CollFrEn: Rich Bilingual English–French Collocation Resource

Beatriz Fisas[1], **Luis Espinosa-Anke**[2], **Joan Codina-Filbà**[1], **Leo Wanner**[3,1]
[1]NLP Group, Pompeu Fabra University, Barcelona
[2]School of Computer Science and Informatics, Cardiff University, UK
[3]Catalan Institute for Research and Advanced Studies (ICREA)
firstname.lastname@upf.edu, espinosa-ankel@cardiff.ac.uk

Abstract

Collocations in the sense of idiosyncratic lexical co-occurrences of two syntactically bound words traditionally pose a challenge to language learners and many Natural Language Processing (NLP) applications alike. Reliable ground truth (i.e., ideally manually compiled) resources are thus of high value. We present a manually compiled bilingual English–French collocation resource with 7,480 collocations in English and 6,733 in French. Each collocation is enriched with information that facilitates its downstream exploitation in NLP tasks such as machine translation, word sense disambiguation, natural language generation, relation classification, and so forth. Our proposed enrichment covers: the semantic category of the collocation (its *lexical function*), its vector space representation (for each individual word as well as their joint collocation embedding), a subcategorization pattern of both its elements, as well as their corresponding BabelNet id, and finally, indices of their occurrences in large scale reference corpora.

1 Introduction

Collocations in the sense of idiosyncratic lexical co-occurrences of two syntactically bound words are central to second language (L2) learning (Hausmann, 1984; Bahns and Eldaw, 1993; Granger, 1998; Lewis and Conzett, 2000; Nesselhauf, 2005; Alonso Ramos et al., 2010) and various NLP applications – including, e.g., word sense disambiguation (Maru et al., 2019), parsing and machine translation (Seretan, 2013), and natural language generation (Wanner and Bateman, 1990; Smadja and McKeown, 1991). However, manually compiled and semantically annotated large scale collocation datasets are scarce.[1] Even more scarce are aligned multilingual collocation resources, which are instrumental for any cross-language application. In what follows, we present a manually compiled and semantically annotated bilingual (English–French) collocation resource. In order to facilitate its uptake in different applications, we enrich the collocations in this resource with additional information: each collocation is assigned its semantic category in terms of a lexical function (Mel'čuk, 1996) and its corresponding relation embedding (Espinosa-Anke et al., 2019). The individual collocation elements are also embedded using Mikolov et al. (2013)'s skipgram algorithm. The functional head of each collocation is furthermore disambiguated against BabelNet (Navigli and Ponzetto, 2012), which facilitates the alignment between the equivalent English and French heads (as, e.g., between Eng. *charges* and Fr. *accusations* in *dismiss the charges* and *rejeter les accusations*). To allow for the consultation of the use of a collocation in context (be it for second language learning or model training), for each collocation, sentences from large scale English and French corpora in which they occur are also released.

The remainder of the paper is structured as follows. In Section 2, we provide some background on the notion of collocation, point to some of the available collocation (or, in more general terms, *multiword expression*) resources, and introduce the concept of lexical function. Section 3 outlines the types of information by which we enrich each collocation in our English and French collocation lists. Section 4

[1] To the best of our knowledge, the largest datasets of this kind are currently the *Lexical Systems* developed at the ATILF Laboratory for several languages https://perso.atilf.fr/apolguere/projects/; cf. (Polguère, 2014) for the theoretical background.

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 1–12
Barcelona, Spain (Online), December 13, 2020.

describes how the resource is organized. Section 5 finally, draws some conclusions from the presented resource and outlines several tasks that we are about to tackle in order to enhance and further enrich it.

2 Background

Despite the fact that an increasing number of works addresses the challenge of collocation extraction and classification, the diverging interpretations of the term "collocation" that underline these works call for a clear statement of what we mean by "collocation", before providing any further details on our resource. In what follows, we thus fist define the notion of "collocation" as we use it. In the following subsections we then review the available resources of multiword expressions, of which collocations are one type, and introduce the *lexical function*-based categorization in terms of which the collocations in our resource are classified.

2.1 On the notion of collocation

The interpretation of the phenomenon of collocation underlying the presented resource is that of an idiosyncratic binary word combination of two syntactically bound and semantically related lexical elements (Kilgarriff, 2006), such that the meaning of one of the elements (the *collocate*) is determined by its co-occurrence with the other element (the *base* or semantic head of the collocation). For instance, in *give* [*an*] *advice*, *take* [*a*] *walk*, or *deliver* [*a*] *speech*, the meaning of *give*, *take* and *deliver* (namely 'perform') is determined by *advice*, *walk* and *speech* respectively. Analogously, the meaning 'intense' of *heavy*, *big*, and *strong* in *heavy storm*, *big surprise* and *strong argument* is determined by *storm*, *surprise* and *storm* respectively.

In the light of the dependence of the meaning of the collocate on the base (e.g., in the collocation *hot topic hot* stands for 'relevant' or 'prominent', in the collocation *hot debate* for 'intense' and in the free combination *hot surface* for 'high temperature'), the value of semantically tagged (or disambiguated) collocation resources for human and machine has been repeatedly emphasized and taken up both in lexicography and in NLP. 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 – even if the categories are not always homogeneous. For instance, in the MacMillan Dictionary, the entries for *admiration* and *affinity* contain the categories 'have' and 'show'; in the entry for *ability*, collocates with the meaning 'have' and 'show' are grouped under the same category; in the entries *problem* and *admiration*, the categories 'cause' and 'show' are explicitly distinguished; and so on.

In computational lexicography, on the other hand, semantic categories of different granularity have been used for automatic classification of collocations; cf., e.g., Wanner et al. (2016), who use 16 categories for automatic 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. In contrast, for instance, Wanner (2004), Wanner et al. (2006), Gelbukh and Kolesnikova (2012), and Garcia et al. (2019) use the most fine-grained semantic typology of collocations available in the field: the typology of lexical functions (LFs) developed in the context of the Explanatory Combinatorial Lexicology (ECL) (Mel'čuk, 1996). LFs have the advantage that due to their level of detail, they can be used as semantic units in semantic structures and, if needed, for particular applications they can be generalized.[2] Moreover, their cross-language consistency has been validated on a large number of language families. Following the tradition in ECL, in the resource we introduce, collocations are categorized according to their LF.

2.2 A glance at available collocation resources

Printed LF dictionaries of limited coverage are available for French (Mel'čuk *et al.*, 1984 1988 1992 1999; Mel'čuk and Polguère, 2007) and Russian (Mel'čuk and Žolkovskij, 1984). For

[2] As a matter of fact, the broader categories used in (Wanner et al., 2016) have been obtained by the generalization of the LF typology.

French https://www.ortolang.fr/market/lexicons/lexical-system-fr and Spanish http://www.dicesp.com, online resources are available, which can be consulted via dedicated web interfaces, but not downloaded for NLP use. Experiments on enriching WordNet with LFs have been reported earlier in the literature (Wanner et al., 2004; Espinosa-Anke et al., 2016). Alonso Ramos et al. (2008) discuss the compilation of an LF-based collocation resource from the FrameNet corpus https://framenet.icsi.berkeley.edu/, but do not make any resource available. In the context of learning Spanish as second language, Alonso Ramos et al. (2015) facilitate web interface-based retrieval of Spanish LF instances extracted from a large newspaper corpus, but, again, without providing any collocation resource. Similarly, for English learners, a number of works discuss the extraction of (semantically unlabeled) collocations from corpora; cf., e.g., (Chang et al., 2008; Liu et al., 2009).

Semantically unlabeled collocation databases have been compiled for a number of languages using Sketch Engine[3]; cf.: https://www.lexicalcomputing.com/language-databases-tools-solutions/collocation-databases/.

Apart from general collocation databases, some resources are available that focus on specific types of multiword expressions, such as, e.g., phrasal verbs (Tu and Roth, 2012) or Light Verb Constructions (Tu and Roth, 2011).

To the best of our knowledge, no resources as proposed in this paper are available.

2.3 Lexical Functions

Formally, a *lexical function* (LF) can be interpreted as a function that provides, for a given lexical item (referred to as 'keyword' or 'base'), the set of its values (= 'collocates') that express the meaning of this LF. In total, about 60 "simple" LFs (including, e.g., 'perform', 'cause', 'realize', 'terminate', 'intense', and 'positive') are distinguished. Simple LFs can be combined into "complex" LFs; see (Kahane and Polguère, 2001) for the mathematical apparatus of this combination. For the sake of brevity and transparency, each LF is labeled by a Latin acronym: 'perform' ≡ "Oper(are)", 'realize' ≡ "Real(is)", 'intense' ≡ "Magn(us)", etc. Consider, for illustration, a few examples of notably frequent LFs in English. Indices indicate the subcategorization patterns of the collocate+base structures ('1': the first semantic argument of the base is realized as the grammatical subject, '2': the second semantic argument of the base is the grammatical subject, etc.).

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

Oper$_1$ ('do', 'perform', 'have'):[4]
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')[5]
Real$_1$(*temptation*) = {*succumb* [to $\sim$],
yield [to $\sim$]}
Real$_1$(*exam*) = {*pass*}
Real$_1$(*piano*) = {*play*}

IncepOper$_1$ ('begin to do B', '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 $\sim$]

Note that the set of simple LFs contains *syntagmatic* and *paradigmatic* LFs. A syntagmatic LF captures a specific idiosyncratic relation between the keyword and the value such that both co-occur with each other. In other words, syntagmatic LFs are genuine collocations. Magn, Oper$_1$, Real$_1$, InceptOper$_1$

[3] https://www.sketchengine.eu/

[4] As already pointed out above, the index indicates the syntactic structure of the collocation. 'i' stands for a structure in which the i-th semantic actant of the base is realized as grammatical subject.

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

and CausOper$_1$ cited above are typical syntagmatic LFs. A paradigmatic LF captures a specific idiosyncratic relation between the keyword and the value, such that one can substitute the other. Examples of paradigmatic LFs are Syn(onymy): **Syn**(*car*) = *automobile*, Mult(itude): **Mult**(*player*) = *team*, Gener(al): **Gener**(*car*) = *vehicle*, and others. In our resource, we currently capture only syntagmatic LFs since we are interested in collocations.

3 Composition of the collocation resource

As shown by Maru et al. (2019)'s experiments with (Espinosa-Anke et al., 2016)'s ColWordNet,[6] the mere annotation of collocational information with LF tags is already useful for word sense disambiguation. However, LF-tagged collocations can be further enriched to be even more useful for state-of-the-art NLP applications. For instance, instead of lexical items as such, it is very common to use their embeddings. Espinosa-Anke et al. (2019) have also shown that the embedded relation vectors of collocations differ from the embedded vectors of other perhaps better known semantic relations such as hypernymy or meronymy. Furthermore, the sentential contexts of the occurrences of collocations in corpora is an additional signal that can (and should) be used, even more so with the breakthrough of language models and their capacity to generate better multiword expression representations thanks to, precisely, observing their textual context. Our goal is thus to provide a bilingual collocation resource that is enriched with all of this information.

3.1 Collocation lists and corpora

The base of our collocation resource are lists of English and French collocations manually tagged with LFs as well as reference corpora for both languages.

3.1.1 Collocation Lists

We start from lists of syntagmatic lexical function instances, i.e., collocations, with the LF labels assigned to them (see Section 2.3), in English and French, retrieved manually over a number of years by I. Mel'čuk from different online sources and printed material. The English list contains in total 7,480 syntagmatic LF instances, almost evenly distributed between verb+noun (50.1%) and noun+adjective/adverbial+verb (49.9%) collocations. Among verb+noun collocations, Oper$_{1/2/3}$ are the most frequent (accounting for 32.9% of all captured verb+noun collocations) and among the noun+adjective/adverbial+verb collocations, Magn (accounting for 74,2% of all captured noun+adjective/adverbial+verb collocations) are the most frequent.

The French list contains 6,733 syntagmatic LF instances, with a distribution 53.6%:45.5%:0.9% between verb+noun, noun+adjective/adverbial+verb, and preposition+noun collocations. Similar to the English list, Oper$_{1/2/3}$ and Magn dominate (with 41% of Oper$_i$ and 76,7% of Magn in the respective syntactic pattern).[7] Table 1 lists the frequencies of the collocations of the 10 most frequent syntagmatic LFs (with their semantic glosses) in both English and French and their "density", i.e., the ratio between the distinct bases that appear in collocations tagged with a specific LF and the total number of distinct bases in our dataset (2,277 for English and 2,444 for French).[8]

Table 2 displays the distribution of the number of collocates across bases for both English and French LF instances in our resource.

3.1.2 Reference corpora

Reference corpora serve us, on the one hand, to obtain the collocation embedding vectors with which we enrich the collocation lists (see Subsections 3.2.2, and 3.2.3 below), and, on the other hand, as source of collocations in use: both the English and French LF instances are linked to their occurrences in such corpora. The occurrence contexts can be used for illustration of the contextualized use of a collocation in

[6] ColWordNet is an extended WordNet enriched with information of eight different LFs.

[7] In addition, the lists contain 2626 English paradigmatic LF instances and 2110 French paradigmatic LF instances. As pointed out above, they are not included so far in our resource; see also future work in Section 5.

[8] Note that '#' and 'ρ' are different because a single base can co-occur with different collocates with the meaning of the same LF. Real$_1$ and Real$_2$ have the same meaning, namely 'fulfil (the role) assigned by the semantic frame of the base'; only that their syntactic structure is different. Cf., Real$_1$(*law*) = [*to*] *enforce*, Real$_2$(*law*) = *abide*.

	LF gloss	English		French	
		#	ρ	#	ρ
Magn	‘intense’	2,758	0.37	2,366	0.35
Oper1	‘perform’	1,040	0.14	1,258	0.19
Real1	‘fulfil’	316	0.04	277	0.04
AntiMagn	‘weak’	304	0.04	207	0.04
IncepOper1	‘begin to perform’	221	0.03	265	0.04
AntiBon	‘negative’	210	0.03	228	0.03
Oper2	‘undergo’	187	0.03	216	0.03
CausFunc0	‘cause existence of’	150	0.02	150	0.02
Real2	‘fulfil’	144	0.02	99	0.01
Bon	‘positive’	137	0.02	113	0.02

Table 1: Most frequent LFs in our resource. ‘ρ’ stands for “density” of an LF.

#collocates	English (% bases)	French (% bases)
1	44	46.5
2	19.4	15.5
3	10.6	8.9
4	6.6	5.1
5	4.1	4.2
6–10	9.9	8.5
>10	5.5	3

Table 2: Distribution of the collocates across the different bases in the English and French LF instances lists

second language teaching contexts or in online collocation dictionaries. They can also serve as targeted training material to fine-tune language models. In other words, they allow for a more varied use of the occurrence contexts than sample sentence copies, as, e.g., in the Spanish online collocation dictionary DiCE http://www.dicesp.com/.

Not all corpora are equally suited for our purposes. For example, it is likely to expect more occurrences of collocations in general discourse than in encylopedia-like corpora such as Wikipedia. This intuition has been evaluated in the past, where two separate vector spaces were learned for bases and collocates, and showed that indeed a less constrained corpus is likely to produce better collocation representations (Rodríguez Fernández et al., 2016). Thus, we use Gigaword for the English portion of our resource, and for French the spoken language corpus ORFÉO (Benzitoun et al., 2016), the Corpus Est Républicain `http://redac.univ-tlse2.fr/corpus/estRepublicain.html` analyzed with Talismane (Urieli, 2013) and the newspaper corpus frWaC corpus from the Wacky corpus collection (Baroni et al., 2009). The English corpus contains about 150 million sentences, while the French corpora contain 62 million of sentences in total.

In total, 6,528 different LF instances from the English list (88% of coverage) and 5,731 different LF instances from the French one (85% of coverage) occur in these corpora. Table 3 summarizes the distribution of the occurrences across the most common 10 LFs. In the English corpus, we have found an average of 4,026 sentences for each collocation, totaling 26.6 million sentences, while in the French corpora we found 1,094 sentences for each collocation, totaling 5 million sentences.

3.2 Enriching lists of collocations

In what follows, we describe the information with which the lists of English and French LF-tagged collocations are enriched.

LF	English	French
Magn	30,0%	15,6%
Oper1	31,2%	52,5%
Real1	4,6%	6,2%
AntiMagn	2,5%	1,3%
IncepOper1	5,9%	4,16%
AntiBon	0,7%	0,5%
Oper2	4,6%	3,7%
CausFunc0	2,8%	2,9%
Real2	1,4%	2,6%
Bon	1,2%	1,1%

Table 3: The distribution of the instances of the most common 10 LFs in the reference corpora.

3.2.1 Subcategorization information

The base or the collocate of a collocation may imply idiosyncratic subcategorization restrictions, which constitute useful information. For instance, in *go* [*for*] [*a*] *walk* the collocate *go* requires the preposition *for*, and the base *walk* an indefinite article.[9] In our resource, the subcategorization restrictions of the collocation elements are captured; cf. a few French and English examples (the information following the following pattern: 'b(ase) | bpos | c(ollocate) | c.subcat'):
boast | ART | *feed* | –
brake | ART | *step* | *on*
habit | ART | *fall out* | *of*
hope | ART | *feed* | *of* ARG1
...
bataille 'battle' | ART | *se lancer*, 'launch o.s.' | *dans* 'in'
observation 'observation' | NULL | *être* 'be' | *sous* 'below'
sommeil 'sleep' | ART | *sortir*, lit. 'exit'| *de* 'from'
virus 'virus' | *le* 'the' | 'catch' | –
...

Such refined information may have an impact, for instance, on the treatment of function words by downstream NLP tasks, better multiword expression single-tokenization, and certainly on a more accurate collocation classification in the context of second language learning.

3.2.2 Embedding of collocation elements

Embeddings are the most common representations of lexical items in modern NLP applications. Therefore, we provide word embedding models for the vocabulary of this resource (all bases and collocates for both languages) obtained using the skip-gram algorithm (Mikolov et al., 2013).

The main idea is to enable further research in NLP and computational lexicography by providing distributional semantic models for individual collocation elements. We anticipate that this can be useful, for example, for improving word-level representations based on how well they capture their relational properties. This could be done, for example, by predicting relation (pairwise) vectors from two individual word embeddings, similarly as it was proposed by Camacho-Collados et al. (2019b).

3.2.3 Collocation relation vectors

Intuitively, a natural representation of a collocation could be the result of a vector composition operation which is applied to the word embeddings of its base and its collocate. Such a vector composition operation is typical for modeling semantic relations in the distributional semantics literature, where well-known operations are vector difference (Mikolov et al., 2013; Vylomova et al., 2015) and concatenation (Roller et al., 2014), and which have been investigated for capturing, among others, hypernymy and

[9] Verbal collocates and their subcategorization restrictions are often referred to as "phrasal verbs".

meronymy. More formally, let us assume w_b and w_c are the two words forming a collocation, i.e., a base and a collocate, and $\mathbf{v}_b$ and $\mathbf{v}_c$ their corresponding vector representations for some predefined word embedding model. Then, their composition can be given either by their average $\frac{\mathbf{v}_b + \mathbf{v}_c}{2}$, component-wise multiplication $\mathbf{v}_b \odot \mathbf{v}_c$, or vector difference $\mathbf{v}_b - \mathbf{v}_c$, among others.

However, it is unclear if such a composition would capture the idiosyncratic properties of collocations. For example, a conflated vector for *heavy* will not account for its different meanings if paired with different head nouns (e.g., *rain*, *metal* or *table*). More importantly, the idiosyncratic (collocational) relation between *heavy* and *rain* (as opposed to the other examples) is not captured in models based on co-ocurrence statistics, and explicit encodings seem necessary to complement the semantic properties of the individual vectors. This phenomenon is discussed by Espinosa-Anke et al. (2019), who show that representing a collocation's context in a dedicated vector space is more desirable (and leads to better results in the relation classification task) than simply operating with individual word vectors.

Based on these findings, we construct a *relation vector* model where each collocation is represented as a dedicated vector. Representing pairs of words is bound to become a popular problem in general, as joint embeddings can complement word representations and make them more powerful in downstream tasks such as lexical semantics modeling, text categorization or textual inference (Joshi et al., 2018; Camacho-Collados et al., 2019a). Our relation vector model of choice is SeVeN (Espinosa-Anke and Schockaert, 2018), where each collocation is represented as a vector $\mathbf{r}_{bc}$ condensing left, middle and right contexts of sentences in which its base and its collocate occur (also in reversed order). This is achieved by averaging their corresponding word embeddings. Specifically, given a sentence s and some context $\mathcal{C}$, we compute

$$\mathcal{C}^s_{v_b v_c} = \frac{1}{k} \sum_{r=1}^{k} \mathbf{v}_{a_r} \tag{1}$$

where a is a word appearing in a sentence in which w_b and w_c are mentioned within a predefined window and k is the number of words in that context ($\mathcal{C}$). We consider six different contexts: 'before w_b' (*pre*), 'between w_b and w_c' (*mid*), and 'after w_c' (*post*) for the occurrence w_b + w_c and 'before w_c' (*pre**), 'between w_c and w_b' (*mid**), and 'after w_b' (*post**) for the reverse occurrence w_c + w_b. Thus, we obtain $\mathbf{r}_{bc} \in \mathbb{R}^{6d}$, where d is the dimensionality of the pre-trained word vectors. We then average $\mathcal{C}$ over the set S_{bc} of all sentences mentioning w_b and w_c:

$$\mathcal{C}_{w_b w_c} = \frac{1}{|S_{bc}|} \sum_{s \in S_{bc}} \mathcal{C}^s_{w_b w_c} \tag{2}$$

Finally, the collocation vector $\mathbf{r}_{bc}$ is given by:

$$\mathbf{r}_{bc} = \mathcal{C}^{pre}_{w_b w_c} \oplus \mathcal{C}^{mid}_{w_b w_c} \oplus \mathcal{C}^{post}_{w_b w_c} \oplus \mathcal{C}^{pre*}_{w_b w_c} \oplus \mathcal{C}^{mid*}_{w_b w_c} \oplus \mathcal{C}^{post*}_{w_b w_c} \tag{3}$$

However, simply weighted averages based on frequencies, as is the case in this model, may ignore the fact that some words contribute more to the relation. For example, in the case of **Magn** we are interested in modeling the notion of intensity, and assuming this is something that can be captured from corpora, not all co-occurring words provide this information equally. Therefore, we apply a *conditional autoencoder* that serves two purposes: (1) dimensionality reduction; and (2) purification of the relation vector, putting less weight on words relevant to the meaning of base and collocate alone, and more on those that refer to the relation[10].

Space properties and size A comprehensive assessment of the intrinsic properties of this collocational embedding space is beyond the scope of this paper. We provide, however, a piece of analysis based on exploring semantic clusters. In Table 4, we list the nearest neighbours for selected target collocation vectors in both English and French. Note that the semantic clusters that emerge group collocations of the same lexical function nearby in the space (e.g., *tragic mistake* and *terrible tragedy* are both **Magn**

[10] We refer to the original SeVeN publication for details of the autoencoder architecture. We used the implementation available at `https://bitbucket.org/luisespinosa/seven`.

MAGN		REAL1		QSYN	
EN	FR	EN	FR	EN	FR
tragic-mistake	**réputation-solide**	**follow-line**	**balle-loger**	**irritate-annoy**	**pluie-ondée**
terrible-tragedy	faible-densité	smoke-pipe	cérémonie-tenir	efficiently-expeditiously	élections-tenir
great-achievement	riche-carrière	buy-store	différence-fair	brazen-brash	sueur-suer
bad-mistake	large-victoire	eat-restaurant	victoire-donner	scorn-disdain	geste-poser
wonderful-person	forte-hausse	aim-goal	lutte-poursuivre	uncouth-rude	fumée-dégager
great-honor	lourde-responsabilité	return-save	crise-désamorcer	evergreen-deciduous	vote-faire

Table 4: Nearest neighbours in English and French for selected relation vectors belonging to three lexical functions.

collocations). However, in those cases where this regularity is not preserved, obvious word-level semantics are prevalent, which suggests that these embeddings could be effectively exploited in downstream applications where either relational or word-level semantics or both are required. In terms of space size, we encode embeddings for all collocations for which sufficient examples have been encountered in the reference corpora. These are 5,844 collocations in English and 4,156 in French.

3.2.4 BabelNet senses

The bases of the collocations in our lists are assigned their BabelNet senses (Navigli and Ponzetto, 2012). This ensures, on the one hand, the disambiguation of the bases, and, on the other hand, the alignment of the bases across the English and French collocation lists. For instance, Eng. *reception* and Fr. *accueil* are both assigned bn:00066506n, Eng. *purchase* and Fr. *achat* bn:00065265n, Eng. *death* and Fr. *mort* bn:00100948a, etc. However, not all bases have a BabelNet id; cf., e.g., *shadow* in the sense of Fr. *ombrage* or *ascendant*.[11] In this case, no cross-language linkage is currently provided. Note that the current version of our resource does not contain the BabelNet senses of the collocates either.

3.2.5 References to the occurrences in the corpora

A great number of sentences in our reference corpora contain collocations; cf. Section 4. Consider five of them for illustration (subcategorization patterns are not highlighted):

(1) Fr. **AntiMagn**: *Une mince*$_{\text{Coll}}$ *chance*$_{\text{Base}}$ *de qualification existe encore ...*
(2) Eng. **AntiVer**: *The White House still continues its baseless*$_{\text{Coll}}$ *accusations*$_{\text{Base}}$ *against ...*
(3) Eng. **Oper1**: *The legendary Olympia Music Hall in Paris bid*$_{\text{Coll}}$ *adieu*Base *to French music ...*
(4) Fr. **Oper2**: *Il va se trouver*$_{\text{Coll}}$ *sous le feu*$_{\text{Base}}$ *de l'actualité pugilistique*
(5) Eng. **Real1**: *The plane crashed because of a problem with lowering*$_{\text{Coll}}$ *its landing gear*$_{\text{Base}}$ *or had ...*

In order to facilitate the illustration of the use of collocations in context and also to provide more targeted material for collocation-related model training, we assign to each collocation the indices of its occurrences in the reference corpora.

4 Corpus Preparation

The development of this resource consisted of a two-step procedure: first, corpus processing and indexing; and second, collocation matching and assembling the resource in the desired output format. In the first step, we apply a syntactic parser to obtain the Part-of-Speech (PoS) tags and dependency relation information for the sentences in the corpus. Once processed, the sentences are indexed in a Solr search engine. In the second step, each collocation is searched for in the index. PoS tags and dependencies are used to retrieve the sentences in which the base and the collocate co-occur distribution-wise and are related syntactically. For this purpose, first, a query is applied to search for a specific syntactic relation between the base and the collocate (e.g., verb-object or head-modifier). This query guarantees a high precision, but could result in a low recall for some collocations due to, e.g., parsing errors or difficulty to

[11] For English, out of the 3065 different bases, 27 do not have a BabelNet id; for French, 262 out of 3148 bases do not have it.

L.id	b	BN.id	bpos	c	LF	st.	end	q	s	sentence fragment
EN	wrong	bn:00104880a	A	terribly	Magn	5	6	4	1	he knew something was terribly wrong .
EN	wreath	bn:00017726n	N	lay	CausFunc2	4	5	4	1	prince charles to lay wreath at graves
EN	wrath	bn:00081680n	N	incur	Oper2	5	7	4	1	the foundation has also incurred the wrath of many in the exile ...
EN	_cold feet_	bn:00020546n	N	get	IncepOper1	9	11	1	1	but in the white house , some are getting cold feet
EN	_turn out_	bn:00085376v	V	well	Bon	3	5	1	1	when things turned out well , they walked away with huge bonuses
…	…	…	…	…	…	…	…	…	…	…

Table 5: The format of the codification of the English and French corpora we release as part of CollFrEn.

determine the head of a multiword base or collocate. To increase the recall, a second query is performed searching for a sequence that combines the lemmas and POS tags of the base and collocate. If both queries do not retrieve any results, a third and fourth queries may also be applied where the conditions are further relaxed (e.g., in terms of the search for words and search for the base and collocate lexical items at a maximum distance of six tokens). It is obvious that the chance to retrieve co-occurrences that do not form a collocation increases as we relax the conditions. In order to ensure the transparency in this respect, each of the retrieved sentences indicates the most restrictive query that was used to obtain it, such that it is straightforward to retain only high precision sentences.

The English corpus is composed of 9,272,395 sentences. 83.1% of it have been obtained with the first query, 16% with the second one, and less than 1% applying more relaxed queries. The French corpus is composed of 3,474,134 sentences, with 86% being retrieved with the first query, 11.7% with the second, 2.1% with the third, and 0.1% with more relaxed queries. Further details about the format of the resulting corpora (which is the same in both English and French) can be found in Table 5.

5 Conclusions

We presented a bilingual English–French collocation resource in which the collocations are tagged with respect to lexical functions and enriched by information that is commonly used by state-of-the-art NLP applications. This information concerns, in particular, subcategorization patterns, embeddings of the collocation elements and embeddings of the collocation relations and indices of the collocation occurrences in the reference corpora. For disambiguation and interlinking of the bases in the English and French collocation lists, we use BabelNet senses.

The presented resource can be used either as an input to NLP applications or as an online collocation dictionary. In this latter interpretation it resembles the online collocation dictionary DiCE of Spanish `http://www.dicesp.com/`. As our resource, DiCE classifies collocations in terms of lexical functions, provides the subcategorization information of both the base and the collocate and cites examples of the use of each collocation extracted from a large scale corpus. However, while DiCE includes only some selected examples of the occurrence of each collocation, in our resource, each collocation is indexed with all of its occurrences in the corpus. Furthermore, DiCE does not contain any embedding-related information on the collocation elements or the collocational relations. The resource is available at `https://github.com/TalnUPF/CollFrEn`.

As part of future work, we plan to align the French and English collocation equivalents (not only the bases), completing the lists when no equivalent is available in the present list. Furthermore, we plan to automatically extend the resource using state-of-the-art collocation extraction and semantic classification techniques, also for other languages than English and French. In this context, the resources created in the PARSEME Cost Action `https://typo.uni-konstanz.de/parseme/` will be also explored.

Acknowledgements

Many thanks to Igor Mel'čuk for making the original lists of the English and French LF instances available to us. We would also like to thank the three anonymous reviewers whose comments helped to significantly improve the paper. This work has been supported by the European Commission in the context of its H2020 Program under the contract numbers 870930-RIA, 825079-STARTS, and 779962-RIA.

References

Margarita Alonso Ramos, Owen Rambow, and Leo Wanner. 2008. Using semantically annotated corpora to build collocation resources. In *Proceedings of LREC*, pages 1154–1158, Marrakesh, Morocco.

Margarita Alonso Ramos, Leo Wanner, Orsolya Vincze, Gerard Casamayor, Nancy Vázquez, Estela Mosqueira, and Sabela 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.

Margarita Alonso Ramos, Roberto Carlini, Joan Codina-Filbà, Ana Orol, Orsolya Vincze, and Leo Wanner. 2015. Towards a learner need-oriented second language writing assistant. In *Proceedings of the European Conference on Computer Assisted Language Learning (CALL)*, Padova, Italy.

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

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet ORFÉO : un corpus d'étude pour le français contemporain. *Corpus*, 15.

Jose Camacho-Collados, Luis Espinosa-Anke, Shoaib Jameel, and Steven Schockaert. 2019a. A latent variable model for learning distributional relation vectors. In *International Joint Conferences on Artificial Intelligence*.

Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2019b. Relational word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3286–3296, Florence, Italy, July. Association for Computational Linguistics.

Y.C. Chang, J.S. Chang, H.J. Chen, and H.C. Liou. 2008. An Automatic Collocation Writing Assistant for Taiwanese EFL learners. A case of Corpus Based NLP Technology. *Computer Assisted Language Learning*, 21(3):283–299.

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

Luis Espinosa-Anke and Steven Schockaert. 2018. Seven: Augmenting word embeddings with unsupervised relation vectors. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2653–2665.

Luis Espinosa-Anke, José Camacho-Collados, Sara Rodríguez-Fernández, Horacio Saggion, and Leo Wanner. 2016. Extending wordnet with fine-grained collocational information via supervised distributional learning. In *Proceedings of COLING*, pages 3422–3432. ACL.

Luis Espinosa-Anke, Steven Schockaert, and Leo Wanner. 2019. Collocation classification with unsupervised relation vectors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5765–5772.

M. Garcia, M. Garcia-Salido, S. Sotelo, E. Mosqueira, and M. Alonso-Ramos. 2019. Pay attention when you pay the bills. a multilingual corpus with dependency-based and semantic annotation of collocations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4012–4019, Florence, Italy.

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

Sylviane 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.

Mandar Joshi, Eunsol Choi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2018. pair2vec: Compositional word-pair embeddings for cross-sentence inference. *arXiv preprint arXiv:1810.08854*.

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

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

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

A. Li-E. Liu, D. Wible, and N.-L. Tsao. 2009. Automated suggestions for miscollocations. In *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 47–50, Boulder, CO.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. Syntagnet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of EMNLP*, pages 3525–3530. ACL.

Igor Mel'čuk and Alain Polguère. 2007. *Lexique actif du français.* De Boeck Supérieur, Brussels.

Igor Mel'čuk and Alexander Žolkovskij. 1984. *Explanatory Combinatorial Dictionary of Modern Russian.* Wiener Slawistischer Almanach, Vienna.

Igor Mel'čuk *et al.* 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain, Volumes I–IV.* Presses de l'Université de Montréal, Montreal.

Igor Mel'čuk. 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.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Pol Moreno, Gabriela Ferraro, and Leo 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.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

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

Alain Polguère. 2014. From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, 27(4):396–418.

Sara Rodríguez Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug. 7-12; Berlin (Germany).[place unknown]: ACL; 2016. Vol. 2, Short Papers; p. 499-505.* ACL (Association for Computational Linguistics).

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1025–1036.

V. Seretan. 2013. On collocations and their interaction with parsing and translation. *Informatics*, 1(1):11–31.

F. Smadja and K.R. McKeown. 1991. Using collocations for language generation. *Computational Intelligence*, 7(4):229–239.

Yuancheng Tu and Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, pages 31–39. Association for Computational Linguistics.

Yuancheng Tu and Dan Roth. 2012. Sorting out the most confusing english phrasal verbs. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 65–69. Association for Computational Linguistics.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit.* Ph.D. thesis, Université de Toulouse II le Mirail.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2015. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. *arXiv preprint arXiv:1509.01692*.

Leo Wanner and John Bateman. 1990. A collocational based approach to salience-sensitive lexical selection. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, Dawson, PA.

Leo Wanner, Margarita Alonso Ramos, and Antonia Martí. 2004. Enriching the Eurowordnet by Collocations. In *Proceedings of LREC*, Lisbon. ELDA.

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

Leo Wanner, Gabriela Ferraro, and Pol Moreno. 2016. Towards distributional semantics-based classification of collocations for collocation dictionaries. *International Journal of Lexicography, doi:10.1093/ijl/ecw002*.

Leo Wanner. 2004. Towards automatic fine-grained semantic classification of verb-noun collocations. *Natural Language Engineering Journal*, 10(2):95–143.