-grained Collocational Resources

Sara Rodríguez-Fernández[1], Roberto Carlini[1], Luis Espinosa-Anke[1], 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|roberto.carlini|luis.espinosa|leo.wanner@upf.edu

Abstract

Collocations such as *heavy rain* or *make [a] decision*, are combinations of two elements where one (the *base*) is freely chosen, while the choice of the other (*collocate*) is restricted, depending on the base. Collocations present difficulties even to advanced language learners, who usually struggle to find the right collocate to express a particular meaning, e.g., both *heavy* and *strong* express the meaning 'intense', but while *rain* selects *heavy*, *wind* selects *strong*. *Lexical Functions* (LFs) describe the meanings that hold between the elements of collocations, such as 'intense', 'perform', 'create', 'increase', etc. Language resources with semantically classified collocations would be of great help for students, however they are expensive to build, since they are manually constructed, and scarce. We present an unsupervised approach to the acquisition and semantic classification of collocations according to LFs, based on word embeddings in which, given an example of a collocation for each of the target LFs and a set of bases, the system retrieves a list of collocates for each base and LF.

Keywords: collocation retrieval, collocation semantic classification, collocation resources, second language learning, word embeddings

1. Introduction

Collocations of the kind *make* [*a*] *suggestion*, *attend* [*a*] *lecture*, *heavy rain*, *deep thought*, etc., are restricted lexical co-occurrences between two syntactically related lexical elements. One of these elements (the *base*) is freely chosen, while the choice of the other (the *collocate*) depends on the base (Hausmann, 1984; Cowie, 1994; Mel'čuk, 1995). For instance, in *make* [*a*] *suggestion*, the choice of *suggestion* is free, while the choice of *make* is restricted; cf., e.g., **take* [*a*] *suggestion*. Such collocations are by their nature idiosyncratic and therefore also language-specific. Thus, you *make* [*a*] *suggestion*, but you *give* [*an*] *advice*, you *attend* [*a*] *lecture*, but you *assist* [*an*] *operation*, *rain* is *heavy* while *wind* is *strong*. In English, you *take* [*a*] *walk*, while in Spanish you 'give' it (*dar* [*un*] *paseo*, and in German you 'make' it ([*einen*] *Spaziergang machen*); in English, *rain* is *heavy*, while in Spanish and German it is 'strong' (*fuerte lluvia*/*starker Regen*). And so on. The idiosyncrasy of such collocations makes them a real challenge even for advanced second language learners (Hausmann, 1984; Bahns and Eldaw, 1993; Granger, 1998; Lewis and Conzett, 2000; Nesselhauf, 2005; Alonso Ramos et al., 2010). Nesselhauf (2005) and Farghal and Obiedat (1995) report that learners of English paraphrase their discourse in order to avoid using collocations that they do not master. In other words, a learner knows the meaning they want to express, but they fail to do it by means of a collocation, which means that they fail to pick the collocate that expresses this meaning. Semantically annotated collocation resources would thus be of great aid. A number of collocation dictionaries already either group collocates semantically (as, e.g., the Oxford Collocations Dictionary and BBI (Benson et al., 2010) for English) or use explicit semantic glosses (as, e.g., the MacMillan Collocations Dictionary for English, the LAF (Mel'čuk and Polguère, 2007) for French, and DiCE http://dicesp.com for Spanish). However, due to the high cost of their compilation, such dictionaries are often of limited coverage[1] and available only for a few languages.

In what follows, we present an unsupervised example-based approach to automatic compilation of semantically typed collocation resources. The typology that underlies our work are the glossed *lexical functions* (LFs) (Mel'čuk, 1996), which are the most fine-grained semantic collocation typology available to date. Using a state-of-the-art continuous word representation, we take as input seed a single representative example of a specific LF to retrieve from a corpus the collocates that are of the same LF (i.e., type) for new bases. So far, we focused in our experiments on Spanish. In the next section, we introduce the LF typology; in Section 3., we describe our methodology for the acquisition of collocation resources. Section 4. outlines the experiments carried out to assess the performance of the implementation of the methodology, and Section 5. discusses their outcome. Section 6., finally, presents some conclusions we draw from the presented work.

2. Lexical Functions: A Semantic Collocation Typology

Collocation dictionaries, such as the Oxford Collocations Dictionary or the MacMillan Collocations Dictionary group collocations in terms of semantic categories to facilitate that language learners can easily retrieve the collocate that expresses the meaning they want to express. However, this categorization (or classification) is not always homogeneous. For instance, in the MacMillan Dictionary, the entries for *admiration* and *affinity* contain the categories 'have' and 'show', each with their own collocates, while for other headwords, such as, e.g., *ability*, collocates with the meaning 'have' and 'show' are grouped under the same

[1]To the best of our knowledge, only for English collocations dictionaries of a reasonable coverage are available

category; in the entry for *alarm*, *cause* or *express* are not assigned to any category, while for other keywords the categories 'cause' and 'show' are used (see e.g., *problem* for 'cause' or *admiration* for 'show'); and so on. On the other hand, in the case of some headwords, the categories are very fine-grained (cf., e.g., *amount*, which includes glosses like 'very large', 'too large', 'rather large', 'at the limit', etc.), while in the case of others, it is much more coarse-grained (cf., e.g., *analogy*, for which collocates with different semantics are included under the same gloss, as, e.g., *appropriate*, *apt*, *close*, *exact*, *good*, *helpful*, *interesting*, *obvious*, *perfect*, *simple*, *useful* that all belong to the category 'good'). This lack of uniformity may confuse learners, who will expect that collocates grouped together share similar semantic features. Still, the use of semantic categories that reveal a sufficient level of detail for the presentation of collocations in dictionaries is meaningful.

In computational lexicography, categories of different granularity have been used for automatic classification of collocations from given lists; cf., e.g., Wanner et al. (in print), who use 16 categories for the classification of verb+noun collocations and 5 categories for the classification of adj+noun collocations; Moreno et al. (2013), who work with 5 broader categories for verb+noun collocations, or Chung-Chi et al. (2009), who also use very coarse-grained semantic categories of the type 'goodness', 'heaviness', 'measures', etc. But all of these categories have the disadvantage to be *ad hoc*. Therefore, we follow a different approach. As already Wanner et al. (2006), Gelbukh and Kolesnikova. (2012) and also Moreno et al. (2013) in their second run of experiments, we use the semantic typology of Lexical Functions (LFs) to classify collocations– assuming that once we obtained LF instances, they can be grouped by lexicographers into more generic coherent semantic categories. However, in contrast to these works, we acquire and classify the collocations simultaneously, while they classify only already given collocations.

As already mentioned above, LFs (Mel'čuk, 1996) are a means to typify meanings of collocates in lexical collocations. In total, about 60 "simple" types (including, e.g., 'perform', 'cause', 'realize', 'terminate', 'intense', and 'positive') are distinguished. The simple types can be combined to "complex" types; see (Kahane and Polguère, 2001) for the mathematical apparatus of the combination. For the sake of brevity, each type is labeled by a Latin acronym: 'perform' ≡ "Oper(are)", 'realize' ≡ "Real(is)", 'intense' ≡ "Magn(us)", etc. Formally, an LF can be interpreted as a function that provides, for a given base, the set of collocates that express the meaning of this LF. Consider a few examples:

Magn ('intense'):
Magn(*thought*) = {*deep*, *profound*}
Magn(*wounded*) = {*sorely*, *heavily*}

Oper$_1$ ('do', perform', 'have'):[2]
Oper$_1$(*lecture*) = {*give*, *deliver*}
Oper$_1$(*search*) = {*carry out*, *conduct*, *do*, *make*}
Oper$_1$(*decision*) = {*make*}
Oper$_1$(*idea*) = {*have*}

Real$_1$ ('realize/ do what is expected with B')[3]
Real$_1$(*temptation*) = {*succumb* [to ∼], *yield* [to ∼]}
Real$_1$(*exam*) = {*pass*}
Real$_1$(*piano*) = {*play*}

IncepOper$_1$ ('begin to do, begin to have B')
IncepOper$_1$(*fire*$_N$) = {*open*}
IncepOper$_1$(*debt*) = {*run up*, *incur*}

CausOper$_1$ ('do something so that B is performed/done')
CausOper$_1$(*opinion*) = *lead* [to ∼]

3. Methodology for the Acquisition of Collocation Resources

Taking as inspiration the Neural Probabilistic Model (Bengio et al., 2006), Mikolov et al. (2013c) proposed an approach for computing continuous vector representations of words from large corpora by predicting words given their context, while at the same time predicting context, given an input word. The vectors computed following the approaches described in (Mikolov et al., 2013a; Mikolov et al., 2013c) have been extensively used for semantically intensive tasks, mainly because of the properties that word embeddings have to capture relationships among words which are not explicitly encoded in the training data. Among these tasks are: Machine Translation (Mikolov et al., 2013b), where a transition matrix is learned from word pairs of two different languages and then applied to unseen cases in order to provide word-level translation; Knowledge Base (KB) Embedding (transformation of structured information in KBs such as Freebase or DBpedia into continuous vectors in a shared space) (Bordes et al., 2011); Knowledge Base Completion (introduction of novel relationships in existing KBs) (Lin et al., 2015); Word Similarity, Syntactic Relations, Synonym Selection and Sentiment Analysis (Faruqui et al., 2015); Word Similarity and Relatedness (Iacobacci et al., 2015); and taxonomy learning (Fu et al., 2014; Espinosa-Anke et al., 2016). From these examples, we can deduce that word embeddings provide an efficient semantic representation of words and concepts, and, therefore, may also be leveraged for the acquisition of collocational resources. Hence, we examine this hypothesis by putting forward an unsupervised algorithm for collocation acquisition which strongly relies in relational properties of word embeddings for discovering semantic relations.

3.1. Exploiting the Analogy Property

In what follows, we describe our unsupervised approach to the acquisition of (*base*, *collocate*) pairs for each individual LF within a given set LF of LFs.

Our algorithm produces, for each LF $\iota \in LF$, and a given base b_ι, a set BC of $(b_\iota, \varsigma_\iota)$ pairs, where ς_ι is a collocate which has been retrieved from a corpus in two stages. In the first stage, the similarity between the relation that ς_ι holds with b_ι and the relation held by the components of a representative collocation $\phi\iota$ (with the base b_ι^ϕ and the collocate ς_ι^ϕ) is computed. In the second stage, a filtering is applied,

[2]The index indicates the syntactic structure of the collocation. Due to the lack of space, we do not enter here in further details; see (Mel'čuk, 1996) for a detailed description.

[3]Here and henceforth 'B' stands for "base" or "keyword".

based on the collocation-specific statistical independence metric $NPMI_C$ (Carlini et al., 2014).

Algorithm 1 Collocate Discovery Algorithm

Input:

1: LF // Set of Lexical Functions
2: B // Set of manually selected bases
3: ε // Word embeddings model

Output:

4: Λ // Final resource
5: $\Lambda = \emptyset$
6: **for** $\iota \in LF$ **do**
7: **for** $b_\iota \in B$ **do**
8: $BC = \emptyset$ // Base and collocates set
9: $C = relSim\,(b_\iota, \phi_\iota, \varepsilon)$
10: **for** $\varsigma \in C$ **do**
11: $conf = NPMI_C\,(b_\iota, \varsigma)$
12: **if** $conf > \theta$ **then**
13: $BC = BC \cup \{(b_\iota, \varsigma)\}$
14: $\Lambda \cup \{BC\}$
return Λ

Algorithm 1 outlines the two stages. The first stage (lines $4-9$) consists, first, in retrieving a candidate set by means of the function $relSim$, which computes the similarity of the relation between b_ι^ϕ and ς_ι^ϕ to the relation between b_ι and a hidden word x. $relSim$ can be thus interpreted as satisfying the well-known analogy "*a* is to *b* what *c* is to x", exploiting the vector space representation[4] of *a*, *b*, and *c* to discover x (Zhila et al., 2013). Specifically, we compute $v_b - v_a + v_c$ in order to obtain the set of vectors closest to v_x by cosine distance. To obtain the best collocate candidate set, we retrieve the ten most similar vectors to x, where x is the unknown collocate we aim to find. This is done over a model trained with *word2vec*[5] on a 2014 dump of the Spanish Wikipedia, preprocessed and lemmatized with Freeling (Atserias et al., 2006).

The second stage (lines $10-14$) implements a filtering procedure by applying $NPMI_C$, an association measure that is based on pointwise mutual information, but takes into account the asymmetry of the lexical dependencies between a base and its collocate. It is computed as:

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

Following Carlini et al. (2014), we calculate $NPMI_C$ over a 7M sentence corpus compiled from Spanish newspaper material, and set the association threshold θ to 0, such that all (b_ι,ϕ_ι) collocation candidates below θ are discarded.

4. Experiments

In what follows, we outline first the setup of our experiments and present then their outcome.

4.1. Experimental Setup

For our experiments, we focus on eight of the most productive LFs in Spanish (see Table 1 for the list, along with their meanings). As mentioned in Section 3., the algorithm requires a seed example as input to the acquisition of collocates of a given LF. Therefore, for each LF, we take a representative collocation, i.e., a collocation whose collocate has a general abstract meaning similar to that of the LF, such as *crear* [*un*] *entorno* 'create [an] environment' for $CausFunc_0$ (whose meaning is 'cause that B begins to exist', 'create'), or *disminuir* [*el*] *precio* 'reduce [the] price' for CausPredMinus (whose meaning is 'decrease'. The seed examples that were chosen for each LF can be seen in Table 1.

LF	Meaning	Representative example
Magn	'intense'	*gran idea* 'great idea'
AntiMagn	'weak'	*leve cambio* 'slight change'
$CausFunc_0$	'create'	*crear* [*un*] *entorno* 'create [an] environment'
$LiquFunc_0$	'put an end'	*romper* [*una*] *amistad* 'break [a] friendship'
CausPredPlus	'increase'	*aumentar* [*el*] *precio* 'increase [the] price
CausPredMinus	'decrease'	*disminuir* [*el*] *precio* 'decrease [the] price
Bon	'good'	*dia bueno* 'good day'
Manif	'show'	*expresar afecto* 'express affection'

Table 1: Seed examples for each LF

For each LF, 20 bases were selected for the use in the experiments. The retrieved candidates for each base and for each of the target LFs were tagged as correct or incorrect by an expert lexicographer, according to two criteria: (1) whether the candidate collocates with the base forming a correct collocation and, if criterium (1) is fullfilled, (2) whether the collocate correctly belongs to the particular LF.

4.2. Outcome of the Experiments

To assess the performance of our approach, we calculated its precision, taking into account: (1) the number of candidates that correctly collocate with the base, and (2) the number of collocates that belong to the given LF. Tables 2 and 3 display the outcome of the experiments. Table 2 shows the number of collocate candidates obtained for each LF after the application of the $NPMI_C$ filter (second column); the number of correct collocations formed by the given bases and the retrieved collocate candidates (third column), and the number of correctly typed retrieved collocations with respect to each LF (fourth column). Table 3 shows the achieved precision during the identification of correct collocations and during the typification of the collocations calculated over all candidates of an LF from Table 2 (first value in the third column) and over the correctly identified collocations (second value in the third column).

In what follows, we present a brief analysis of the results. Some of the collocates that were retrieved for each LF can be seen in Table 4.

5. Discussion

With a p = 0.946, the system's performance for Magn is close to human judgement as far as the identification of collocations is concerned. The precision of the correct recognition of a collocation as Magn is somewhat lower (p = 0.797). Most of the erroneous typifications as Magn are due to two reasons: (1) semantic similarity of the collocate to the Magn (as, e.g., *creciente* 'growing'), and (2) the failure of word embeddings to distinguish Magn-collocates

[4] We denote the vector of a word as v, e.g., v_a.

[5] http://word2vec.googlecode.com/

LF	Retrieved examples
Magn	*lluvia torrencial* 'torrential rain', *viento huracanado* 'hurricane-force winds', *ruido ensordecedor* 'deafening noise', *valor incalculable* 'inestimable value'
CausFunc$_0$	*desatar* [*una*] *epidemia* 'to spark [a] pandemic', *desencadenar* [*una*] *crisis* 'to trigger [a] crisis', *redactar* [*un*] *informe* 'to draft [a] report', *promulgar* [*un*] *edicto* 'to issue [an] edict
LiquFunc$_0$	*demoler* [*un*] *edificio* 'to demolish [a] building', *apagar* [*un*] *fuego* 'to extinguish [a] fire', *resolver* [*un*] *problema* 'to solve [a] problem', *anular* [*un*] *acuerdo* 'to nullify [an] agreement'
CausPredPlus	*mejorar* [*la*] *estabilidad* 'to improve stability', *incrementar* [*la*] *cobertura* 'to increase coverage', *fortalecer* [*el*] *liderazgo* 'to strengthen leadership', *estimular* [*la*] *economía* 'to stimulate [the] economy'
CausPredMinus	*minimizar* [*un*] *valor* 'to minimize [a] value', *reducir* [*una*] *tasa* 'to reduce [a] rate', *reducir* [*un*] *salario* 'to reduce [a] salary', *minimizar* [*un*] *coste* 'to minimize costs'
Bon	*posición excelente* 'excellent position', *carrera impecable* 'impeccable career' *resultado satisfactorio* 'satisfactory result', *forma perfecta* 'perfect shape'
Manif	*manifestar* [*una*] *preocupación* 'to manifest [a] concern', *reflejar alegría* 'to reflect joy', *evidenciar* [*una*] *mejoría* 'to show improvement'

Table 4: Examples of correctly retrieved collocates for each LF

LF	#candidates	#collocations	#correct LFs
Magn	74	70	59
AntiMagn	17	12	0
CausFunc$_0$	64	49	44
LiquFunc$_0$	56	42	15
CausPredPlus	70	61	42
CausPredMinus	44	40	6
Bon	67	47	24
Manif	26	15	10

Table 2: Number of collocations found for each LF

LF	Precision (p) (identif. collocations)	Precision (p) (LFs)
Magn	0.946	0.797|0.843
AntiMagn	0.706	0.000|0.000
CausFunc$_0$	0.766	0.687|0.898
LiquFunc$_0$	0.750	0.268|0.357
CausPredPlus	0.871	0.600|0.688
CausPredMinus	0.909	0.136|0.150
Bon	0.701	0.358|0.511
Manif	0.577	0.385|0.667

Table 3: Performance of the acquisition and classification of collocations with respect to LFs

from their antonyms (as, e.g., *mínimo* 'minimal'). However, the fact that starting from *gran idea* 'great idea' we obtained such Magn-collocations as *lluvia torrencial* 'torrential rain', *viento huracanado* 'hurricane-force winds', *ruido ensordecedor* 'deafening noise' or *valor incalculable* 'inestimable value', etc. shows the potential of our approach.

In the case of CausFunc$_0$, several free combinations were judged as collocations; cf., e.g., *unificar* [*un*] *sistema* 'to unify [a] system' or *idear* [*un*] *sistema* 'to design [a] system'. Still, almost 70% of the obtained candidates are correctly typified as CausFunc$_0$; cf., e.g., *desatar* [*una*] *epidemia* 'to spark [a] pandemic', *desencadenar* [*una*] *crisis*, 'to trigger [a] crisis', *redactar* [*un*] *informe*, 'to draft [a] report' or *promulgar* [*un*] *edicto*, 'to issue [an] edict', etc.

In the case of CausPredPlus, the system performs well when detecting collocations among all the possible word combinations retrieved in the first stage. Among the collocations with the correct meaning, we obtained *mejorar* [*la*] *estabilidad* 'to improve stability', *incrementar* [*la*] *cobertura* 'to increase coverage', *fortalecer* [*el*] *liderazgo* 'to strengthen leadership', and *estimular*, *reactivar* [*la*] *economía* 'to stimulate, revive [the] economy'. However, the number of collocates that do not convey the target meaning is considerably higher for CausPredPlus. Unsurprisingly, most of them convey exactly the opposite meaning ('decrease'), which can be easily explained by the semantic similarity that antonyms show when represented by word embeddings.

As with CausFunc$_0$, with Bon and Manif, a handful of free combinations were judged as collocations. Some examples of these are *actitud sincera* 'sincere attitude', *intención indudable* 'unquestionable intention' or *aspecto inusual* 'unusual aspect' for Bon; and *reafirmar* [*el*] *apoyo* 'to reassert support', *definir* [*una*] *característica* 'to define [a] feature' or *justificar* [*un*] *temor* 'to justify [a] fear' for Manif. Precision as to whether a candidate belongs to a particular LF is somehow low for Bon and Manif. Two main aspects could be the cause of this decrease of the performance of

the approach: that the chosen seed example is not sufficiently common or general, and therefore not representative enough for the LF, or that these LFs present a wider meaning, and are thus more difficult to attain. Further research is needed to assess these issues.

Finally, as far as AntiMagn, LiquFunc_0 and CausPredMinus, whose meanings are opposite to Magn, CausFunc_0 and CausPredPlus, respectively, are concerned, the number of candidates retrieved by the system that are correct collocates remains high. However, the precision of the classification with respect to the target LFs drops significantly when compared to their *positive* counterparts. This occurs because, as stated above, word embeddings fail to distinguish between antonyms, considering words with opposite meanings as actual synonyms. Most of the collocates found for AntiMagn, LiquFunc_0 and CausPredMinus are correct instances of Magn, CausFunc_0 and CausPredPlus. For instance, we obtained *luz cegadora* 'blinding light' and *daño severo* 'severe damage' for AntiMagn; *levantar* [*un*] *edificio* 'to erect a building' and *encender* [*un*] *fuego* 'to light [a] fire' were found for LiquFunc_0; and for CausPredMinus cases such as *incrementar* [*un*] *salario* 'to increase wages' or *aumentar* [*el*] *valor* 'to increase [a] value' were obtained.

6. Conclusions

We presented a language-independent approach to automatic acquisition and fine-grained semantic classification of collocation resources with respect to Lexical Functions. Such resources are crucial for second language learning as well as for computational applications related to language production, e.g., natural language generation (Smadja and McKeown, 1990; Wanner, 1992) or machine translation (Mel'čuk and Wanner, 2001). Although there has been a large body of work on automatic retrieval of collocations (Choueka, 1988; Church and Hanks, 1989; Smadja, 1993; Kilgarriff, 2006; Evert, 2007; Pecina, 2008; Bouma, 2010), and also some works on the semantic classification of collocations (Wanner et al., 2006; Gelbukh and Kolesnikova., 2012; Moreno et al., 2013; Wanner et al., in print), to the best of our knowledge, this is the first proposal to retrieve and classify collocations simultaneously in an unsupervised manner. In our future work, we will aim to improve the precision of the classification procedure with respect to the "difficult" LFs such as AntiMagn,CausPredMinus, LiquFunc_0, etc.

7. 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), by the European Commission under the contract number H2020-RIA-645012, and by the ICT PhD program of Universitat Pompeu Fabra through a travel grant.

8. Bibliographical References

Alonso Ramos, M., Wanner, L., Vincze, O., Casamayor, G., Vázquez, N., Mosqueira, E., and Prieto, S. (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.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., and Padró, M. (2006). Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of LREC*, pages 48–55.

Bahns, J. and Eldaw, M. (1993). Should we teach EFL students collocations? *System*, 21(1):101–114.

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. (2006). Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer.

Benson, M., Benson, E., and Ilson, R. (2010). *The BBI Combinatory Dictionary of English: Your guide to collocations and grammar, Third Edition*. Benjamins Academic Publishers, Amsterdam.

Bordes, A., Weston, J., Collobert, R., and Bengio, Y. (2011). Learning structured embeddings of knowledge bases. In *AAAI*.

Bouma, G. (2010). Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010, Short paper track*, Uppsala.

Carlini, R., Codina-Filba, J., and Wanner, L. (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.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*, pages 34–38.

Chung-Chi, H., Kao, K. H., Tseng, C. H., and Chang, J. S. (2009). A thesaurus-based semantic classification of English collocations. *Computational Linguistics and Chinese Language Processing*, 14(3):257–280.

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

Cowie, A. (1994). Phraseology. In R.E. Asher et al., editors, *The Encyclopedia of Language and Linguistics, Vol. 6*, pages 3168–3171. Pergamon, Oxford.

Espinosa-Anke, L., Saggion, H., Ronzano, F., and Navigli, R. (2016). Extasem! extending, taxonomizing and semantifying domain terminologies. In *AAAI*.

Evert, S. (2007). Corpora and collocations. In A. Lüdeling et al., editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

Farghal, M. and Obiedat, H. (1995). Collocations: A neglected variable in efl. *IRAL-International Review of Applied Linguistics in Language Teaching*, 33(4):315–332.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E. H., and Smith, N. A. (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.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., and Liu, T. (2014). Learning semantic hierarchies via word embeddings. In *ACL (1)*, pages 1199–1209.

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

Granger, S. (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.

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

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). Sensembed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105.

Kahane, S. and Polguère, A. (2001). Formal foundation of lexical functions. In *Proceedings of the ACL '01 Workshop COLLOCATION: Computational Extraction, Analysis and Exploitation*.

Kilgarriff, A. (2006). Collocationality (and how to measure it). In *Proceedings of the 12th EURALEX International Congress*.

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

Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pages 2181–2187.

Mel'čuk, I. and Polguère, A. (2007). *Lexique actif du français*. de boeck, Brussels.

Mel'čuk, I. and Wanner, L. (2001). Towards a lexicographic approach to lexical transfer in machine translation (illustrated by the german-russian language pair). *Machine Translation*, 16(1):21–87.

Mel'čuk, I. (1995). Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, et al., editors, *Idioms: Structural and Psychological Perspectives*, pages 167–232. Lawrence Erlbaum Associates, Hillsdale.

Mel'čuk, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. *Lexical functions in lexicography and natural language processing*, 31:37–102.

Mikolov, T., Yih, W.-T., and Zweig, G. (2013a). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Moreno, P., Ferraro, G., , and Wanner, L. (2013). Can we determine the semantics of collocations without using semantics? In I. Kosem, et al., editors, *Proceedings of the eLex 2013 conference*.

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

Pecina, P. (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.

Smadja, F. and McKeown, K. (1990). Automatically extracting and representing collocations for language generation. In *Proceedings of the Annual ACL Conference*, pages 252–259.

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

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

Wanner, L., Ferraro, G., and Moreno, P. (in print). Towards distributional semantics-based classification of collocations for collocation dictionaries. *International Journal of Lexicography*.

Wanner, L. (1992). Lexical choice and the organization of lexical resources in text generation. In *Proceedings of the European Conference on Artificial Intelligence*.

Zhila, A., Yih, W.-T., Meek, C., Zweig, G., and Mikolov, T. (2013). Combining heterogeneous models for measuring relational similarity. In *HLT-NAACL*, pages 1000–1009.