Improving Collocation Correction by Ranking Suggestions Using Linguistic Knowledge

Roberto Carlini[1], *Joan Codina-Filba*[1], *Leo Wanner*[2,1]

(1) Natural Language Processing Group, Department of Information and Communication Technologies
Pompeu Fabra University, Barcelona
(2) Catalan Institute for Research and Advanced Studies (ICREA)

roberto.carlini@upf.edu, joan.codina@upf.edu, leo.wanner@upf.edu

ABSTRACT

The importance of collocations in the context of second language learning is generally acknowledged. Studies show that the “collocation density" in learner corpora is nearly the same as in native corpora, i.e., that use of collocations by learners is as common as it is by native speakers, while the collocation error rate in learner corpora is about ten times as high as in native reference corpora. Therefore, CALL could be of great aid to support the learners for better mastering of collocations. However, surprisingly few works address specifically research on CALL-oriented collocation learning assistants that detect miscollocations in the writings of the learners and propose suggestions for their correction or that offer the learner the possibility to verify a word co-occurrence with respect to its correctness as collocation and obtain suggestions for its correction in case it is determined to be a miscollocation. This disregard is likely to be, on the one hand, due to the focus of the CALL research so far on grammatical matters, and, on the other hand, due to the complexity of the problem. In order to be able to provide an adequate correction of a miscollocation, the collocation learning assistant must “guess" the meaning that the learner intended to express. This makes it very different from grammar or spell checkers, which can draw on grammatical respectively orthographic regularities of a language. In this paper, we focus on the problem of the provision of a ranked list of correction suggestions in a context in which the learner submits a collocation for verification and obtains a list of correction suggestions in the case of a miscollocation. We show that the retrieval of the suggestions and their ranking benefits greatly from NLP techniques that provide the syntactic dependency structure and subcategorization information of the word co-occurrences and a weighted Pointwise Mutual Information (PMI) that reflects the fact that in a collocation, it is the base that is subject of the free choice of the speaker, while the occurrence of the collocate is restricted by the base, i.e., that collocations are per se asymmetric.

KEYWORDS: CALL, collocations, miscollocation correction, syntactic dependencies, subcategorization, PMI.

Roberto Carlini, Joan Codina-Filba and Leo Wanner 2014. Improving collocation correction by ranking suggestions using linguistic knowledge. *Proceedings of the third workshop on NLP for computer-assisted language learning.* NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107: 1–12.

1 Introduction

The importance of collocations in the context of second language learning is beyond any doubt; see, among others, (Granger, 1998; Lewis, 2000; Nesselhauf, 2004, 2005; Lesniewska, 2006; Alonso Ramos et al., 2010) for studies in this respect. Orol and Alonso Ramos (2013) even show in their study that the "collocation density" in learner corpora is nearly the same as in native corpora, i.e., that the use of collocations by learners is as common as it is by native speakers. At the same time, they also find that the collocation error rate in learner corpora is about 32% (compared to about 3% by native speakers). That is, *Computer Assisted Language Learning* (CALL) could be of great aid to support the learners for better mastering of collocations. However, surprisingly few works address specifically CALL-oriented collocation learning assistants that detect miscollocations in the writings of the learners and propose suggestions for their correction or that offer the learner the possibility to verify using an interactive interface a word co-occurrence with respect to its correctness as collocation and obtain suggestions for its correction in case it is determined to be a miscollocation; cf. (Wanner et al., 2013) for an overview. This is likely to be, on the one hand, due to the focus on grammatical matters that CALL research had so far, and, on the other hand, due to the complexity of the problem. The problem envisaged by a collocation learning assistant is that in order to be able to provide an adequate correction of a miscollocation, it must "guess" the meaning that the learner intended to express. Thus, if the learner writes *assume an exam*, we do not know a priori (especially not when the learner uses an interface for the verification of *assume an exam* as an isolated word co-occurrence; cf., e.g., http://miscollocation-richtrf.rhcloud.com/ for illustration), whether she wants to say *take an exam* or *pass an exam*. This makes collocation checkers very different from grammar or spell checkers, which can draw on grammatical respectively orthographic regularities of a language.

In what follows, we focus on the problem of the collocation learning assistants of the verification of isolated word co-occurrences introduced by the learner with respect to their collocation status and the provision of a ranked list of correction suggestions in case a co-occurrence is considered to be a miscollocation. In the state-of-the-art proposals, the correction suggestions are ranked in terms of occurrence frequency or point-wise mutual information (PMI). Both measures do not take into account the essential linguistic features of collocations, and, in particular, their dependency structures and subcategorization and their asymmetric nature that results from the fact that the *base* element of a collocation is subject of the free choice of the speaker, while the occurrence of the *collocate* element is restricted by the base.

In the next section, we introduce the linguistic considerations on collocations that motivate our proposal. Section 3 presents how we draw upon these linguistic considerations to rank collocation correction suggestions. Section 4 outlines the experiments we carried out to verify our proposal, and Section 5 discusses the outcome of these experiments. In Section 6, a short summary of the related work is presented, before Section 7 draws some conclusions.

2 The Linguistic nature of collocations

The term "collocation" as introduced by Firth (1957) and cast into a definition by Halliday (1961) encompasses the statistical distribution of lexical items in context: lexical items that form high probability associations are considered collocations. It is this interpretation that underlies most works on automatic identification of collocations in corpora; see, e.g., (Choueka, 1988; Church and Hanks, 1989; Pecina, 2008; Evert, 2007; Bouma, 2010). However, in contemporary lexicography and lexicology an interpretation that stresses the idiosyncratic nature of collocations prevails. Thus, Benson (1989) states that "collocations should be defined not just as 'recurrent word combinations', but as 'ARBITRARY recurrent word combinations' ". "Arbitrary" as opposed to "regular" means that collocations are unpredictable and language-specific. For instance, in English, one *takes a walk*, while in French, German and Italian one 'makes' it: *faire une promenade*, *einen Spaziergang machen*, *fare una passeggiata*, and in Spanish one 'gives' it: *dar un paseo*. In English, one *gives a lecture*, in German and Italian one 'holds' it: *eine Vorlesung halten*, *tenere una lezione*, and in Russian one 'reads' it: *čitat′ lekciju*.

According to Hausmann (1984), Cowie (1994), Mel'čuk (1995) and others, a collocation is a binary idiosyncratic co-occurrence of lexical items between which a direct syntactic dependency holds and where the occurrence of one of the items (the *base*) is subject of the free choice of the speaker, while the occurrence of the other item (the *collocate*) is restricted by the base. Thus, in the case of *take* [*a*] *walk*, *walk* is the base and *take* the collocate, in the case of *high speed*, *speed* is the base and *high* the collocate, etc. It is this understanding of the term "collocation" that we find reflected in general public collocation dictionaries and that we follow since it seems most useful in the context of second language acquisition. However, this is not to say that the two main interpretations of the term "collocation", the distributional and the idiosyncratic one, are disjoint, i.e., necessarily lead to a different judgement with respect to the collocation status of a word combination. On the contrary: two lexical items that form an idiosyncratic co-occurrence are likely to occur together in a corpus with a high value of *Pointwise Mutual Information* (PMI) (Church and Hanks, 1989):

$$PMI = \log\left(\frac{P(a \cap b)}{P(a)P(b)}\right) = \log\left(\frac{P(a|b)}{P(a)}\right) = \log\left(\frac{P(b|a)}{P(b)}\right) \tag{1}$$

The PMI indicates that if two variables a and b are independent, the probability of their intersection is the product of their probabilities. A PMI equal to 0 means that the variables are independent; a positive PMI implies a correlation beyond independence; and a negative PMI signals that the co-occurrence of the variables is lower than the average. Two lexemes are thus considered to form a collocation when they have a positive PMI, i.e., they are found together more often that this would happen if they would be independent variables.

PMI has been a standard collocation measure throughout the literature since Church and Hank's proposal in 1989. It can be used not only for collocation detection, but also for ranking miscollocation correction suggestions: collocations of the base of the miscollocation retrieved from a reference corpus are ranked such that those with a higher PMI appear higher in the correction list than those with a lower PMI. Since for lexemes with largely different individual probabilities (the probabilities can be measured in terms of the number of sentences that contain these words) the PMIs cannot be compared in magnitude, a normalization has been suggested by Bouma (2009):

$$NPMI_{CB} = \frac{\left(\log \frac{P(a,b)}{P(a)P(b)}\right)}{-\log P(a,b)} \tag{2}$$

However, a mere use of PMI, $NPMI_{CB}$ or any similar measure does not consider two central linguistic features of collocations:

1. The lexical dependencies between the base and the collocate are not symmetric, while PMI is commutative, i.e., $PMI(a,b) = PMI(b,a)$.

2. Between the base and the collocate of a collocation, always a direct syntactic dependency holds whose sub-categorization structure depends on the base (as the semantic head of the collocation).

2.1 Lexical asymmetry of collocations

Collocations are lexically asymmetrical. Consider, for instance, the collocation *take an exam*. The base *exam* is far less frequent than the collocate *take*. Thus, if we suppose that in a collection of 10000 sentences *take* appears 1000 times, *exam* 10 times, and *take an exam* 5 times, we obtain:

$$P(take|exam) = 0.5 \gg 0,005 = P(exam|take)$$

with the PMI:

$$PMI = \log\left(\frac{P(take \cap exam)}{P(take)P(exam)}\right) = \log\left(\frac{P(take|exam)}{P(take)}\right) = \log\left(\frac{P(exam|take)}{P(exam)}\right) \quad (3)$$

$$PMI = \log\left(\frac{0.0005}{0.1 \cdot 0.001}\right) = \log\left(\frac{0.5}{0.1}\right) = \log\left(\frac{0.005}{0.001}\right) = \log(5)$$

On the other hand, analyzing the co-occurrence of *exam* with a less frequent verb such as *cheat*, and considering that in the same collection *cheat* appears 10 times and *cheat on an exam* 5 times, we obtain:

$$PMI = \log\left(\frac{P(cheat \cap exam)}{P(cheat)P(exam)}\right) = \log\left(\frac{P(cheat|exam)}{P(cheat)}\right) = \log\left(\frac{P(exam|cheat)}{P(exam)}\right)$$

$$PMI = \log\left(\frac{0.0005}{0.001 \cdot 0.001}\right) = \log\left(\frac{0.5}{0.001}\right) = \log\left(\frac{0.5}{0.001}\right) = \log(500)$$

In both cases, the PMI is positive, such that we can consider both co-occurrences to be valid collocations. However, the PMI of *take* [*an*] *exam* is much smaller than the PMI of *cheat* [*on an*] *exam*. This means that when using the PMI as criterion for ranking collocation suggestions, *take* [*an*] *exam* is ranked much lower than *cheat* [*on an*] *exam* — although *take* [*an*] *exam* is a very common collocation and should be ranked higher.

To address this inconvenience, Bouma (2009) normalizes in Eq. (2) the collocation PMI (i.e., for $PMI > 0$) with a neutral (or symmetric) $p(collocate \cap base)$. However, Eq. (2) does not compensate the penalization of highly frequent collocates: In general, it can be observed that when the collocate is a very common word, the $NPMI_{CB}$ is still penalized. This is because $NPMI_{CB}$ is symmetric with respect to the base and the collocate.

In order to account for this problem, we propose an asymmetric normalization that uses $p(collocate)$:

$$NPMI_C = \frac{PMI(collocate, base)}{-log(p(collocate))} \quad (4)$$

Normalizing with *p(collocate)* and replacing the PMI computed with conditional probabilities, as done in Eq. ((1)), we obtain:

$$NPMI_C = \frac{PMI(collocate, base)}{-\log(P(collocate))} \tag{5a}$$

$$= \frac{\log\left(\frac{P(collocate|base)}{P(collocate)}\right)}{-log(P(collocate))} \tag{5b}$$

$$= \frac{\log(P(collocate|base)) - \log(P(collocate))}{-\log(P(collocate))} \tag{5c}$$

$$= -\frac{\log(P(collocate|base))}{\log(P(collocate))} + 1 \tag{5d}$$

$$= 1 - \log_{P(collocate)}(P(collocate|base)) \tag{5e}$$

In Eq. ((5e)), we can observe that the $NPMI_C$ is the logarithm of the conditional probability, with the probability of the collocate as its base. Figure 1 shows that in the interval [0,1], $NPMI_C$ is always above $NPMI_{CB}$. Furthermore, $NPMI_C$ is much less influenced by high frequencies of the collocate. In the next sections, we will analyze how this changes the ratings of the collocation correction suggestions in some sample cases.

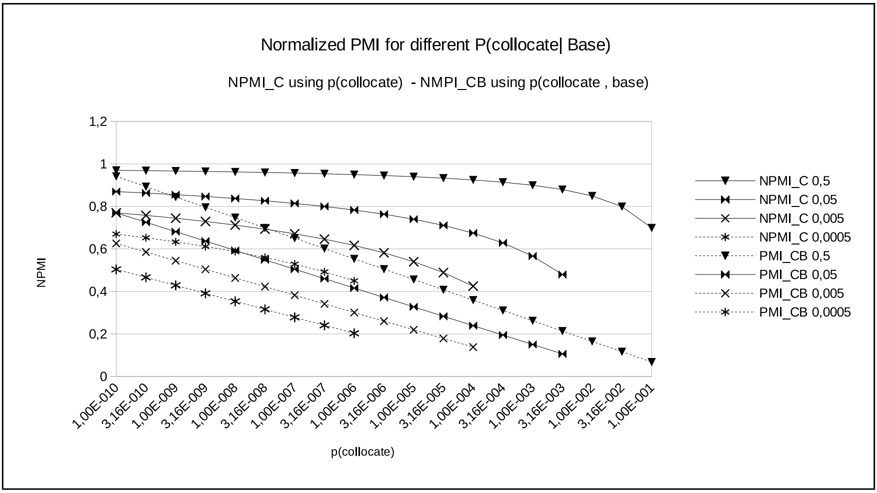

Figure 1: Graph showing the differences between the $NPMI_C$ (continuous line) and $NPMI_{CB}$ (dashed line). Each line shows the variation of the $NPMI$ for different values of $p(collocate|base)$ as the probability of the collocate changes, given a probability of the base of 10^-10.

2.2 Syntactic dependencies in collocations

As argued in lexicography (see the citations above), an essential feature of a collocation is that a direct dependency holds between the base and the collocate. Thus, in *take* [*a*] *walk* as in *I* *took a walk with Mary*, the base *walk* is the direct object of the collocate *take*. In contrast, in *Wexford followed him through the pleached walk and they entered the house by a glazed garden door,*[1] *pleached* is a participle modifier of *walk*. It may thus be considered as an adjective–noun collocation, but not as a verb–direct object collocation *pleach* [*a*] *walk* — as assumed by the MUST collocation checker.

In *I enjoyed the privacy during my walk with Mary*, *enjoy* and *walk* cannot form a collocation at all, even if they may have a rather high PMI, because there is no direct syntactic dependency between them.

Apart from syntactic dependencies, the sub-categorization information of the base must be considered. We say *apply for* [*a*] *job*, *apply to college* and *apply* [*a*] *theory*; *nominate for presidency*, and *nominate* [*the*] *candidate*; and so on. Governed prepositions must be clearly distinguished from semantically full prepositions; cf. *to* in *go to the concert* or *through* in *drive through the city*.

That is, when searching for correction suggestions for a given miscollocation, we need to be aware that a collocation is not a mere prominent co-occurrence of words, as argued in early works in the field; see, e.g., (Choueka, 1988; Church and Hanks, 1989). Rather, it implies dependency information that needs to be taken into account.

3 Ranking of Collocation Suggestions

The basic source that allows us to check a sequence of words provided by the user for its collocation status and to come up with corrections suggestions is a large reference dependency treebank. For each pair of POS-tagged tokens between which a relevant dependency relation holds, the PMI, $NPMI_{CB}$ and $NPMI_C$ are calculated; proper nouns, determiners, conjunctions, numerals, etc. are excluded. As “relevant", we consider relations that have been observed to hold between collocation elements; the most prominent of them are: direct object, indirect object, subject, modifier, and adverbial. When the noun is not a direct dependant of the verb because there is a preposition in between, the preposition is considered part of the relation.

In order to check the validity of a word combination introduced by the user as collocation and to propose correction suggestions for a miscollocation, the following procedure is followed:

1. Check the $NPMI_C$ value of the combination in the database:
   (a) If the combination is found in the reference corpus, the collocation is considered correct or wrong depending on whether the $NPMI_C$ value is positive or negative respectively.
   (b) If the combination does not exist in the reference corpus, we consider its elements as mutually exclusive and, accordingly, the combination as a miscollocation.[2]
2. If the collocation is considered wrong, we attempt to find candidate suggestions for its correction: any valid combination with the base or the collocate of the miscollocation that fits the dependency profile of the miscollocation are retrieved from the corpus.

[1] Example offered by the MUST collocation checker (http://miscollocation-richtrf.rhcloud.com/) on October, 1st 2014 for *pleach walk*, which has been proposed by MUST as one of the collocations the learner should consider along with *take* [*a*] *walk*.

[2] This criterion can be improved by checking whether the individual words exist in the reference corpus. If the collocate or the base are missing, we could pre-calculate PMI values using semantic information from EuroWordNet and search for the missing word using its hypernym.

3. Retrieved candidate suggestions are ranked according to their $NPMI_C$. Only the first items of the list are shown, but the user can ask to go down in the list. Also, the user can solicit sample sentences from the corpus in which the corresponding collocation occurs (as in the MUST collocation checker) to help them to understand the correct use of it.

4 Experiments

In order to validate our model, we carried out experiments on Spanish material, taking as starting point some miscollocations with the bases *siesta* ‘nap’ (Table 1), *meta* ‘finish line , target, objective, goal, goalkeeper’ (Table 3), *examen* ‘exam’ (Table 2) , and *teléfono* ‘phone’ (Table 4) from the learner corpus CEDEL2 (Lozano, 2009). As reference treebank, we use a treebank of Spanish newspaper material. The sentences of the treebank are indexed in a Solr (Lucene) index for more efficient access. The index allows us to retrieve directly the list of tokens, the list of lemmas and all the tokens related with another token by a given relation type. It is also used to retrieve examples to be shown to the user.

The table of each base lists the most common collocates, indicating the frequencies in the corpus, the PMI, $NPMI_{CB}$ and $NPMI_C$; for *meta* the table contains two parts, a list for *meta* acting as direct object and another for *meta* being the subject. That is, the tables can be considered as ranked lists of collocation suggestions.

collocate	$Freq_C$	$Freq_{CB}$	PMI	$NPMI_{CB}$	$NPMI_C$
dormir	612	69	3,611	0,716	0,881
echar	5847	42	2,415	0,459	0,775
hacer	165124	18	0,597	0,106	0,358
estar	13349	3	0,911	0,142	0,33
haber	149464	9	0,339	0,057	0,198

Table 1: Table of PMI and normalized PMIs for the base *siesta*.

$collocate$	$Freq_C$	$Freq_{CB}$	PMI	$NPMI_{CB}$	$NPMI_C$
pasar	23170	228	1,69	0,373	0,671
superar	19861	119	1,475	0,307	0,57
aprobar	12676	82	1,508	0,303	0,542
realizar	28999	99	1,231	0,252	0,508
suspender	6241	32	1,407	0,262	0,455
preparar	11526	35	1,18	0,221	0,418
someter	1927	11	1,454	0,249	0,404
hacer	165124	139	0,623	0,131	0,373
ordenar	6494	15	1,061	0,186	0,345
terminar	4906	11	1,048	0,179	0,328
efectuar	4909	11	1,048	0,179	0,328
practicar	5091	11	1,032	0,177	0,325
afrontar	8994	11	0,785	0,134	0,268

Table 2: Table of PMI and normalized PMIs for the base *examen*.

The miscollocations have been entered in sequence via an interface comparable to the MUST collocation checker (http://miscollocation.appspot.com/).

$collocate$	$Freq_C$	$Freq_{CB}$	PMI	$NPMI_{CB}$	$NPMI_C$
cruzar^{1}	4608	223	2,442	0,538	0,758
alcanzar^{1}	25412	227	1,708	0,377	0,689
fijar^{3}	8803	43	1,446	0,275	0,492
batir^{2}	2066	17	1,672	0,296	0,468
perforar^{2}	217	5	2,12	0,343	0,466
lograr$^{2\ 3}$	21357	50	1,126	0,217	0,441
conseguir	25486	50	1,05	0,202	0,423
regatear^{2}	342	3	1,7	0,265	0,391
encarar^{2}	1200	6	1,456	0,238	0,383
inquietar^{2}	520	3	1,518	0,237	0,364
marcar^{2}	15636	26	0,978	0,179	0,363
cumplir^{3}	19295	26	0,887	0,162	0,341
perseguir^{3}	3462	8	1,121	0,187	0,335
rebasar^{3}	958	3	1,253	0,195	0,321
conquistar^{3}	2033	5	1,148	0,186	0,321
$collocate$	$Freq_C$	$Freq_{CB}$	PMI	$NPMI_{CB}$	$NPMI_C$
ser^{1,2,3,4}	784559	388	0,558	0,13	0,564
desviar^{4}	532	6	1,917	0,314	0,461
parar^{4}	1999	6	1,342	0,22	0,374
salvar^{4}	2045	6	1,332	0,218	0,373
alcanzar^{3,4}	11168	15	0,992	0,174	0,35
estar^{1,2,3,4}	171062	80	0,534	0,107	0,323
fijar^{3}	3394	5	1,033	0,167	0,308
tocar^{4}	3227	4	0,958	0,152	0,284

Table 3: Table of PMI and normalized PMIs for the base *meta*. The upper part of the table captures the figures for *meta* as direct object and the lower part for *meta* as subject. Each collocate corresponds to a given sense of *meta*: ‘^{1}’ stands for ‘the finish line’ , ‘^{2}’ for ‘goal’ in football, ‘^{3}’ for ‘objective’ and ‘^{4}’ for ‘goalkeeper’.

$collocate$	$Freq_C$	$Freq_{CB}$	PMI	$NPMI_{CB}$	$NPMI_C$
pinchar	403	77	2,789	0,558	0,652
descolgar	230	61	2,931	0,575	0,648
sonar	2218	105	2,183	0,449	0,617
coger	4627	123	1,932	0,403	0,6
intervenir	1294	55	2,136	0,415	0,566
colgar	1723	61	2,057	0,403	0,564
llamar	12957	111	1,44	0,298	0,519
desconectar	234	14	2,285	0,398	0,506
contestar	3066	41	1,634	0,31	0,481
atender	8563	57	1,331	0,259	0,451
usar	6286	46	1,372	0,263	0,444
habilitar	760	14	1,773	0,309	0,443

Table 4: Table of PMI and normalized PMIs for the base *teléfono*.

5 Discussion

In the tables, we can appreciate the differences produced by the different measures used to calculate the co-occurrence strength between the collocation elements. The main differences occur with verbs that are very common, such as *hacer* '[to] make' as collocate of *siesta* or *examen*. In both cases, $NPMI_C$ considers it more important than other verbs. Thus, PMI and $NPMI_{CB}$ rank *hacer* in co-occurrence with *examen* as lowest (0.623 respectively 0,131), while $NPMI_C$ keeps it in the middle of the table, ranking it higher than *efectuar* '[to] effect' or *afrontar* '[to] face', which are much less common (which makes the $NPMI_C$ ranking more appropriate).

In co-occurrence with the base *teléfono,* the verb *llamar* '[to] call' appears in the middle of the list when ranked by $NPMI_C$, while when ranked by PMI or $NPMI_{CB}$ it appears down in the least, even if it is almost the most common collocate of phone (5% of the cases).

Table 3 shows that the verb *ser* '[to] be' is the most common (33 of the cases) for *meta* as subject; the $NPMI_C$ upgrades it, ranking it higher, even if is a very common verb and has a low PMI. Looking at the list, we see that there is also *estar* '[to] be' with similar PMI. However, $NPMI_C$ does not promote it. Analyzing the data more deeply, we can observe that $p(meta|ser) \sim p(meta|estar)$, but $p(ser|meta) \gg p(estar|meta)$. That is, the penalization of *estar* by $NPMI_C$ is correct.

The tables show the relationship between the base and the collocate when the collocate is a verb and the base its direct object. When the base is subject (or a different kind of object), different collocations may appear. Table 3 shows that when the base *meta* has a different grammatical function in the sentence, it often also has a different sense. It tends to mean 'goal', 'finish line' or 'objective' when is the direct object, but when appears as subject it often stands for 'goalkeeper'.

There are some coincidences between the two lists of verbs in Table 3. This is because of their use as both passive and active. For the moment, we do not make any distinction between passive and active forms.

6 Related Work

A number of works deal with detection of miscollocations and collocation error correction in learners' writings. However, only a few allow for the validation of isolated word co-occurrences with respect to their collocation status and provide ranked lists of correction suggestions. One of them is (Chang et al., 2008). They check a V–N co-occurrence provided by a learner against a collocation list obtained before from a reference corpus. Co-occurrences not found in this collocation list are variegated in that their verbal elements are substituted by all English translations of their L1 (Chinese, in this case) counterpart in an electronic dictionary. The variants are again matched against the collocation list. The finally matching co-occurrences that contain the noun of a non-matching co-occurrence are offered as correction suggestions. The Mutual Reciprocal Rank (MRR) of the correction list is reported to reach 0.66.

Dahlmeier and Ng (2011), who deal with the detection of miscollocations in writings also exploits L1 interference in learners. They work with confusion sets of semantically similar words. Given an input text in L2, they generate L1 paraphrases, which are then looked up in a large parallel corpus to obtain the most likely L2 co-occurrences. For this strategy, they report a precision of 38%.

Futagi et al. (2008) target the detection of miscollocations in learner texts, leaving the correction aside. Unlike the above proposals, they are not restricted to V+N co-occurrences. But similar to (Chang et al., 2008), they extract the co-occurrences from a learner text, variegate them and then look up the original co-occurrence and its variants in a reference list to decide on its status. To obtain the variants, they apply spell checking, vary articles and inflections and use WordNet to retrieve synonyms of the collocate. Wu et al. (2010) use a classifier to provide a number of collocate corrections. The classifier takes the learner sentence as lexical context. The probability predicted by the classifier for each suggestion is used to rank the suggestions. According to the evaluation included in (Wu et al., 2010), an MRR of 0.518 for the first five correction suggestions has been achieved. Liu et al. (2009) retrieve miscollocation correction suggestions from a reference corpus using three metrics: (i) mutual information (Church and Hanks, 1989), (ii) semantic similarity of an incorrect collocate to other potential collocates based on their distance in WordNet, and (iii) the membership of the incorrect collocate with a potential correct collocate in the same “collocation cluster”'. A combination of (ii)+(iii) leads to the best precision achieved for the suggestion of a correction: 55.95%. A combination of (i)+(ii)+(iii) leads to the best precision of 85.71% when a list of five possible corrections is returned.

Ferraro et al. (2014) suggest a two stage strategy for correction of miscollocations in Spanish. The first stage is rather similar to the one proposed by Futagi et al. (2008): it retrieves the synonyms of the collocate in the miscollocation in question from a number of auxiliary resources (including thesauri, bilingual L1-L2 dictionaries, etc.) and combines them with the base of the miscollocation to candidate corrections. The candidate corrections that are valid collocations of Spanish are returned as correction suggestions. If none of them is, the second stage applies a metric to retrieve correction suggestions. Three metrics have been experimented with: affinity metric, lexical context metric and context feature metric. The context feature metric, which uses the contextual features of the miscollocation (tokens, PoS tags, punctuation, grammatical functions, etc.), performed best in that it achieved an MRR of the top five suggestions of 0.72.

However, none of the above mentioned works considers in detail the asymmetric nature of collocations as captured by $NPMI_C$ and none of them takes into account that the co-occurrence strength between tokens (as captured by the (normalized) PMIs) needs to be calculated differentiating between different dependency relations and different sub-categorization frames.

7 Conclusions

In this paper, we have shown that the asymmetric nature of collocations requires an “asymmetric" normalization of the commonly used PMI measure and that any co-occurrence measure should be applied to co-occurrences of the same syntactic profile, i.e., with the same syntactic dependency relation and the same sub-categorization frame. The consideration of these characteristics of collocations allows for a more accurate ranking of correction suggestions for miscollocations.

Acknowledgments

The work presented in this paper has been supported by the Spanish Ministry of Economy and Competitiveness under the contract number FFI2011-30219-C02-02 in the framework of the HARENES Project, carried out in collaboration with the DICE Group of the University of La Coruña.

References

Alonso Ramos, M., Wanner, L., Vincze, O., Casamayor, G., Vázquez, N., Mosqueira, E., and Prieto, S. (2010). Towards a motivated annotation schema of collocation errors in learner corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 3209–3214, La Valetta, Malta.

Benson, M. (1989). The structure of the collocational dictionary. *International Journal of Lexicography*, 2(1):1–13.

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

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

Chang, Y., Chang, J., Chen, H., and Liou, H. (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.

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

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

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

Dahlmeier, D. and Ng, H. (2011). Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Edinburgh, Scotland.

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

Ferraro, G., Nazar, R., Ramos, M. A., and Wanner, L. (2014). Towards advanced collocation error correction in Spanish learner corpora. *Language Resources and Evaluation*, 48(1):45–64.

Firth, J. (1957). Modes of meaning. In Firth, J., editor, *Papers in Linguistics, 1934-1951*, pages 190–215. Oxford University Press, Oxford.

Futagi, Y., Deane, P., Chodorow, M., and Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21(1):353–367.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and Formulae. In Cowie, A., editor, *Phraseology: Theory, Analysis and Applications*, pages 145–160. Oxford University Press, Oxford.

Halliday, M. (1961). Categories of the theory of grammar. *Word*, 17:241–292.

Hausmann, F.-J. (1984). Wortschatzlernen ist kollokationslernen. zum lehren und lernen französischer wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(1):395–406.

Lesniewska, J. (2006). Collocations and second language use. *Studia Lingüística Universitatis lagellonicae Cracoviensis*, 123:95–105.

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

Liu, A. L.-E., Wible, D., and Tsao, N.-L. (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.

Lozano, C. (2009). CEDEL2: Corpus escrito del español L2. In Bretones Callejas, C., editor, *Applied Linguistics Now: Understanding Language and Mind*, pages 197–212. Universidad de Almería, Almería.

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

Nesselhauf, N. (2004). How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In Aston, G., Bernardini, S., and Stewart, D., editors, *Corpora and language learners*, pages 109–124. Benjamins Academic Publishers, Amsterdam.

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

Orol, A. and Alonso Ramos, M. (2013). A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish. *Procedia–Social and Behavioural Sciences*, 96:563–570.

Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57, Marrakech.

Wanner, L., Verlinde, S., and Alonso Ramos, M. (2013). Writing assistants and automatic lexical error correction: word combinatorics. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., and Tuulik, M., editors, *Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 conference*, pages 472–487, Tallinn & Ljubljana. Trojina, Institute for Applied Slovene Studies & Eesti Keele Instituut.

Wu, J.-C., Chang, Y.-C., Mitamura, T., and Chang, J. (2010). Automatic collocation suggestion in academic writing. In *Proceedings of the ACL Conference, Short paper track*, Uppsala.