Corpus-based Methodology for an Online Multilingual Collocations Dictionary: First Steps

Adriane Orenha-Ottaiano[1], Marcos Garcia [2], Maria Eugênia Olímpio de Oliveira Silva[3], Marie-Claude L'Homme[4], Margarita Alonso Ramos[5], Carlos Roberto Valêncio[6], William Tenório[7]

[1] São Paulo State University (UNESP), Brazil
[2] Universidade de Santiago de Compostela, Galiza, Spain
[3] University of Alcalá, Spain
[4] OLST, Université de Montréal, Québec, Canada
[5] Universidade da Coruña, Spain
[6] São Paulo State University (UNESP), Brazil
[7] São Paulo State University (UNESP), Brazil
E-mail: adriane.ottaiano@unesp.br, marcos.garcia.gonzalez@usc.gal, eugenia.olimpio@uah.es, mc.lhomme@umontreal.ca.ca, margarita.alonso@udc.es, carlos.valencio@unesp.br, williamtenoriotenorio@gmail.com

Abstract

This paper describes the first steps of a corpus-based methodology for the development of an online Platform for Multilingual Collocations Dictionaries (PLATCOL). The platform is aimed to be customized for different target audiences according to their needs. It covers various syntactic structures of collocations that fit into the following taxonomy: verbal, adjectival, nominal, and adverbial. Part of its design, layout and methodological procedures are based on the *Bilingual Online Collocations Dictionary Platform* (Orenha-Ottaiano, 2017). The methodology also relies on the combination of automatic methods to extract candidate collocations (Garcia et al., 2019a) with careful post-editing performed by lexicographers. The automatic approaches take advantage of NLP tools to annotate large corpora with lemmas, PoS-tags and dependency relations in five languages (English, French, Portuguese, Spanish and Chinese). Using these data, we apply statistical measures (Evert et al., 2017; Garcia et al., 2019b) and distributional semantics strategies to select the candidates (Garcia et al., 2019c) and retrieve corpus-based examples (Kilgarriff et al., 2008). We also rely on automatic definition extraction (Bond & Foster, 2013) so that collocations can be more effectively organized according to their specific senses.

Keywords: collocations; collocations dictionary; online platform; automatic extraction; lexicography

1. Introduction

In the past two decades, collocations have been high on the agenda of foreign language teaching and learning (Nesselhauf, 2005; Alonso-Ramos, 2008, 2019; Laufer, 2011; Orenha-Ottaiano, 2021; Torner & Bernal, 2017, among others). Despite this fact, when it comes to the translation of collocations, the number of studies that can contribute to better comprehension of the difficulties regarding the complexity of translation of such combinations is not as significant (Kenny, 2001; Bernardini, 2007; Gregorio-Godeo & Molina, 2011; Orenha-Ottaiano, 2009, 2012, forthcoming).

Additionally, even though several authors emphasise the importance of compiling dictionaries with a special focus on collocations or for the building of specific collocations dictionaries (Alonso-Ramos, 2001; Atkins & Rundell, 2008; Moon, 2008; Orenha-Ottaiano, 2013, 2015, 2017; Kilgarriff, 2015, etc.), the number of online or electronic collocations dictionaries available is still scarce, especially when it comes to bilingual or multilingual collocations dictionaries for general language.

The work described in this paper aims to fill this gap. We describe a methodology for the design and compilation of an online platform for multilingual collocations dictionaries (English, Portuguese, French, Spanish and Chinese). The collection of relevant collocations is corpus-based and semi-automated (automatic extraction with human validation). Furthermore, the design of the platform takes into consideration users' needs as suggested by the principles of the function theory of lexicography (Bothma & Tarp, 2012; Fuertes-Olivera & Tarp, 2014; Tarp, 2015).

Besides the introduction, the paper is structured as follows. Section 2 addresses the motivational aspects for the development of a corpus-based methodology of multilingual collocations dictionaries and an online platform. Section 3 outlines the methodological steps used in this research. Section 4 explores the Multilingual Collocations Dictionary's structure and design. Finally, Section 5 presents the concluding remarks and highlights some ideas for further work.

2. Motivation

One of the main motivations for carrying out this research is that collocations require specific pedagogical attention. Concerning lexicographical work, excellent monolingual collocations dictionaries for learners of English as a second or foreign language are available, such as the *Longman Collocations Dictionary and Thesaurus* (2013), *Macmillan Collocations Dictionary for Learners of English* (Rundell, 2010), *Oxford Collocations Dictionary for Students of English* (Mcintosh et al., 2009), *LTP Dictionary of Selected Collocations* (Hill; Lewis, 1999) and *The BBI Combinatory Dictionary of English* (Benson et al., 1997), with the last two are only available in paper format.

In Portuguese, to the best of our knowledge, the only online and corpus-based dictionary of collocations is the one developed by Orenha-Ottaiano (2017). As it is bi-directional, and users can consult it both as a monolingual (either Portuguese or English) or as a bilingual (English-Portuguese and Portuguese-English).

In Spanish, the *Diccionario combinatorio práctico del español contemporáneo* (Bosque, 2006) is a corpus-based dictionary for native or foreign language speakers of Spanish, which focuses not only on collocations but also on other phraseologisms, such as idioms (*locuciones fijas*). The *Diccionario de colocaciones del español* (*DiCE*; Alonso-Ramos, 2004) is available online and encodes collocations according to the principles of the Meaning-Text Theory.

In French, Beauchesne's the *Dictionnaire des Cooccurrences* (2001) is an example of a printed and online monolingual collocations dictionary, but it is not corpus-based. The *DiCouèbe* (Jousse & Polguère, 2005) is an online French combinatorial dictionary in which collocations are all encoded with Lexical Functions.

In Chinese, we can mention the *Modern Chinese Collocation Dictionary* (Mei, 1999) and *Dictionary of Chinese Common Word Collocations* (Yang, 1990).

As far as bilingual dictionaries are concerned, as previously mentioned, Orenha-Ottaiano (2016, 2017) built an online platform of bilingual Collocations Dictionary (English-Portuguese and Portuguese-English), which has recently been changed into a platform of multilingual collocations dictionaries, as discussed in this paper. Alegro et al. (2010) published a printed dictionary containing 3,000 adjectival collocations (Portuguese-English), but it is neither corpus-based nor in an electronic or online format.

The *DiCoEnviro* (L'Homme et al., 2018) and the *DiCoInfo* (L'Homme, 2008) are online terminological dictionaries in English, French and Spanish (a few Portuguese, Italian and Chinese terms are also listed) that focus on specialized terms, encodes specialized collocations and explain the meaning of collocates using the system of lexical functions (Mel'čuk, 1996).

Finally, another bilingual dictionary worth mentioning is *The Oxford Collocations Dictionary* (English-Chinese), both printed and app versions.

A lot of research has taken place on corpus-based and online bilingual or multilingual collocations dictionaries in other languages, such as the Dictionary of Collocations of European Portuguese (Pereira & Mendes, 2002), a dictionary of Italian collocations (Spina, 2010), an investigation on the automatic construction of a multilingual dictionary of collocations (Garcia et al., 2019a), and a bilingual English-Italian dictionary of collocations (Berti & Pinnavaia, 2014), among others. Nevertheless, there is still a gap in the availability or publication of online dictionaries themselves as they are research proposals and have not been published yet.

Another motivational aspect of this project concerns the possibility of developing a platform offering a higher degree of customisation of the structure of the dictionaries. It aims at the development of an innovative lexicographical methodology and model for a multilingual collocations dictionary, as well as the design of a collocations software and platform, the PLATCOL[1]. Moreover, it targets the setting up of a useful and large

[1] The Platform for Multilingual Collocations Dictionaries (PLATCOL) is the practical result of the project *A phraseographical methodology and model for an online corpus-based Multilingual Collocations Dictionary Platform*, sponsored by The São Paulo Research Foundation (FAPESP). It is a two-year project with a partnership between São Paulo State University (Brazil), responsible for English and Portuguese languages, the University of Montréal (French), University of Granada (Chinese), University of Coruña and University of Alcalá (Spanish), and University of Santiago de Compostela, for the automatic retrieval of corpus data.

resource for semi-automatic collocations retrieval, as well as automatic extraction of good examples, definitions and translation.

3. Methodology

The methodology to build the dictionary is based on the automatic approach described in Garcia *et al.* (2019a), enriched with sense information of the bases and a manual review and validation of the extracted data made by lexicographers.

3.1 Corpora

We compiled a large corpus for each of the five languages of the project using different source data, as Table 1 below shows:

Language	Sources	Size (tokens)
Portuguese	Jornal do Brasil, Wikipedia/Wikibooks, Paracrawl, CHAVE (Santos & Rocha, 2004), CBras, BrWaC (Wagner Filho et al., 2018)	4B
Spanish	EuroParl (Kohen, 2005), Literature (short stories/romances) (Garcia et al., 2019a), Wikipedia/Wikibooks	1,2B
English	EuroParl, Wikipedia/Wikibooks	1.6B
French	FrWaC (Baroni et al., 2009), Wikipedia/Wikibooks	2.5 B
Chinese	Wikipedia, Wikibooks, and literary texts	600M

Table 1: Corpora Size and Sources

The corpora were parsed with UDPipe (Straka & Straková, 2017) using the latest models (v2.7) trained on the UD corpora (de Marneffe *et al.*, 2021). Previous to this syntactic analysis, we tokenized and PoS-tagged the data using the same UDPipe models for English and French, LinguaKit (Gamallo *et al.*, 2018) for Portuguese and Spanish, and the Stanford CoreNLP suite (Manning *et al.*, 2014) for the Chinese texts.

3.2 Definition and extraction of keywords

We focus on collocation types with three morphosyntactic classes of bases: nouns, verbs, and adjectives. Due to the large size of the corpora, we attempt to extract basic vocabulary lists for each class and language. Therefore, we automatically extracted the lemmas of the nouns with a minimum frequency of one occurrence per million tokens in each corpus, annotating them as *known* or *unknown* if they appear in large lexica[2]. We used the dictionaries provided by FreeLing (Padró & Stanilovsky, 2012) for each language (English, Portuguese, French and Spanish), except for Chinese. We didn't use any lexicon for Chinese because we are not aware of any free dictionary for this language.

[2] Due to the lower frequency of verbs and adjectives, we used frequency=>0.5 in these cases.

After the automatic extraction, which took place for each language separately, the lists of keywords were submitted to the lexicographers to filter out noise (e.g., lemmas with typos, entries wrongly processed, etc.) and to select the most frequent lemmas, then used to extract candidate collocations. Besides, each keyword has been enriched with the potential senses present in WordNet, using the Open Multilingual WordNet (Bond & Foster, 2013) by means of the interface provided by the NLTK package (Bird & Klein, 2009).

Table 2 shows a sample of keywords in French as an example, sorted by descending order of frequency. Candidates marked NO by lexicographers were removed from the list.

Base-candidate	Frequency	Frequency per million	Validation
adulte	89630	34.028372142183656	OK
chasse	89494	33.97673922227585	OK
instance	89227	33.87537165157449	OK
pêche	89163	33.85107380691199	OK
administrateur	89149	33.84575865339207	OK
qu	**89146**	**33.84461969192351**	**NO**
orbite	89097	33.82601665460379	OK
session	89026	33.79906123318133	OK
précision	89017	33.79564434877567	OK
tension	88916	33.75729931266766	OK
litre	88904	33.75274346679344	OK
entraîneur	88696	33.67377547164032	OK
parlement	88579	33.62935597436669	OK
canal	88443	33.57772305445888	OK
leader	88393	33.5587403633163	OK
vocation	88308	33.52646978837392	OK
appartement	88193	33.482809598745995	OK
copie	88114	33.452816946740725	OK

Table 2: Results of validation in French

After having manually validated the base candidates in each language separately, we reached the following results for English, French and Portuguese, shown in Table 3.

	Automatically extracted candidates			Validated candidates		
	French	**Portuguese**	**English**	**French**	**Portuguese**	**English**
Nouns	9,754	10,307	10,545	8,361	8,690	8,713
Verbs	4,895	5,573	5,502	2,902	3,817	3,982

Table 3: Number of automatically extracted and validated candidates

As can be noted in Table 3, about 15% of nouns were discarded in French, 16% in Portuguese, and 18% in English. As for the verbs, 40% of them were discarded in French, 32% in Portuguese, and 28% in English. These results highlight the importance of post-editing in all lexicographical phases.

3.3 Identification of collocations and example sentences

Following Garcia *et al.* (2017) we extract pairs of the target dependency relations using the manually validated keywords and restricting the potential collocates for their morphosyntactic category. Thus, for noun bases we extract the following syntactic relations:[3] *obj* (verb-noun collocations), *nsubj* (instances of noun-verb), *obl* (verb-preposition-noun), *amod* (adjective-noun), and *nmod* and *compound* (both including noun-noun or noun-prep-noun instances). For verb bases we extract *xcomp* (verb-adjective collocations) and *advmod* (verb-adverb). Finally, for adjective bases, we extract *advmod* examples (adjective-adverb candidates).

For each triple (base;collocate;relation) we follow the syntactic co-occurrence method described in Evert (2008) to compute, apart from frequency data, the following statistical values: PMI, Dice, log-likelihood, t-score, z-score, ^{2}, and simple-ll (together with ΔP (Gries, 2013). In order to reduce the large size of the candidates sets we remove those combinations with a normalized frequency lower than one per million, and sort the remaining ones by t-score (Garcia *et al.*, 2019b).

Then, we collect up to eight sentences for each candidate collocation, selected by a set of GDEX-inspired heuristics (Kilgarriff et al., 2008). We have implemented a basic strategy using some of the proposals of Kosem et al. (2019a) for English and for Portuguese (the latter were also used for the other romance languages): sentences with less than six tokens are discarded, and those with more than 30 tokens are incrementally penalized. Furthermore, sentences with punctuation, proper nouns, words with more

[3] https://universaldependencies.org/u/dep/all.html

than 12 characters, and strange characters (e.g., in other alphabets and encodings) are also penalized. Other heuristics in the literature were not implemented as they require language-specific resources or are computationally very expensive.

This automatically extracted information is then used by language experts to select the collocations for the final resource. For each candidate, the lexicographers decide which combinations are going to be incorporated into the dictionary, and select the appropriate sense for the base and a set of five examples to be shown on the platform. The tables below show examples of automatically retrieved data in English (Tables 4 and 5) and in Portuguese (Tables 6 and 7) from *noun* bases, showing collocates, frequencies, some of the statistical score results and examples (four out of eight) – the first example has collocations highlighted manually.

base	collocate	deprel	freq base	freq collocate	freq	freq norm	MI	di	ll	ts	zs
bond	double	amod	18052	60424	1871	33.24	64,069,650,805,298	393,761,171,440,127	427,453,180,843,812	0.092838	682,620,358,777,046
interval	time	compound	7546	258128	1334	34.10	450,163,454,373,253	166,162,814,372,741	349,116,460,482,685	0.143626	32,223,953,217,268
language	programming	compound	141852	36647	457	11.68	175,969,516,098,485	277,214,778,124,125	150,645,396,615,149	0.002271	242,231,087,565,422
bond	single	amod	18052	299877	588	10.45	256,306,795,734,007	489,730,681,878,011	201,454,580,106,199	0.026216	580,005,558,729,172
compound	organic	amod	28611	25825	3395	60.31	767,481,981,008,368	828,828,726,010,929	579,814,814,566,719	0.105615	159,191,454,287,767
group	functional	amod	309637	19037	1867	33.17	401,247,091,282,802	162,828,195,759,869	405,314,891,160,725	0.005653	370,591,810,572,147
file	media	compound	53420	69899	516	13.19	241,027,046,499,507	425,204,881,062,829	184,423,554,202,936	0.007778	453,193,910,956,383
role	play	obj	228878	453350	99651	3023.61	417,611,948,069,146	126,793,007,163,871	298,213,079,557,869	0.289431	283,512,452,652,812
question	answer	obj	74554	39164	15712	476.73	670,790,303,510,026	126,934,999,895,782	124,148,471,289,531	0.172872	711,565,652,867,556

Table 4: Automatically retrieved data from the English corpus – base = noun

base	collocate	example 1	example 2	example 3	example 4
bond	double	This reaction can be used to determine the position of a **double bond** in an unknown alkene.	This makes the Br closest to the double bond slightly positive and therefore an electrophile.	The IUPAC numerical prefixes are used to indicate the number of double bonds.	That is, hydrogen ends up on the more substituted carbon of the double bond.
interval	time	Whoever measures a particular space-**time interval** will get the same value, no matter how fast they are travelling.	The space-time interval, formula_19, is invariant.	The second consequence of the invariance of the space-time interval is that clocks will appear to go slower on objects that are moving relative to you.	Solutions Consider the formula for average velocity in the formula_1 direction, formula_49 , where formula_18 is the change in formula_1 over the time interval formula_52 .
language	programming	Since computer **programming languages** have so much in common, it is generally easy to learn a new programming language once you have mastered another.	Pascal is an influential computer programming language named after the mathematician .	Many people think they must choose a specific programming language in order to become a programmer, believing that they can only do that language.	D is a programming language being designed as a successor to C++.
bond	single	What of having two double bonds separated by a **single bond**?	What of having a compound that alternates between double bond and single bond?	Remember that single bonds can rotate in space if not impeded.	Each ending point and bend in the line represents one carbon atom and each short line represents one single carbon-carbon bond.
compound	organic	This number also applies to other **organic compounds** which have hydrogen atoms at similar distances from each other.	The IUPAC system is necessary for complicated organic compounds.	Hydrocarbons are organic compounds that contain carbon and hydrogen only.	Heterotrophs require at least one organic nutrient to make other organic compounds.
group	functional	These parts of organic molecules are called **functional groups**.	There are many functional groups of interest to organic chemists.	The identification of functional groups and the ability to predict reactivity based on functional group properties is one of the cornerstones of organic chemistry.	Just as elements have distinctive properties, functional groups have characteristic chemistries.
file	media	Where not otherwise noted, non-text **media files** are available under various free culture licenses, consistent with the .	Typically the in using an image or other media file is to it to .	See for details about which media files can be uploaded.	Please view the media description page for details about the license of any specific media file.
role	play	Wave packets will **play** a central **role** in what is to follow, so it is important that we acquire a good understanding of them.	Usually hydrogen plays the role of the electrophile; however, hydrogen can also act as an nucleophile in some reactions.	The ideas of bond polarity and dipole moment play important roles in organic chemistry.	Smooth ER plays an important role in lipid emulsification and digestion in the cell.
question	answer	You should now be able to **answer** the following **questions** from your previous knowledge.	Use the content in this chapter and/or from external sources to answer the following questions.	This is the Reading room where raise and answer Wikibooks-related questions and concerns regarding technical issues, policies, or other aspects of our community.	If you answered the above questions correctly you should find this next section easy!

Table 5: Automatically retrieved data from the English corpus - examples

base	collocate	deprel	freq base	freq collocate	freq	freq norm	mi	dl	ll	ts	zs
direito	ter	obj	560822	11188299	183090	1306.77	160,048,855,495,172	0.165837224384871	160,804,071,456,546	499,409,455,757,521	0.030225
contato	entrar	obl	139840	1081619	84533	1107.33	462,428,604,203,138	0.362541768192905	379,698,081,324,801	138,535,656,820,408	0.121584
rede	social	amod	451341	1510216	176610	1182.83	463,777,831,408,155	0.271090759214873	796,451,699,849,827	201,260,093,158,997	0.152594
atenção	chamar	obj	412559	374632	171188	1221.82	623,414,529,184,781	0.290572130437827	114,164,059,303,372	354,210,236,916,932	0.303104
diferença	fazer	obj	260225	5263149	79253	565.65	261,413,494,243,881	0.195799523645682	154,592,588,917,987	582,811,846,552,285	0.027897
quantidade	grande	amod	213140	5140962	84160	563.66	301,599,685,612,871	0.24858108135078	204,365,955,457,986	723,098,185,697,314	0.030479
acesso	ter	obj	309489	11188299	147993	1056.27	199,934,115,026,427	0.24337875729177	188,232,852,359,215	576,828,284,897,299	0.025097
destaque	grande	amod	59884	5140962	24434	163.65	306,634,489,037,465	0.255333171494298	608,313,819,034,328	398,396,784,243,363	0.009309
direito	humano	amod	640881	739743	95389	638.86	453,375,817,538,921	0.124578155368098	416,989,915,076,945	142,226,851,904,013	0.121406

Table 6: Automatically retrieved data from the Portuguese corpus – base = noun

base	collocate	example 1	example 2	example 3	example 4
direito	ter	Vc desconhece completamente que vc e a idiota que aceitou ter um filho seu **tem** os mesmos **direitos** e obrigações perante a prole comum .	Todo e qualquer trabalhador tem o direito , inclusive os que atuam em cargo de confiança .	Tenho o direito de a gurda de a minha filha ?	Portanto , até que isso aconteça , nenhum servidor tem direito adquirido a a nova regra .
contato	entrar	br, é necessário **entrar** em **contato** com o Registro .	apenas inquéritos sérios devem entrar em contato para obter mais detalhes	Ganhei a sentença , e o banco ainda não entrou em contato .	Entraram em contato com a médica e realizei o exame .
rede	social	Monte a sua ou a de alguém conhecido , ou crie um personagem novo para compartilhar em as **redes sociais** .	Em a página de o evento em a rede social , os organizadores explicam quais são suas reivindicações .	Depois de um bocado de relutância , me rendi a os encantos de a rede social azul .	De acordo com analistas norte-americanos , o resultado , um empate virtual , teve influência direta de o grande movimento em as redes sociais .
atenção	chamar	**chamar** a **atenção** a uma situação cotidiana e simples , mas imperceptível por os juristas .	já chamava atençao por o estilo hippie-chique com que se vestia .	em maio o material começou a chamar a atenção de os grandes portais .	Essa atitude chamou a atenção de os políticos de o Piaui , que reivindicaram esse território .
diferença	fazer	Temos certeza de que somente este detalhe **fará** toda a **diferença** por o astral de o seu cantinho .	Os acessórios de cozinha são os detalhes que fazem a diferença .	Enfim , pesquisar , estudar qual sua atuação pode fazer muita diferença .	Em o dia a dia elas fazem toda a diferença !
quantidade	grande	muitas influências negativas sobre a saúde , como o cigarro, inatividade física e **grandes quantidades** de gordura corporal também estão associadas a a riqueza de o Ocidente .	existe uma grande quantidade de nomes para o segundo nível , mas .	Mas o grande destaque foi a grande quantidade de palestras de conscientização que atingiu públicos de todas idades , tanto homens quanto mulheres .	Destaque para as pérolas com acabamento ABS , que possuem uma maior quantidade de camadas de banho , tornando- se mais resistentes .
acesso	ter	Para quem quiser experimentar ainda mais , existe a possibilidade de comprar um Passaporte para **ter acesso** a atividades extras --	Os usuários passaram a ter livre acesso a o acervo .	Em aquela altura , os imigrantes ainda não tinham grande acesso a a terra .	Assim , temos acesso exclusivo a softwares , certificações e outros serviços .
destaque	grande	Seu estilo foi classificado como arte naïf e teve **grande destaque** até a década de 1980 .	Mas o grande destaque foi a grande quantidade de palestras de conscientização que atingiu públicos de todas idades , tanto homens quanto mulheres .	Escolha peças que auxiliem em um destaque maior de os seus arranjos e de o cômodo .	Em o final de o século XX , o pintor Leonilson foi o maior destaque cearense em a pintura .
direito	humano	violação de os **direitos humanos** , desvio de dinheiro público , corrupção ativa e passiva , etc .	e possíveis casos de violações de direitos humanos quando de as negociações individuais .	Sua linha política , entretanto , está mais voltada para o assistencialismo do que para a defesa de os direitos humanos de as pessoas trans ou homossexuais .	Parece fora de dúvidas que o Brasil evoluiu bastante em a adoção de mecanismos para proteção de os direitos humanos .

Table 7: Automatically retrieved data from the Portuguese corpus - examples

The volume of the automatically retrieved data is very large. We set a filter of 20 occurrences per million, in the same syntactic dependence, following Evert (2008). This filter has given, on average, 20,000 candidates with base = name, and 8,000 with base = verb, for example. The post-editing phase is still in progress and may last a few months as data have been manually validated, evaluated and also revised by at least two lexicographers. As collocations are being revised, they are directed to the following phase of automatic translation into other languages, as described in the next section, according to the pairs we have previously set (please see subsection *4.3*)

3.4 Translation of collocations

Once the monolingual collocations are inserted in the platform, we will use an unsupervised approach to retrieve candidate translations among the languages of the project. The strategy, inspired by Garcia et al. (2019c), can be summarized as follows:

We first train monolingual *word2vec* models (Mikolov et al., 2013) using processed corpora and representing each word as a pair of lemma and PoS-tag (e.g., "house_NOUN"). Then, these models are mapped in a shared vector space with *vecmap* (Artetxe et al., 2018). Finally, we create a compositional vector for a given collocation in language A, and search for similar candidates (in terms of cosine similarity) in language B (Garcia et al., 2019c). The candidate translations are ranked by the confidence of the models, and they will be manually validated by lexicographers in further work.

4. The Multilingual Collocations Dictionary Structure and Design

The Multilingual Collocations Dictionaries[4] (PLATCOL) proposed here aim at fulfilling users' needs regarding language encoding, and, as such, are considered to be a production dictionary. Besides helping users produce more authentic texts, PLATCOL also has the purpose of developing users' collocational competence, which is intrinsically connected with fluency. The wider the repertoire of collocations, the greater fluency a learner can achieve. Moreover, the platform is intended to have an easy-to-use layout that offers the possibility of being customized.

Since foreign language learners or dictionary users in general encounter challenges in using collocations in their native language, and PLATCOL is also designed to display monolingual dictionaries. Thus, it will serve as a monolingual, bilingual or multilingual dictionary (English, Portuguese, French, Spanish and Chinese), also taking into account that collocations are automatically activated for each language covered by the platform, as the presentation screen of PLATCOL's prototype illustrates (Figure 1).

[4] We use the term *dictionaries* as we mean that users can opt to activate monolingual, bilingual or even multilingual dictionaries, according to their needs and languages they want to search for.

Figure 1: Screenshot of PLATCOL's Presentation Screen Prototype

The new site is under construction, as it will be adjusted to the new languages (French, Spanish and Chinese)[5], with a more ambitious and interactive design as well as more detailed and enhanced lexicographical features and methodology.

4.1 User Profile and Needs

In any lexicographic work, reference is made to the following topics: typology of users, their needs and skills. Thus, in many studies, users' "problem" and needs are the main focus. However, as Fuertes Olivera and Tarp (2014) clearly state, this concern does not bear fruit, since it does not materialise in concrete theoretical and practical decisions, but instead researchers tend to approach the problem in a more general way and do not go into further discussion. Consequently, it is proposed that a better approach is to differentiate between two types of lexicography: a contemplative and a transformative one.

[5] A site used to host the *Bilingual Collocations Dictionary* (Orenha-Ottaiano 2017) and was modified for PLATCOL (http://www.institucional.grupogbd.com/dicionario/ index?locale=pt), where users can find information about the platform. However, a new software is being developed under the new methodology and an updated microstructure will be inserted in the near future.

In contemplative lexicography, dictionaries are analysed and users questioned about their use of existing dictionaries to date. In transformative lexicography, theoretical analyses of the potential user situations, the respective user conditions and needs are used to develop new approaches for compiling new dictionaries, typically monofunctional dictionaries (Bergenholtz, Bothma & Gouws, 2011: 34-35).

Generally speaking, the first type can be related to the so-called general theory of lexicography; the second type, in turn, is linked to functional theory. Our proposal is in line with this last perspective and thus the following constitute essential points that guide the development of the platform:

a) The prior definition of the users' profiles to which the proposal is addressed, a crucial step before its elaboration. These are the profiles that have already been defined:

Language Learners	Non-native users, students of an additional language of intermediate or advanced level (from B1 level on, according to the Common European Framework of Reference for Languages: Learning, teaching, assessment), in any environment (university studies, language courses, and so on)
Pre-Service Teachers	Language learners (student teachers) from higher education institutions trained to become professional language teachers
In-Service Teachers	Additional language teachers, native or non-native ones, with specific training or degree in languages
Translators	Learner or professional translators, native or non-native, of non-specialized texts
Material Developers	Authors of manuals and teaching materials aimed at teaching and learning additional languages
Researchers or Lexicographers	Researchers in general, especially linguists, phraseologists and lexicographers

Table 8: User profiles.

b) The consideration of specific extra-lexicographic or social situations that would motivate the use of the platform: "to determine which type of needs a specific type of user may have in each type of situation" (Bergenholtz & Tarp, 2003:173):

We start from the idea that the different target audiences of a lexicographic work have a series of information and consultation needs (Fuertes Olivera & Tarp, 2014). These needs can only be met if users have quick and easy access to a set of lexicographic data prepared according to their profile. This way, users should be able to extract the information they need, so that they can employ it later, according to their purposes. These purposes, in turn, are always related to the extra-lexicographic contexts and situations that gave rise to these needs (Tarp, 2015).

Considering the profile of potential users of the platform, we acknowledge that the

lexicographically relevant social situations, among the four defined within functional theory, are as follows: 1. Communicative, in which users try to solve problems related to production, reception, translation, proofreading and correction of written or oral texts; and 2. Cognitive, when users need or want to expand their knowledge of something. This typology could be applied to the profile of all indicated users; however, recognizing the limitations of the proposal, it is necessary to establish some restrictions, as Table 9 shows.

Language Learners	Communicative situations are limited to the production of written texts. With regard to cognitive situations, these would be related to the context of language learning; the user would consult PLATCOL with the aim of translating, revising or correcting a text
Pre-Service Teachers	In the case of non-native pre-service teachers, communicative situations are connected to production, translation and proofreading written or oral texts. Regarding cognitive situations, our goal is that pre-service teachers use the platform to develop collocational competence, improving their ability to solve doubts about the use of collocations and helping them to understand the problems posed by their didactics
In-Service Teachers	In this case, communicative situations are related to text correction and review. In the case of non-native teachers, cognitive situations may also occur, mainly related to the preparation of teaching materials
Translators	In this case, the Platform could be useful in many communicative situations - both in the reception and in the transfer, reproduction and revision of texts -, as well as in cognitive situations, to assist translators who need specific lexicographic data related to the frequency or context of using a collocation, for example
Material Developers	The communicative situations relevant to these users refer, above all, to text review and correction. Also, in this case, cognitive situations related to the preparation of manuals and teaching-learning materials may occur
Researchers or Lexicographers	In the case of non-native speakers, communicative situations may occur in situations related to text production, revision and correction. Native speakers, in turn, can find themselves in contexts in which the platform can be useful to access certain information about collocations, such as examples, contexts of use, classification, etc.

Table 9: User profiles related to lexicographically relevant social situations and some restrictions.

c) the determination of the platform's lexicographic functions:

A lexicographic function must be understood as "the assistance provided by the dictionary to meet a certain type of user's specific needs in a certain type of extra-lexicographical situation"[6] (Fuertes Olivera & Tarp, 2008: 80, the translation is ours). Our proposal must be considered to be multifunctional, since, according to the extra-lexicographic situations discussed, it must fulfill two functions: a communicative and cognitive one. Given the recommendations of functional theory and considering that users' abilities in dictionary use cannot be determined in advance, we must ensure that access to information is quick and easy.

For this reason, the dictionaries' macrostructure includes a systematic introduction and

[6] "...la asistencia que presta el diccionario para satisfacer el tipo específico de necesidades que tiene un determinado tipo de usuarios en un determinado tipo de situación extra-lexicográfica" (Fuertes Olivera & Tarp, 2008: 80)

usage guide. Likewise, the design of the dictionaries' microstructure has been made taking into account users' profile and needs. The features here described about users' needs and profiles are based on our considerable experience of translation, translation training, foreign language teaching and teacher training. In the near future, we intend to carry out research on users' needs among the target groups.

4.2 Dictionaries' microstructure

The compilation of a collocations dictionary, an already complex task, becomes even more challenging when multiple languages are taken into consideration. The organization of the microstructure, as explained below, is especially daunting.

PLATCOL's entries include nouns, verbs, and adjectives which correspond to the bases of the collocations (see more about the collocations structures in this section).

In a collocations dictionary, the headwords can be organized according to at least two different principles. One of the views in the treatment of collocations is statistically based. Collocations are defined under a statistical approach iwith regard to their frequent co-occurrence. This way, the headword can be either the base or the collocate, depending on the frequency of co-occurrence in the corpus.

The other view follows Hausmann's approach (1985, 1989), using the concept of the base, the element usually known by users, and of the collocate, the element they are searching for, that is to say, what learners and translators, for example, need to find.

In this project, we opted for the latter view (Hausmann 1985, 1989), claiming that it is more user-friendly and effective with regard to most user profiles, besides being the starting point for most users. Moreover, users will be able to perform either base or collocate searches in the platform search bar.

The entries of the multilingual collocation dictionaries consist of the following elements:

A **headword**, which corresponds to the basis of the collocations. Headwords can be nouns, verbs or adjectives
A **word class**: a word class is placed right after the headword (the base of the collocation). In the case of these collocation dictionaries, they will be either a noun (n.), a verb (v.) or an adjective (adj.). If a word belongs to more than one word class, such as *abstract* (n.), *abstract* (v.) and *abstract* (adj.), each word class appears in separate entries, so that the collocations, collocations structures and other pieces of information are easily organized
Frequency of each headword
A **definition**: a brief definition of the different senses of the base will be provided. The decision of including a definition is that the collocations can be duly organized according to each sense of the headword. Hence, users will be able to have quicker access to the collocations they are searching fo

Table 10: Entry elements of the Multilingual Collocation Dictionaries

The collocations are structured as follows:

Collocation syntactic structure: depending on the part of speech of the entry and the language of this collocation, collocations are organized according to the syntactic structure below (Hausmann 1985, 1989, Orenha-Ottaiano 2009, 2016, 2017)
Collocation taxonomy: verbal, nominal, adjectival and adverbial
In each section of each headword and definition, users can choose collocations to be either displayed in alphabetical order or by frequency or salience (ranked according to their statistical score). For more specialized users, such as researchers and lexicographers, collocations can be ranked by t-score, MI score etc.
Incorporation of usage examples: to illustrate how collocations are used based on a specific meaning. Users will have the chance to choose from displaying from 1 to 5 examples

Table 11: Collocations' organization

Below, Table 12 shows a summarized entry structure:

ENTRY
<headword> plan</headword>
part-of-speech > noun
gramrel 1 > verb + NOUN develop plan
collocate 1> develop (develop plan)
Collocation frequency (Advanced Options)
Statistical measure (Advanced Options)
example (up to 5)
collocate 2 > come up with (come up with plan)
Collocation frequency (Advanced Options)
Statistical measure (Advanced Options)
example (up to 5 - Advanced Options)
collocate n > propose (propose plan)
Collocation frequency (Advanced Options)
Statistical measure (Advanced Options)
example (up to 5)
gramrel 2 > NOUN + verb plan cover
collocate 1 > covers
Collocation frequency (Advanced Options)
Statistical measure (Advanced Options)
example (up to 5 - Advanced Options)
collocate 2 > cover
Collocation frequency (Advanced Options)
Statistical measure (Advanced Options)
example (up to 3 - - Advanced Options)

Table 12: Microstructure adapted and expanded from Orenha-Ottaiano et al. (2020).

According to the type of collocation and language, the collocations will have the following syntactic structures applied to English:

Verbal	Adverbial
verb $_{collocate}$ + noun $_{base}$ noun $_{base}$ + verb $_{collocate}$ verb $_{collocate}$ + prep. + noun $_{base}$ verb $_{collocate}$ + adverbial particle + noun $_{base}$	adverb $_{collocate}$ + adjective $_{base}$ verb $_{base}$ + adverb $_{collocate}$ adverb $_{collocate}$ + verb $_{base}$
Nominal	**Adjectival**
noun $_{base}$ + noun $_{collocate}$ noun $_{base}$ + prep. + noun $_{base}$	adjective $_{collocate}$ + noun $_{base}$

Table 13: Collocations' Taxonomy and Syntactic Structures.

The syntactic structures or order of the elements of collocations may vary from one language to the other. For example, adjectival collocations in Portuguese, Spanish and French can have two different syntactic structure orders, depending on the meaning the speaker wishes to convey:

Noun $_{base}$ + Adjective $_{collocate}$
Adjective $_{collocate}$ + Noun $_{base}$

Users will then have free access to PLATCOL's basic microstructure dictionaries, without having to sign in (as shown in Figure 2).

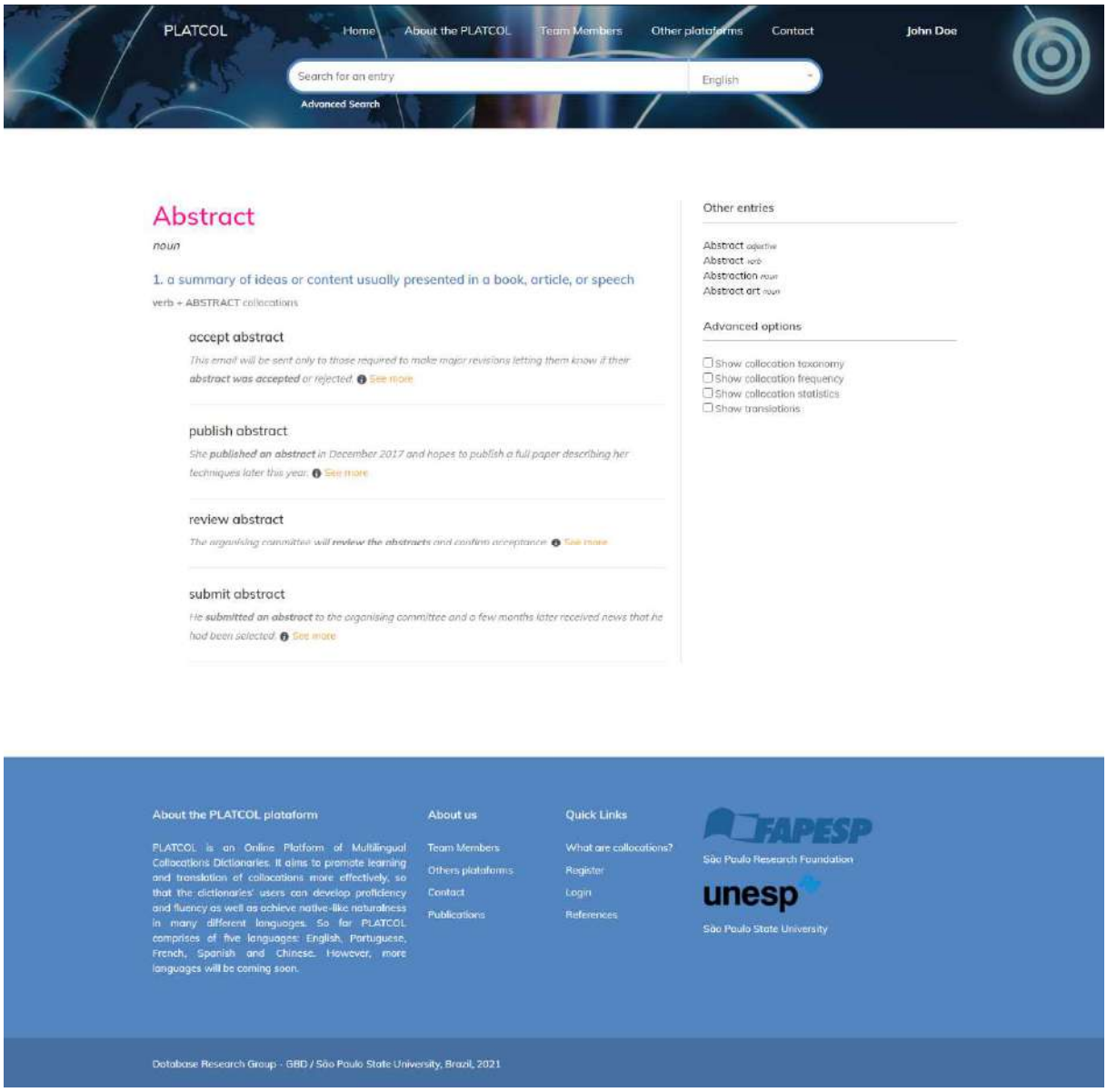

Figure 2: Screenshot of basic structure of an entry.

Besides the basic microstructure, *Advanced options* will be available if a user opts to sign in, according to their profile.

A new dictionary structure will be available so users can choose from items in a *Menu* containing the following elements:

☐	Collocation frequency
☐	Collocation's statistical information: it provides users with statistical measures so that they can check or analyze the results of each collocation's frequency of co-occurrence
☐	Taxonomy of Collocation
☐	Translation of collocations

Table 14: Dictionary's menu options.

Figures 3 and 4 show the dictionary structure generated by the items chosen from a menu in *Advanced options*.

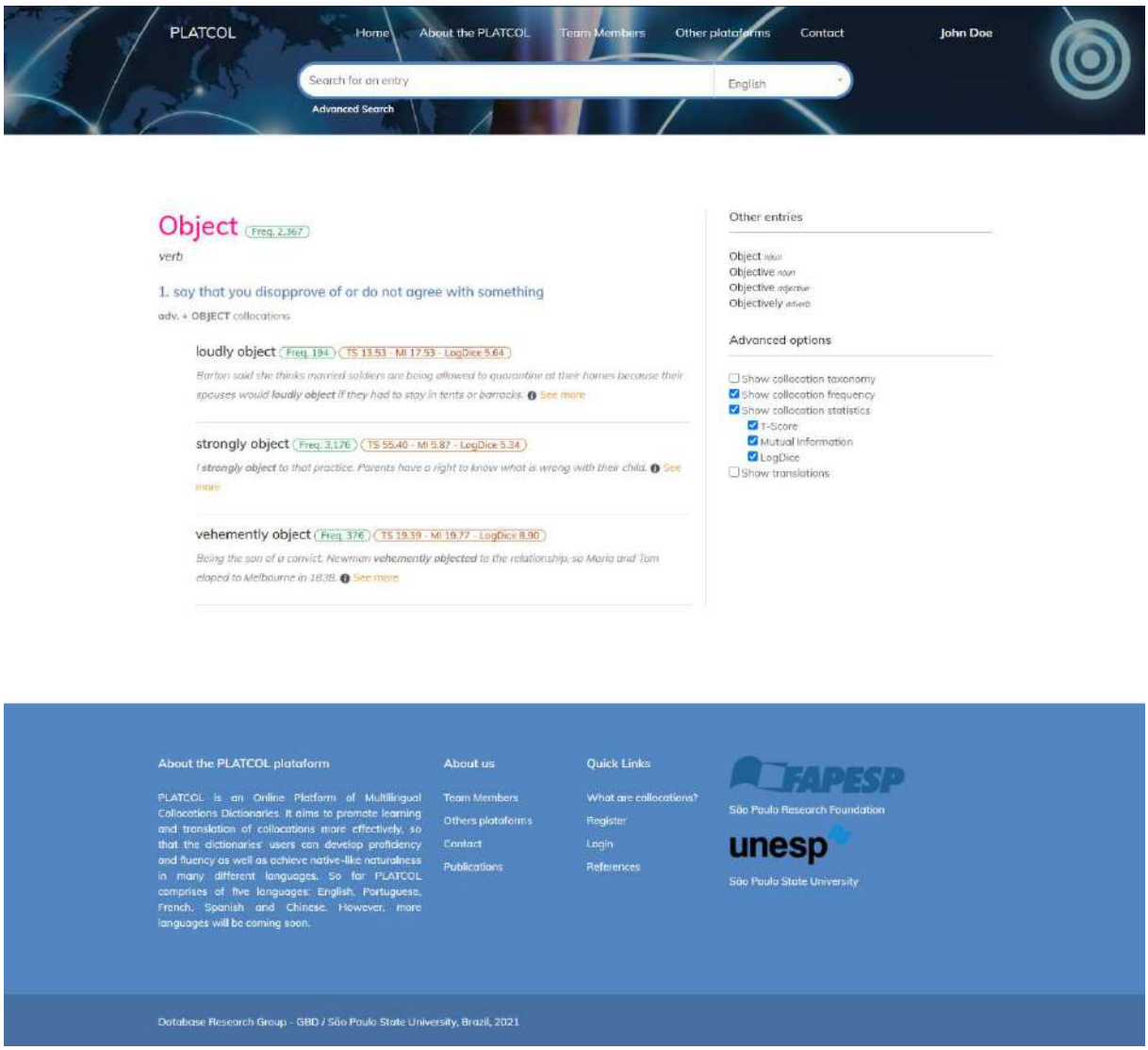

Figure 3: Screenshot of the advanced option microstructure (the entry is a verb).

Figure 4: Screenshot of the advanced option microstructure (the entry is an adjective).

Additionally, a user may opt to click on *Advanced options* and choose to see the translation equivalents of the sought entry (*plan*) and its collocations in, for example, two more languages of the platform, Portuguese and Spanish (Figure 5).

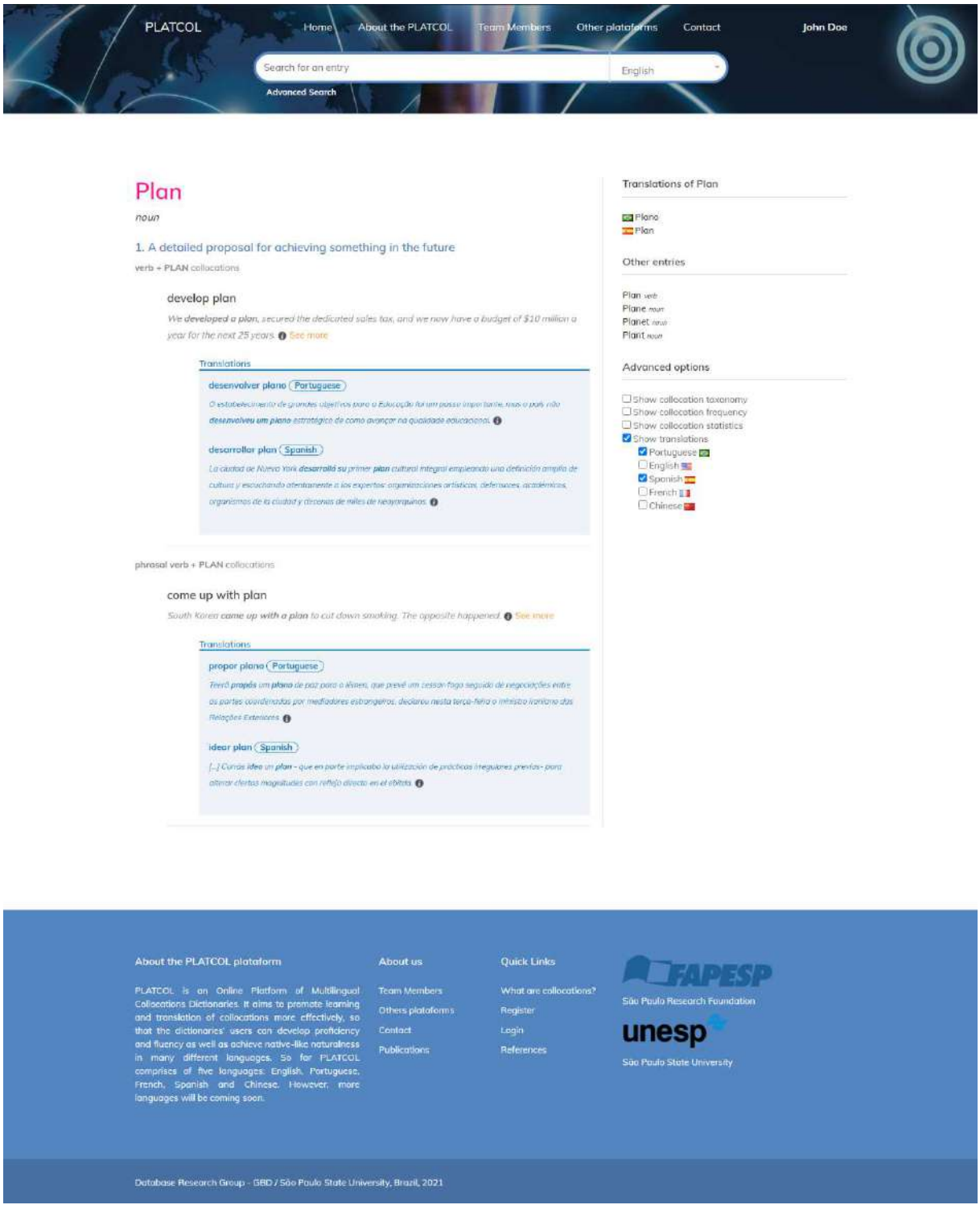

Figure 5: Screenshot of a user's choice for a translation equivalent of the entry *plan*.

Of course, future developments of the platform will take into account user feedback.

With respect to post-editing and validation of entry structures, the research will undertake the following three phases (traffic lights phases), indicating to users their status:

PHASE 1	Data automatically inserted into the Platform, and not revised yet, will be displayed with a red icon beside it.
PHASE 2	Data revised by one member of the team (reviewer 1), but may still need a second evaluation and/or some adjustments or corrections. This time, there will be an orange icon.
PHASE 3	Data checked by a second reviewer (reviewer 2) and now considered to be correct. There will be a green icon beside it. In case, if, for any reason, it cannot be validated, it keeps the same status, that is to say, with an orange icon, and reviewer 1 will have to make some adjustments or corrections.

Table 15: Phases for post-editing and validation of entry structures

This strategy allows users to have access to all entries, collocations and automatically extracted data without having to wait until the whole validation process is over.

As this is an ongoing project, some methodological aspects as well as macro and microstructure decisions may still be changed or reshaped, with a view to best adjust the platform to the new languages investigated as well as to users' different lexicographical needs. Matters regarding the number of collocations or the amount of data to be displayed on the collocation dictionaries' screen as well as types of filter (Kosem et al., 2019b), aiming to help users find relevant information according to their profile and needs, are still being investigated and will be further discussed in future work.

4.3 Dictionary typology and directionality

Regarding the coverage of languages, the platform can display monolingual, bilingual or multilingual dictionaries. With regard to directionality, collocations are retrieved from all corpora languages and will be automatically translated and post-edited in the following directions:

- from English into Portuguese;
- from Portuguese into English;
- from Spanish into Portuguese;
- from Spanish into English;
- from Chinese into Spanish.

These directions serve only for research purposes. It is worth mentioning that another pair or group of languages can be chosen since the corresponding settings are manually entered into the system, regardless of the automatic retrieval process. Once a collocation in a given language is registered, translations into other languages can also be manually defined in the system.

Once translation pairs between collocations are identified and registered in the system,

making up a multilingual database, it becomes possible to identify and automatically suggest new translations among other languages. This process occurs through an inference-based algorithm, built from an inference hypothesis related to the composition of multiple translation dictionaries: if word A translates into word B which in turn translates into word C, what is the probability that C is a translation of A? Studies developed under this hypothesis (e.g. Mausam et al., 2010), presented significant results in relation to the analysis via inference of translation pairs between different languages. In this process, the algorithm performs the analysis of previously registered translations, identifies other translation pairs via inference, and shows lexicographers the possibilities of translations, who must analyze the reliability and quality of the translation found.

For example, the collocations "develop a plan", in English, and "desenvolver um plano", in Portuguese, are equivalents. Similarly, the collocations "desenvolver um plano", in Portuguese, and "desarrollar un plan", in Spanish, also have a translation relationship. This way, even if it has not been previously identified in the automatic extraction process, the relationship between the collocations "develop a plan", in English, and "desarrollar un plan", in Spanish, will be automatically inferred.

4.4 The Dictionaries and CEFR levels

Second language teachers have classified collocations into different CEFR levels, but this classification is not common in collocation dictionaries. Even in learners' English dictionaries which include the level of CEFR, such as *Cambridge*, the level is assigned to the headword, but there is no information about the collocations under the headword. For example, the noun *crime*, assigned as B1. There is no information about collocations such as *commit crime*, *charged of crimes* or *alleged crimes* which appear as examples and do not seem to belong to the same level. We are interested in the relevance of collocations for all levels and, therefore, this dictionary should include collocations for all CEFR learners.

This claim leads to the challenge of establishing criteria to assign collocations to a specific level. There are different approaches. The *English Vocabulary Profile* (Capel, 2010) adds data from learner corpora to frequency information obtained from English corpora or vocabulary lists to determine the lexicon non-native speakers should know at a given level. DICI-A (*Dizionario delle Collocazioni Italiane per Apprendenti*), on the other hand, takes a corpus of native speakers as a reference point (Spina, 2016) and uses a set of parameters to determine the level of collocations it includes: the frequency and dispersion of a collocation in the corpus, its function (expressions with descriptive meaning versus marks of textual organization and pragmatic elements) and the topic with which the collocation in question is associated. As for Spanish collocations, García-Salido and Alonso (2018) choose frequency in the corpus to level the collocations of the *DiCE*, but taking as a point of departure the collocations included in the *Plan Curricular del Instituto Cervantes* (Instituto Cervantes, 1997-2016). By means of

analysis of a sample of collocations included in both the dictionary and the *Plan Curricular del Instituto Cervantes*, a negative correlation was found between the levelling proposed for those collocations in the *Plan Curricular* and the corpus frequency; that is, higher levels correspond to lower frequencies, and vice versa.

A challenge for assigning CEFR levels in a multilingual collocation dictionary is to find the equivalence between different languages. For instance, according to frequency criterion, a given collocation in a language could be assigned to B1 level, however, its equivalent in another language could be classified into a lower or higher one, according to the same criterion. For example, even though the collocations *black coffee*, *café solo*, *café noir*, and *café preto* could be considered translation equivalents, they are not found equally in different language corpora and may not be assigned to the same CEFR level.

5. Conclusion and further work

This paper outlined a corpus-based methodology for the development of the Online Platform for a Multilingual Collocations Dictionary, PLATCOL. It described the lexicographical features developed to compile PLATCOL's collocations dictionaries and presented their macro and microstructure.

We also discussed the automatic approaches to annotate corpora with lemmas, PoS-tags and dependency relations in the five languages of PLATCOL. Automatic methods to extract candidate collocations were also explained as well as statistical measures and distributional semantics strategies to select the candidates described, highlighting the relevance of post-edition in the lexicographical process.

The collocations dictionaries' prototypes were presented to illustrate PLATCOL's customized design, layout and lexicographical features, stressing the importance of developing an innovative customization methodology tailored to users' needs and specifically designed for a collocations dictionary. Hence, we hope to contribute to future lexicographical and phraseological/phraseographical research.

For future work, we will take advantage of the strategy presented by Garcia *et al.* (2019c) to gather candidate translations for each selected collocation. This approach generates lists of bilingual collocation equivalents, which will be then reviewed by those lexicographers with a good proficiency in each language pair, approving those proper equivalents which have been automatically extracted by the system, and providing new translations when necessary.

6. Acknowledgements

We gratefully acknowledge the financial support provided by The São Paulo Research Foundation (FAPESP), Process ner 2020/01783-2.

7. References

Alonso-Ramos, M. (2001). Construction d'une base de données des collocations bilingue français-espagnol. *Langages,* 35(143), pp. 5-27.

Alonso-Ramos, M. (2004). *DiCE: Diccionario de Colocaciones del Español*. Universidade da Coruña. Accessed at: http://dicesp.com. (12 April 2021).

Alonso-Ramos, M. (ed.). (2008). Papel de los diccionarios de colocaciones en la enseñanza de español como L2. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*, Barcelona: IULA/Documenta Universitaria, pp. 1215-1230.

Alonso-Ramos, M. & García-Salido, M. (2019). Testing the Use of a Collocation Retrieval Tool Without Prior Training by Learners of Spanish. *International Journal of Lexicography*, 32(4), pp. 480-497.

Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Artetxe, M., Labaka, G. & Agirre, E. (2018). A robust self-learning method for fully unsu-pervised cross-lingual mappings of word embeddings. In I. Gurevych & Y. Miyao (eds.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, pp. 789–798. Available at: https://www.aclweb.org/anthology/P18-1073.pdf.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3), pp. 209-226.

Beauchesne, J. (2001). *Dictionnaire des cooccurrences*. Montréal: Guérin.

Benson, M., Benson, E. & Ilson, R. (1997). *The BBI Combinatory Dictionary of English*. Amsterdam/Philadelphia: John Benjamins.

Bergenholtz, H. & Tarp, S. (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes. Journal of Linguistics,* 31, pp. 171-196.

Bergenholtz, H., Bothma, T. & Gouws, R.H. (2011). A model for integrated dictionaries of fixed expressions. In I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st century: New Applications for New Users. Proceedings of eLex 2011.* Bled, Slovenia, pp. 34-42. Available at: https://elex2011.trojina.si/elex2011_proceedings.pdf.

Bernardini, S. (2007). Collocations in Translated Language: Combining Parallel, Comparable and Reference Corpora. In M. Davies, P. Rayson, S. Hunston & P. Danielsson (eds.) *Proceedings of the Corpus Linguistics Conference* (CL2007). Birmingham, UK, pp. 1-16. Available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/15_Paper.pdf.

Berti, B. & Pinnavaia, L. (2014). Creating a Bilingual Italian-English Dictionary of Collocations. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI Euralex International Congress*: The User in Focus. Bolzano/Bozen, pp. 515-524. Available at: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX_Part_3.

pdf.

Bird, S., Loper, E. & Klein, E. (2009). *Natural Language Processing with Python.* Sebastopol, CA*:* O'Reilly Media Inc. Available at: http://www.datascienceassn.org/sites/default/files/Natural%20Language%20Processing%20with%20Python.pdf.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135-146.

Bond, F. & Foster, R. (2013). Linking and extending an open multilingual wordnet. In H. Schuetze, P. Fung & M. Poesio (eds.) *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria, pp. 1352-1362. Avalaible at: https://www.aclweb.org/anthology/P13-1133.pdf.

Bosque, I. (2006). *Diccionario combinatorio práctico del español contemporáneo.* Madrid: SM.

Bothma, T.J.D. & Tarp, S. (2012). Lexicography and the Relevance Criterion. *Lexikos*, 22, pp. 86-108.

Capel, A. (2010). A1-B1 Vocabulary: Insights and issues arising from the English Profile Wordlists Project. *English Profile Journal*, 1, pp. 1-11.

Corpas Pastor, G. (2017) Collocations in e-Bilingual Dictionaries: from Underlying Theoretical Assumptions to Practical Lexicography and Translation Issues. In S. Torner & E. Bernal (eds.) *Collocations and other Lexical Combinations in Spanish. Theoretical and Applied Approaches*. London: Routledge, pp. 139-160,

de Marneffe, M.C., Manning, C.D., Nivre, J. & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), pp. 1-52. DOI: https://doi.org/10.1162/coli_a_00402.

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

Evert, S., Uhrig, P., Bartsch, S. & Proisl, T. (2017). E-VIEW-affilation–A large-scale evaluation study of association measures for collocation identification. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Proceedings of eLex 2017–Electronic lexicography in the 21st century*: *Lexicography from Scratch.* Leiden, the Netherlands, pp. 531-549. Available at: https://elex.link/elex2017/proceedings/eLex_2017_Proceedings.pdf.

Fuertes-Olivera, P.A. & Tarp, S. (2014). *Theory and Practice of Specialised Dictionaries. Lexicography versus Terminography.* Berlín/Boston: Walter de Gruyter.

Fuertes-Olivera, P.A. & Tarp, S. (2008). La Teoría Funcional de la Lexicografía y sus consecuencias para los diccionarios de economía del español. *Revista de Lexicografía*, XIV, pp. 75-95.

Gamallo, P., Garcia, M., Piñeiro, C., Martinez-Castaño, R. & Pichel, J. C. (2018). LinguaKit: a Big Data-based multilingual tool for linguistic analysis and

information extraction. In Institute of Electrical and Electronics Engineers (ed.). *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS).* Valencia, Spain, pp. 239-244.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2017). Using bilingual word-embeddings for multilingual collocation extraction. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (eds.) *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017).* Valencia, Spain, pp. 21–30. Available at: https://www.aclweb.org/anthology/W17-1703.pdf.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019a). Towards the automatic construction of a multilingual dictionary of collocations using distributional semantics. In I. Kosem, T. Z. Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Proceedings of eLex 2019: Smart Lexicography.* Sintra, Portugal, pp. 747-762. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_42.pdf.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019b). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In A. Savary, C. Parra Escartín, F. Bond, J. Mitrović & V. B. Mititelu (eds.) *Proceedings of the Joint Workshop on Multiword Expressions and WordNet* (MWE-WN 2019). Florence, Italy, pp. 49-59. Available at: https://www.aclweb.org/anthology/W19-5107.pdf.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019c). Weighted compositional vectors for translating collocations using monolingual corpora. In G. Corpas Pastor & R. Mitkov (eds.) *Computational and Corpus-Based Phraseology.* Cham, Switzerland: Springer, pp. 113-128.

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

Hausmann, F.J. (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz & J. Mugdan (eds.) *Lexikographie und Grammatik.* Tübingen: Niemeyer, pp. 118-129.

Hausmann F.J. (1989). Le dictionnaire de collocations. In F.J. Hausmann, O. Reichmann, H.E. Wiegand & L. Zgusta (eds) *Wörterbücher: ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires.* Berlin/New-York: De Gruyter, pp. 1010-1019.

Jousse, A.L. & Polguère, A. (2005). *Le DiCo et sa version DiCouébe. Document descriptif et manuel d'utilisation.* Université de Montréal: Observatoire de linguistique Sens-Texte (OLST).

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the 13th EURALEX International Congress.* Barcelona: Institut Universitari de Linguistica Aplicada/Universitat Pompeu Fabra, pp. 425–432.

Koehn, P. (2005). Europarl: a parallel corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit.* Phuket, Thailand, pp. 79–

86. Available at: https://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf.

Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J. & Tiberius, C. (2019a). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119–137.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C. (2019b). Collocations Dictionary of Modern Slovene. Proceedings of the 18th EURALEX International Congress: lexicography in global contexts. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 989-997.

Laufer, B. (2011). The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language. *International Journal of Lexicography*, 24(1), pp. 29–49.

L'Homme, M.C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire,* 217, pp. 78-103.

L'Homme, M.C., Robichaud, B. & Prévil, N. (2018). Browsing the Terminological Structure of a Specialized Domain: A Method Based on Lexical Functions and their Classification. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *11th Language Resources and Evaluation, LREC 2018*. Miyazaki, Japon, pp. 3079-3086. Available at: https://www.aclweb.org/anthology/L18-1487.pdf.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In K. Bontcheva & J. Zhu (eds.) *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland, pp. 55-60.

Mausam, S., Etzioni, S., Weld, O., Reiter, D.S., Skinner, K. Sammer, M. & Vessier, S. (2010). Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174(9-10), pp. 619-637.

Mayor, M. (ed.) (2013). *Longman Collocations Dictionary and Thesaurus*. Harlow: Pearson Education.

Mei, J. (ed.) (1999). *Xiandai hanyu dapei cidian* (1st ed.). Shanghai: hanyu da cidian chuban she.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (eds.) *Workshop Proceedings of the International Conferenceon Learning Representations (ICLR 2013)*. Scottsdale, AZ, USA. Available at: https://arxiv.org/pdf/1301.3781v3.pdf.

Orenha-Ottaiano, A. (2020). The creation of an Online English Collocations Platform to help develop collocational competence. *PHRASIS. Rivista di Studi Fraseologici e Paremiologici,* 4, pp. 59-81.

Orenha-Ottaiano, A. (forthcoming). Escolhas colocacionais a partir de um Corpus de Aprendizes de Tradução e a importância do desenvolvimento da competência colocacional. *Cadernos de Fraseoloxía Galega*.

Orenha-Ottaiano, A. (2017). The compilation of an Online Corpus-Based Bilingual Collocations Dictionary: motivations, obstacles and achievements. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Proceedings of eLex 2017–Electronic lexicography in the 21st century*: *Lexicography from Scratch.* Leiden, the Netherlands, pp. 458-473. Available at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper27.pdf.

Orenha-Ottaiano, A. (2013). The proposal of an electronic bilingual dictionary based on corpora. In O. Karpova (ed.) *Life Beyond Dictionaries. X International School on Lexicography.* Florence, Italy, pp. 405-408.

Orenha-Ottaiano, A., Kuhn, T.Z. & Valêncio, C.R. (2020). The building of an Online Platform for Monolingual Dictionaries of Academic Collocations in Portuguese and English. Paper presented at the *56th Linguistics Colloquium*, online.

Oxford Collocations Dictionary (English-Chinese) (2nd ed.). (2006). Oxford: Oxford University Press.

Oxford Collocations Dictionary (English-Chinese) App version. Accessed at: https://play.google.com/store/apps/details?id=hk.com.oup.dicts&hl=en_US&gl=US.

Padró, Ll. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Language Resources and Evaluation Conference (LREC 2012).* Istanbul, Turkey, pp. 2473-2479. Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf.

Pereira, L.A.S. & Mendes, A. (2002). An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. In A. Braasch & C. Povlsen (eds.) *Proceedings of the 10th EURALEX International Congress*, v. II. Copenhagen, Denmark, pp. 841-849.

Rundell, M. (2010). *Macmillan Collocations Dictionary for Learners of English.* Oxford: Macmillan Publishers Ltd.

Santos, D. & Rocha, P. (2004). The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck & B. Magnini (eds.) *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum* (CLEF 2004), Bath, UK, pp. 821-832.

Spina, S. (2010). The Dictionary of Italian Collocations: Design and Integration in an Online Learning Environment. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.) *Conference Proceedings of the International Conference on Language Resources and Evaluation* (LREC 2010). Valletta, Malta, pp. 3202-3208. Available at: http://www.lrec-conf.org/proceedings/lrec2010/pdf/681_Paper.pdf.

Spina, S. (2016). Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations. In B. Sanromán Vilas (ed.) *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching* (Mémoires de la Société Néophilologique de Helsinki).

Helsinki: Société Néophilologique, pp. 219–244.

Straka, M. & Straková, J. (2017). Tokenizing, POS-tagging, lemmatizing and parsing UD 2.0 with UDPipe. In J. Hajič & D. Zeman (eds.) *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada, pp. 88-99. Available at: https://www.aclweb.org/anthology/K17-3009.pdf.

Tarp, S. (2015). La teoría funcional en pocas palabras. *Estudios de Lexicografía*, 4, pp. 31-42.

Torner, S. & Bernal, E. (eds.) (2017). *Collocations and Other Lexical Combinations in Spanish*. London: Routledge.

Wagner Filho, J. A., Wilkens, R., Idiart, M. & Villavicencio, A. (2018). The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *11th Language Resources and Evaluation, LREC 2018*. Miyazaki, Japon, pp. 4339-4344. Available at: https://www.aclweb.org/anthology/L18-1686.pdf.

Wiegand, H.E. (1984). On the Structure and Contents of a General Theory of Lexicography. In R.R.K. Hartmann (ed.) *LEXeter'83 Proceedings. Papers from the International Conference on Lexicography. Exeter, 9-12 September 1983*. Tübingen: Niemeyer, pp. 13-30.

Yang, T. (1990). *Hanyu changyongci dapei cidian*. (1st ed.). Beijing: Waiyu jiaoxue yu yanjiu chuban she.