Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 57-64 recibido 29-03-2016 revisado 18-04-2016 aceptado 03-05-2016

Semantics-Driven Collocation Discovery

Descubrimiento de Colocaciones Utilizando Semántica

Sara Rodríguez-Fernández[1], Luis Espinosa-Anke[1], Roberto Carlini[1], Leo Wanner[1,2]
[1]NLP Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra
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: *Collocations* are combinations of two lexically dependent elements, of which one (the *base*) is freely chosen because of its meaning, and the choice of the other (the *collocate*) depends on the base. Collocations are difficult to master by language learners. This difficulty becomes evident in that even when learners know the meaning they want to express, they often struggle to choose the right collocate. Collocation dictionaries, in which collocates are grouped into semantic categories, are useful tools. However, they are scarce since they are the result of cost-intensive manual elaboration. In this paper, we present for Spanish an algorithm that automatically retrieves for a given base and a given semantic category the corresponding collocates.
Keywords: collocations, collocation recognition, collocation semantic classification, second language learning, word embeddings, transformation matrix

Resumen: Las *colocaciones*, entendidas como combinaciones de dos elementos entre los cuales existe una dependencia léxica, es decir, donde uno de los elementos (la *base*) se escoge libremente por su significado, pero el otro (*colocativo*) depende de la base, suelen ser difíciles de utilizar por los hablantes no nativos de una lengua. Esta dificultad se hace visible en que estos, a menudo, aún sabiendo el significado que quieren expresar, tienen problemas a la hora de elegir el colocativo adecuado. Los diccionarios de colocaciones, donde los colocativos son agrupados en categorías semánticas son una herramienta muy útil, pero son recursos escasos y de costosa elaboración. En este artículo se presenta, para el español, un algoritmo que proporciona, dada una base y una categoría semántica, colocativos pertinentes a dicha categoría.
Palabras clave: colocaciones, reconocimiento de colocaciones, clasificación semántica de colocaciones, aprendizaje de lenguas, word embeddings, matriz de transformación

1 Introduction

Collocations such as *do* [*a*] *favour*, *take advice*, *take* [*a*] *picture*, *deep breath*, *close examination*, etc., are idiosyncratic co-occurrences of two lexical items with a direct syntactic dependency between them. One of the items (the *base*) is freely chosen by the speaker, while the selection of the other item (the *collocate*) is restricted by the base (Hausmann, 1984; Cowie, 1994; Mel'čuk, 1995). For instance, in *do* [*a*] *favour*, the choice of *favour* is free, while the choice of *do* is restricted; cf., e.g., **make* [*a*] *favour*, **take* [*a*] *favour*. The idiosyncratic nature of collocations makes them language-specific. Thus, while in English and French a picture is 'taken' (*take* [*a*] *picture*, *prendre* [*une*] *photo*), in Spanish it is 'made' (*hacer* [*una*] *foto*); in English you *spend time*, but you 'pass' it in Spanish and French (*pasar tiempo*, *passer du temps*).

Collocations are a key element in foreign language learning. They are difficult to master even for advanced students due to their idiosyncrasy (Hausmann, 1984; Bahns and Eldaw, 1993; Granger, 1998; Lewis and Conzett, 2000; Wible et al., 2003; Nesselhauf, 2005; Alonso Ramos et al., 2010). Wible et al. (2003) show that collocation errors are the most frequent of all errors in students' writings. Even when learners know the meaning they want to express, they often fail to do it by means of collocations, which, as a rule, means that they fail to select the col-

locate that expresses the intended meaning. For this reason, collocation resources that group collocations semantically can significantly contribute to second language learning. A few dictionaries of this kind already exist, see, among others, the Oxford Collocations Dictionary, MacMillan Collocations Dictionary, BBI (Benson, Benson, and Ilson, 2010), *Lexique actif du français* (LAF) (Mel'čuk and Polguère, 2007), and *Diccionario de Colocaciones del Español* (DiCE, http://dicesp.com). Some of them use explicit semantic glosses; cf., e.g., the MacMillan Collocations Dictionary for English, the LAF for French, or the DiCE for Spanish. However, since they are hand-crafted resources, the cost of their compilation is high, which explains why collocation dictionaries are usually of a limited coverage and are available only for a few languages.[1].

In this paper, we describe a *word embeddings*-based approach (Mikolov, Yih, and Zweig, 2013; Levy, Goldberg, and Ramat-Gan, 2014) to automatic compilation of semantically motivated collocation resources for Spanish. We build on the intuition that there is a linear relation between semantically similar words in embedding spaces. We exploit this linear relation to train a *function* (or *transition matrix*) that learns a semantic relation between bases and collocates (i.e., types or glosses of collocations).

The remainder of the paper is structured as follows. Section 2 presents a brief review of related work. In Section 3, the semantic glosses that are used to typify the collocations are presented. Section 4 describes the methodology for the acquisition of the resources, while Section 5 presents the performed experiments and the evaluation of their outcome. In Section 6, we discuss then the performance of the implementation of our approach. Finally, Section 7 draws some conclusions and outlines possible future work in the context of semantically-motivated automatic collocation classification.

2 Related work

In the last decades, a large body of work on automatic retrieval of collocations has been produced. Some approaches exploit statistical evidence to measure word distribution in corpora, both in isolation and in combination with other words (Choueka, 1988; Church and Hanks, 1989; Evert, 2007; Pecina, 2008). Other works combine statistical measures with syntactic information, under the assumption that only those co-occurring words that form a syntactic structure can also form a collocation (Smadja, 1993). More recently, contexts in which a pair of words co-occurs have been taken into account Bouma (2009).

With regard to the semantic classification of collocations, there seems to be a common trend to use supervised machine learning techniques for the classification of collocations against their corresponding target semantic categories, leveraging as training data lists of collocations (Wanner, Bohnet, and Giereth, 2006; Gelbukh and Kolesnikova, 2012; Moreno, Ferraro, and Wanner, 2013; Wanner, Ferraro, and Moreno, 2016).

In our previous work (Rodríguez-Fernández et al., 2016), we developed an approach for automatic extraction of collocations, which accounted for the underlying semantics of each word by means of their distributional representation. This allowed us to perform a joint process of extraction and semantic typification of collocations. The approach is based on the representation of individual words as word vectors and takes in an unsupervised setting advantage of semantic properties of word embeddings (Mikolov, Yih, and Zweig, 2013), which may be defined in terms of the well-known vector operations such as summation and subtraction. Specifically, given a particular meaning and a base, the algorithm retrieves collocates that have in combination with the given base this particular meaning. The discovery of new collocates is thus done by means of an analogy, e.g., it would attempt to discover $x = vec(\text{deafening})$ in the analogy $vec(\text{strong}) - vec(\text{wind}) + vec(\text{noise}) = x$.

As already, in (Rodríguez-Fernández et al., 2016), in our current work, we retrieve and classify collocations in semantic terms simultaneously. However, while in (Rodríguez-Fernández et al., 2016) this was done using an unsupervised learning model, here we draw upon a supervised model.

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

3 Semantic collocation typology

It is common that different bases prompt for different collocates to express a given meaning. For instance, to express that a *disease* is 'intense', the collocates *serious* or *dangerous* can be used. To express that a person 'is affected by' a *disease*, *suffer* or *have* is used. If someone 'starts having' a disease, *catch, get* or *contract* are preferable, while when there is a person 'causing' a disease in someone else, *give, transmit* or *pass on* will be used, and so on. In Spanish, an 'intense' *disease* (*enfermedad*) is *grave* 'grave'. *Sufrir* 'suffer', *padecer* 'endure' or *tener* 'have' can be used to express that a person 'is affected by' a *disease*. *Contraer* 'contract' is preferred for 'start having', while for 'cause' *contagiar* 'pass on' or *transmitir* 'transmit' should be used instead.

As already mentioned above, collocation dictionaries, such as the Oxford Collocations Dictionary or the MacMillan Collocations Dictionary for English, or the *Diccionario de Colocaciones del Español* (DiCE) for Spanish classify collocations into semantic categories such that language learners can find more easily the collocate that communicate the meaning they intend to express. Categories of different granularity are used in each case. Similarly, different works on automatic classification of collocations use as target classes categories of different granularity. For instance, Wanner, Ferraro, and Moreno (2016) use 16 categories to classify verb+noun collocations and 5 for adj+noun collocations; Moreno, Ferraro, and Wanner (2013) and Chung-Chi et al. (2009) classify collocations into broader categories; Wanner, Bohnet, and Giereth (2006), Gelbukh and Kolesnikova (2012) and also Moreno, Ferraro, and Wanner (2013) in their second run of experiments use the semantic typology of *Lexical Functions* (LFs) (Mel'čuk, 1996), also used in DiCE.

In our experiments, we use a subset of ten LFs . For all of these LFs, we define semantic glosses similar to those used in the MacMillan Collocations Dictionary, in order to make the LFs more transparent to users. Some examples are 'intense', 'perform', 'increase' and 'show'; cf. Table 1 for the list of glosses and examples that illustrate them.

4 Methodology

As argued by Mel'čuk (1996), the meaning of collocates across collocations can be captured in a generic semantic (lexical function, LF) typology. For convenience, Mel'čuk defines for each LF a Latin acronym (such as 'Oper', 'Func', 'Magn', etc.), but, in general, for each LF also a semantic gloss is available. 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. Similarly, in Spanish, *ensordecedor* 'deafening' in *ruido ensordecedor* 'deafening noise', *alta* 'high' en *alta estima* 'high esteem' or *fuerte* 'strong' in *fuerte golpe* 'strong blow', can be glossed as 'intense', and so on. Our goal is to capture the relation that holds between the training bases and collocates that share the same gloss, such that, given a new base and a gloss, we can retrieve the corresponding collocate(s) of this new base pertinent to 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*. Our approach is based on Mikolov, Le, and Sutskever (2013)'s translation matrix, where word vector representations between two analogous spaces are found to be linearly related. In Mikolov et al.'s original work, which describes the potential of this property for Machine Translation, 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, and in the same language. To obtain the word vector representations in $\mathcal{B}$ and $\mathcal{C}$, we use Mikolov, Yih, and Zweig (2013)'s *word2vec*.[2]

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 corresponding base and collocate of a collocation $t_i \in \mathbf{T}$. Then, we may denote a *base* matrix as $B_\tau = [b_{t_1}, b_{t_2} \dots b_{t_n}]$, and a *collocate* matrix as $C_\tau = [c_{t_1}, c_{t_2} \dots c_{t_n}]$, given by the corresponding vector representations of each collo-

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

Semantic gloss	Example	# instances
'intense'	*sumo cuidado* 'extreme care'	174
'weak'	*cantidad insignificante* 'negligible amount'	23
'perform'	*dar* [*un*] *abrazo* 'to give [a] hug'	319
'begin to perform'	*tomar posesión* 'to take possession'	67
'stop performing'	*renunciar* [*a un*] *papel* 'to abandon [a] role'	3
'increase'	*fortalecer* [*el*] *control* 'to strengthen control'	22
'decrease'	*bajar* [*un*] *impuesto* 'to lower [a] tax'	16
'create', 'cause'	*escribir* [*una*] *carta* 'to write [a] letter'	181
'put an end'	*apagar* [*un*] *fuego* 'to extinguish [a] fire'	31
'show'	*expresar disconformidad* 'to express disagreement'	5

Table 1: Semantic glosses and size of training set

cation component. Together, they constitute a set of training examples Φ_τ composed by vector pairs $\{b_{t_i}, c_{t_i}\}_{i=1}^{n}$.

We learn a linear transformation matrix from Φ_τ, denoted as $\Psi_\tau \in \mathbb{R}^{\mathcal{B}\times\mathcal{C}}$. Specifically, and following the notation in (Tan et al., 2015), this transformation may be depicted as:

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

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

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

At the end of this procedure, each time our algorithm observes a novel base b_{j_τ}, a novel list of ranked collocate candidates is retrieved by applying $\Psi_\tau b_{j_\tau}$. The obtained list of candidates is filtered in terms of part of speech (only plausible PoS patterns are admitted as candidates) and in terms of the $NPMI$ metric. The $NPMI$ metric is an association measure based on pointwise mutual information that factors in the semantic asymmetry between the base and the collocate (Carlini, Codina-Filba, and Wanner, 2014):

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

Such a combination of heterogeneous models has been used before and proved to be effective to discover other types of relationships between word pairs (Zhila et al., 2013).

5 Experiments

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

5.1 Experimental setup

We carried out our experiments on the ten semantic collocate glosses listed in the first column of Table 1: eight verbal collocate glosses in verb+noun collocations and two property glosses in adj+noun collocations, first without filtering the obtained candidate list and then applying the PoS and $NPMI$ filters. The training examples for each of the glosses in our experiments were taken from a three thousand sentence corpus in which collocations were manually annotated and classified with respect to LFs.[3] Duplicates were removed. However it was common to find more than one collocate for each base. Ten instances for each gloss were set apart for testing. Since the distribution of collocations with different glosses is not homogeneous (e.g., collocations conveying the idea of 'intense' are used more often than those conveying the idea of 'weak', and those meaning 'perform' are more used than those meaning 'stop performing'), in our data, the number of instances per gloss also varies significantly (see Table 1 for the number of training instances for each gloss).

Both bases and collocates were modeled by training their word vectors over a 2014 dump of the Spanish Wikipedia. For the calculation of $NPMI$ during the postprocessing stage, a seven million sentence newspaper corpus was used.

5.2 Evaluation

To assess the outcome of the experiments, the correctness of each candidate from the top-10 that were retrieved for each test base was verified. Given that a base can have different collocates to express a meaning, the evaluation was not performed automatically against

[3] Recall that our glosses correspond to lexical functions.

the collocates found in the corpus; instead, each candidate was manually judged as correct or incorrect. For the outcome of each experiment, we computed both *Precision* (P) as the ratio of collocates with the targeted gloss retrieved 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 ith 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 compared the performance of our setup to the accuracy achieved in our previous work (Rodríguez-Fernández et al., 2016) (see also Section 2 above), which serves us as baseline, also in two variants: without and with PoS+$NPMI$ filters. The results of our experiments are shown in Table 2 ('S1' stands for the baseline configuration in which all top-10 retrieved candidates are kept; 'S2' for the baseline configuration with PoS+$NPMI$ filtering; 'S3' for our current transformation matrix-based setup without filtering; and 'S4' for the matrix-based setup with PoS+$NPMI$ filtering.

6 Discussion

As we can observe from the number of instances in Table 1, certain glosses seem to possess less linguistic variability, requiring a lower number of instances for building the transformation matrix from bases to collocates. For example, the transformation function of 'stop performing', trained with only 3 instances, achieves the second best results both for P and MRR.

Comparing the unfiltered configurations of both the baseline and our approach to their filtered counterparts, an evident increase of precision can be seen. This means that the incorporation of a filtering module, especially the $NPMI$, improves the performance of the algorithms substantially. For example, *suscitar* 'raise', as candidate for the base *infección* 'infection', was discarded by the $NPMI$, while *provocar* 'provoke' and *producir* 'produce', with $NPMI$ 0.30 and 0.31, were kept. Similarly, for the base *velocidad* 'speed', *amplio* 'wide' was discarded, while *máxima* 'maximum', *gran* 'great', *vertiginosa* 'vertiginous' and *alta* 'high', with $NPMI$ 0.46, 0.51, 0.51 and 0.77, were kept.

After close examination of the candidates, we found that a great number of the candidates retrieved and filtered by our system were actually correct collocates. However, their meaning was somewhat different to that of the semantic gloss. A very common source of error are antonym words, since our approach is based on word embeddings, i.e. vector representations of words based on their contexts. Antonyms often share the same linguistic context and are therefore considered as similar words by the model. Consider the following examples as illustration:

(1) *voz tenue*, 'faint voice' (belongs to 'weak' instead of 'intense')

(2) *fuerte tensión*, 'strong tension' (belongs to 'intense' instead of 'weak')

(3) *aumentar* [*una*] *tasa*, 'to increase [a] rate' (belongs to 'increase' instead of 'decrease')

(4) *derribar* [*un*] *templo*, 'to demolish [a] temple' (belongs to 'put an end' instead of 'create', 'cause')

(5) *plantear* [*una*] *duda*, 'to raise [a] question' (belongs to 'create', 'cause', instead of 'put an end')

However, the fact that we are able to obtain such 'intense' collocations as *velocidad vertiginosa* 'vertiginous speed', such 'put an end' collocations as *resolver* [*una*] *duda* 'solve [a] doubt', or such 'increase' collocations as *encarecer* [*un*] *precio* 'to increase [a] price' shows the potential of our approach.

A look at Table 2 may furthermore give the impression that the overall numbers are still rather low (e.g., for 'begin to perform' we achieve only 0.15 of precision, for 'decrease' only 0.19, etc.). In this respect, it should be noted that in our evaluation, the retrieved collocate candidates were considered correct only if they were both correct collocates and belonged to the target semantic gloss. In other words, for a candidate to be correct it was required not only to *collocate* with the base, but also to belong to the target semantic category. However, it is well-known that it is by far not always clear whether a given co-occurrence forms a collocation or a free word combination. If we relax our evaluation in the

	Precision				Mean Reciprocal Rank			
Semantic gloss	S1	S2	S3	S4	S1	S2	S3	S4
'intense'	0.25	0.12	0.17	**0.44**	0.52	0.10	0.31	**0.42**
'weak'	0.0	0.0	0.10	**0.45**	0.00	0.00	**0.75**	0.60
'perform'	0.09	0.00	**0.20**	0.16	0.19	0.00	**0.44**	0.25
'begin to perform'	**0.15**	0.00	0.03	**0.15**	**0.29**	0.00	0.04	0.08
'stop performing'	0.04	0.04	0.06	**0.44**	0.1	0.07	0.35	**0.53**
'increase'	0.20	0.08	0.25	**0.50**	0.51	0.17	0.63	**0.67**
'decrease'	0.04	0.09	0.08	**0.19**	0.07	0.10	0.35	**0.43**
'create', 'cause'	0.07	0.13	**0.21**	0.20	0.38	0.13	**0.57**	0.38
'put an end'	0.06	0.05	0.16	**0.23**	0.26	0.02	0.32	**0.43**
'show'	0.02	0.00	0.31	**0.33**	0.20	0.00	**0.85**	0.55

Table 2: Precision and MRR for the baselines (S1 and S2) and the two configurations of our approach (S3 and S4)

Semantic gloss	**Base**	**Retrieved candidates**
'intense'	*velocidad*	*alto, máximo, constante, gran, considerable, vertiginoso*
	'speed'	'high, maximum, constant, great, considerable, vertiginous'
'weak'	*plazo*	*breve, corto, largo, prorrogable*
	'period'	'brief, short, long, extendable'
'perform'	*viaje*	*hacer, embarcar, efectuar, realizar, iniciar, preparar, topar*
	'trip'	'make, load, carry out, make, initiate, prepare, bump into'
'begin to perform'	*éxito*	*alcanzar, medir, suponer, rebasar, propiciar, presumir, presagiar*
	'success'	'attain, measure, suppose, overflow, propiciate, boast, foretell'
'stop performing'	*escondite*	*abandonar*
	'hiding place'	'abandon'
'increase'	*producción*	*incentivar, fomentar, promover, alentar, potenciar, fortalecer*
	'production'	'incentive, foster, promote, encourage, improve, strenghten'
'decrease'	*pérdida*	*reducir, moderar, frenar, compensar, disminuir, elevar*
	'loss'	'reduce, moderate, brake, compensate, decrease, increase'
'create', 'cause'	*templo*	*construir, erigir, levantar, edificar, derribar*
	'temple'	'build, erect, raise, build, demolish'
'put an end'	*duda*	*resolver, solventar, plantear, zanjar*
	'doubt'	'solve, resolve, set out, settle'
'show'	*opinión*	*expresar, manifestar, reflejar, resumir, plasmar, exponer*
	'opinion'	'express, manifest, reflect, summarize, express, expound'

Table 3: Examples of retrieved collocations

sense that a candidate is judged to be correct if it belongs to the target semantic category, no matter whether it is considered to form with the base a collocation in the strict sense or not,[4] the precision is likely to increase. For instance, for English, we observe that for 'intense', 'put an end' and 'show', it increases 0.1, 0.18 and 0.15 points, respectively. For other glosses, the increase is minor, as, e.g., in the case of 'begin to perform' or 'stop performing', for which the increase is only 0.04 and 0.02.

7 *Conclusions and future work*

We have presented an approach to automatic compilation of semantically-motivated collocation resources. Our technique is grounded in Mikolov, Yih, and Zweig (2013)'s word embeddings and the assumption that semantically related words in two different vector representations are related by linear transformation (Mikolov, Le, and Sutskever, 2013). This property has also been exploited for other tasks, such as word-based translation (Mikolov, Le, and Sutskever, 2013), learning

[4]Thus, combinations such as *gran* in *gran tamaño*, *hacer* in *hacer* [*un*] *movimiento* or *bajo* in *salario bajo* would be considered correct collocates of the glosses 'intense', 'perform' and 'weak', respectively, while *amplia* in *amplia velocidad*, *hacer* in *hacer* [*una*] *decisión*, or *suscitar* in *suscitar* [*una*] *infección* would be rejected as collocates of the glosses 'intense', 'perform' and 'cause', respectively.

semantic hierarchies (hyponym-hypernym relations) in Chinese (Fu et al., 2014), or modeling linguistic similarities between standard and non-standard language (Tan et al., 2015). For our task of collocation discovery, we learn a series of *transition matrices* (one for each target semantic gloss) over a handful of collocation examples, where collocates share the same gloss, and then apply these matrices to discover, for any previously unseen base, new collocates that belong to the same semantic type. In the paper, we discussed the outcome of the experiments with ten different glosses such as 'do / perform', 'increase' or 'intense', and show that for most glosses, an approach that combines a stage of the application of a gloss-specific *transition matrix* and a pruning stage based on statistical evidence outperforms baselines which exploit only one of these stages or a baseline that is based on the embeddings property for drawing analogies (Rodríguez-Fernández et al., 2016).

Here, we focused on Spanish and only on a small amount of collocations. However, since our approach requires only big unannotated corpora, it is highly scalable and portable to other languages. 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.

In the future, we plan to investigate whether increasing the number of training instances, and using word embeddings trained on a corpus richer in collocations may affect the performance of the system. We also plan to extend our work by increasing the number of semantic glosses, thus generating more complete resources.

8 Acknowledgements

The present work has been 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 the Maria de Maeztu Excellence Program (MDM-2015-0502).

References

Alonso Ramos, M., 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.

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

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

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In C. Chiarcos, R. Eckart de Castilho, and M. Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference*. Gunter Narr Verlag, Tübingen, pages 31–40.

Carlini, R., 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.

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

Church, K. and P. Hanks. 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 and J.M.Y. Simpson, editors, *The Encyclopedia of Language and Linguistics, Vol. 6*. Pergamon, Oxford, pages 3168–3171.

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

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

Gelbukh, A. and O. Kolesnikova. 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*. Oxford University Press, Oxford, pages 145–160.

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

Levy, O., Y. Goldberg, and I. Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180.

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

Mel'čuk, I. 1995. Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, editors, *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, Hillsdale, pages 167–232.

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.

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

Mikolov, T., Q.V. Le, and I. Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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

Moreno, P., 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*.

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.

Rodríguez-Fernández, S., R. Carlini, L. Espinosa-Anke, and L. Wanner. 2016. Example-based acquisition of fine-grained collocation resources. In *Proceedings of LREC*, Portoroz, Slovenia.

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

Tan, L., H. Zhang, C.LA. Clarke, and M.D. Smucker. 2015. Lexical comparison between wikipedia and twitter corpora by using word embeddings. *Volume 2: Short Papers*, page 657.

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

Wanner, L., 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.

Wible, D., C.H. Kuo, N.L. Tsao, A. Liu, and H.L. Lin. 2003. Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 19(1):90–102.

Wu, J.-C., 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.

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