A comparison of statistical association measures for identifying dependency-based collocations in various languages.

Marcos Garcia
Universidade da Coruña
Grupo LyS, Dpto. de Letras
Campus da Zapateira, Coruña
Universidade da Coruña, CITIC
Campus de Elviña, Coruña

Marcos García-Salido
Universidade da Coruña
Grupo LyS, Dpto. de Letras
Campus da Zapateira, Coruña

Margarita Alonso-Ramos
Universidade da Coruña
Grupo LyS, Dpto. de Letras
Campus da Zapateira, Coruña
Universidade da Coruña, CITIC
Campus de Elviña, Coruña

{marcos.garcia.gonzalez,marcos.garcias,margarita.alonso}@udc.gal

Abstract

This paper presents an exploration of different statistical association measures to automatically identify collocations from corpora in English, Portuguese, and Spanish. To evaluate the impact of the association measures we manually annotated corpora with three different syntactic patterns of collocations (*adjective-noun*, *verb-object* and *nominal compounds*). We took advantage of the PARSEME 1.1 Shared Task corpora by selecting a subset of $155k$ tokens in the three referred languages, in which we annotated $1,526$ collocations with their Lexical Functions according to the Meaning-Text Theory. Using the resulting gold-standard, we have carried out a comparison between frequency data and several well-known association measures, both symmetric and asymmetric. The results show that the combination of dependency triples with raw frequency information is as powerful as the best association measures in most syntactic patterns and languages. Furthermore, and despite the asymmetric behaviour of collocations, directional approaches perform worse than the symmetric ones in the extraction of these phraseological combinations.

1 Introduction

Although there is no agreement about the linguistic properties of collocations, it is commonly accepted that the automatic identification of this type of multiword expressions (MWEs) is crucial for many natural language processing tasks such as natural language understanding, or machine translation (Sag et al., 2002; Wehrli and Nerima, 2018).

From a statistical point of view collocations are recurrent co-occurrences of word pairs given a short span of text (Firth, 1957; Benson, 1990; Sinclair, 1991). Thus, they are often identified by applying association measures (AMs, e.g., log-likelihood, pointwise mutual information, etc.) on co-occurrence counts in windows of different sizes (Pecina, 2010). However, the phraseological tradition states that collocations are idiosyncratic asymmetric combinations of syntactically related pairs of words (Hausmann, 1989; Benson, 1989). In this regard, their asymmetry derives from the fact that one of the elements of a collocation (the BASE, e.g., *cab* in *take a cab*) is freely selected due to its meaning, while the choice of the other (the COLLOCATE, e.g., *take* in the previous example) is restricted by the former (Mel'čuk, 1995, 1998). Following this perspective, the process for extracting collocations should take advantage of syntactic parsing (Seretan, 2011). Moreover, and with a view to capture the asymmetry of these expressions, directional AMs have been proposed (Carlini et al., 2014). To evaluate the impact of each extraction method, some researchers perform a manual revision of a ranked list of collocation candidates (Seretan and Wehrli, 2006), while others collect a set of gold-standard collocations (from corpora or dictionaries) to evaluate their identification methods (Krenn and Evert, 2001; Pearce, 2002; Pecina, 2010; Evert et al., 2017).

Notwithstanding, most studies focus only on one language or just on a collocation pattern, and most of them use very different gold-standards (e.g., considering idioms or proper nouns as a type of collocations), so that their results are not comparable and cannot be generalized to other languages or collocational schemes.

This paper presents a systematic evaluation of twelve AMs —both symmetric and directional— which have been proposed for collocation extraction. The experiments are carried out us-

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 49–59
Florence, Italy, August 2, 2019.

ing three syntactic patterns (*adjective-noun*, *verb-object*, and *nominal compounds*) in English, Portuguese, and Spanish. To obtain accurate recall and precision values, we have created gold-standard corpora containing $1,526$ collocations labeled in context in these languages.[1] The annotation was performed following a phraseological approach, which not only identifies each collocation but also classifies it according to a lexical function in the Meaning-Text Theory (Mel'čuk, 1998).

The results of the performed experiments show that, to extract these dependency-based collocations, frequency data behaves similarly to the best association measures, and that directional measures obtain worse results than symmetric ones. Moreover, these findings are general tendencies in the three languages that have been evaluated.

The rest of this paper is organized as follows. First, Section 2 introduces some related work on the use of AMs for extracting collocations. Then, we briefly present the gold-standard corpora in Section 3. The evaluation and discussion of the results are addressed in Sections 4 and 5, while some conclusions are drawn in Section 6.

2 Related Work

There is a rich variety of studies dealing with the automatic extraction of collocations from corpora. In this respect, several papers addressed this task applying different AMs to short sequences of ngrams (Smadja, 1993) or syntactic dependencies (Lin, 1999; Seretan and Wehrli, 2006). Other studies, such as Krenn and Evert (2001) and Evert and Krenn (2001) took advantage of POS-tags to focus on particular collocational patterns.

Pearce (2002) compared previous statistical approaches to identify collocational bigrams, showing that the different definitions of collocations involve divergences in the results. In this respect, papers such as Thanopoulos et al. (2002) include named entities in some of the gold-standards.

Pecina and Schlesinger (2006) and Pecina (2010) carry out a large comparison of dozens of statistical metrics to identify collocations (including idioms) in Czech corpora, also proposing several combinations of AMs which improve the performance of single measures. A recent comparison of various AMs, using different corpus sizes and two different gold-standards in English can be found in Evert et al. (2017), which also evaluate surface-based and dependency-based approaches. In Uhrig et al. (2018), the authors analyze the impact of several dependency parsers and syntactic schemes in the same task.

The asymmetric properties of collocations have been taken into account, for instance, in Gries (2013), which proposed directional measures (ΔP) to better capture the behaviour of these expressions. Correspondingly, and following an approach similar to ours, Carlini et al. (2014) propose another asymmetric measure ($NPMI_c$) based on a normalized version of mutual information (Church and Hanks, 1990).

A related task consists of automatically classifying the semantic properties of collocations by means of lexical functions or glosses. In this regard, some studies apply machine learning methods to train classifiers (Wanner et al., 2006, 2016; Gelbukh and Kolesnikova, 2012), while others use distributional semantics to identify a collocate given a base and a lexical function (Rodríguez-Fernández et al., 2016).

Among the many studies that evaluate AMs to extract collocations from corpora, most of them focus only on one language (usually in English, German, Czech, or Spanish) and use different approaches (surface-based or syntactic dependencies). Moreover, different interpretations of collocations make most studies not comparable. Taking the above into account, our evaluation is carried out in a new dataset in three languages and with different syntactic patterns, which has been manually annotated from a phraseological viewpoint following the Meaning-Text Theory.

3 Gold-standard multilingual corpora of collocations.

This section summarizes the annotation process of the corpora and its results.[2] Before that, we introduce the main characteristics of collocations in the phraseological viewpoint adopted in this paper.

3.1 Collocations

As said, we understand collocations as asymmetric combinations of two syntactically related lexical units (Hausmann, 1989). In this regard, one of the elements that form the collocation (the base) is chosen by the speaker due to its meaning. The base, in turn, restricts the selection of the other lex-

[1] https://github.com/marcospln/collocations

[2] See Garcia et al. (2019) for a detailed explanation of the annotation process.

ical unit (the collocate), which conveys a particular meaning in function of a given base.

Under the Meaning-Text Theory, the concept of collocation was formalized as follows:

> A COLLOCATION **AB** of **L** is a semantic phraseme of **L** such that its signified 'X' is constructed out of the signified of the one of its two constituent lexemes –say, of **A**– and a signified 'C'['X'='A⊕C'] such that the lexeme **B** expresses 'C' contingent on **A**.

where **A** is the base, **B** the collocate, **L** a language, and 'C' and 'X' the meanings of the collocate (in this context) and of the collocation, respectively. From this perspective, lexical restrictions are more important than the co-occurrence frequency of the combination, which does not play any role in this definition (Mel'čuk, 1998). This theory also proposes the concept of Lexical Functions (LFs), a tool to represent the relation between the base and a set of potential collocates which convey a given meaning (Wanner, 1996; Mel'čuk, 1996). Thus, using the *Magn* LF (which means 'intensification') we could define *Magn(breath)=deep* and *Magn(effort)=great* to respectively represent the collocations *deep breath* and *great effort*.

3.2 Source corpora and annotation

To create our multilingual gold-standard corpus we used three subcorpora of the PARSEME Shared Task 1.1 (Ramisch et al., 2018). In this respect, some of the MWEs labeled by the PARSEME community (namely the light-verb constructions, LVCs) actually intersect with our objectives, so we took advantage of these annotations. We selected the *test* splits for Portuguese and Spanish ($58k$ and $39k$ tokens, respectively), and the *train* dataset for English (with $53k$ tokens). These resources are annotated with Universal Dependencies (Nivre, 2015).

Guidelines: We defined specific guidelines for each collocation type following Mel'čuk (1995). Besides, we attempted to be compatible with the PARSEME guidelines with a view to combining both annotations. Since we use dependency parsing to retrieve candidate collocations, we annotated the following three syntactic patterns, here exemplified with some of the included LFs:

Verb-object (*obj*): this collocation type refers to predicative nouns depending of verbs which do not contribute to the meaning of the combination ($Oper1$: *fazer aparição*; '[to] appear' in Portuguese), express causation ($CausOper1$: *conceder autorización*; '[to] give permission' in Spanish) or a particular meaning with this specific base ($NonStandard$: *[to] shake hands*). Most of these cases were covered by the LVC category in the PARSEME guidelines, so besides annotating some new collocations, we revised each LVC and added their LFs.

Adjective-noun (*amod*): in these collocations the adjective may express the meaning of 'intensification' ($Magn$: *excelente calidad*, 'excellent quality' in Spanish), or 'attenuation' ($AntiMagn$: *baixo rendimento*; 'low performance' in Portuguese), convey a positive or negative evaluation of the speaker (Bon: *great film*, $AntiBon$: *dura realidade*; 'harsh reality' in Portuguese), or have a specific sense when modifying the noun ($NonStandard$: *agua dulce*; 'fresh water' in Spanish).

Nominal compounds (*nmod* and *compound*): in a nominal compound, the head of the relation may express the concept of 'head of a collective' (Cap: *police chief*), 'a part of' ($Sing$: *membro [do] grupo*; 'group member' in Portuguese), or of a 'set' or the 'totality' of the dependent ($Mult$: *ramo de rosas*; 'bouquet of roses' in Spanish).

Procedure: To carry out the annotation process, we first extracted every instance of the target relations (*obj*, *amod*, *nmod*, and *compound*) from the source corpora, then organized as *base;collocate;relation* lemmatized triples. This process generated $12,496$ candidates ($\approx 5k$ *amod*, $\approx 3,5k$ *nmod/compound*, and $\approx 4k$ *obj*).

Using these data, all the triples were arranged into nine sheets (one per language and dependency relation) including for each candidate a link to an automatically generated HTML page with actual examples from the corpora.

Then, a group of three experts revised the candidates on the sheets, classifying each combination as *collocation*, *non-collocation*, or *doubt*. After that, a final sheet for each relation and language was generated, with the most frequent label for each combination. The dubious cases (those with more than one doubt, or with total disagreement

between the annotators) were revised and classified by the whole team of language experts.

Finally, the gold annotations were added to the initial corpora and transferred to WebAnno (Eckart de Castilho et al., 2016). Then, we used this tool to correct some special cases (e.g., collocations including internal MWEs) and to perform a general revision of the corpora before converting the WebAnno data into the final *.conllu* format.

3.3 Final resources and results

From the initial $\approx$ $12.5k$ unique dependency triples, the annotation process yielded a total of $1,394$ collocations ($\approx$ 11%). The *amod* pattern was the most productive one (620 different examples) followed by *obj* (579). Nominal compounds were the less frequent type, with 195 labeled collocations. The multi-k inter-annotator agreement (Davies and Fleiss, 1982) produced values between $k = 0.37$ (*nmod*) and $k = 0.71$ (*obj*). During the annotation process, 447 combinations were marked as doubt, and out of these, 260 were finally considered collocations by the language experts. Even if we do not make explicit use of lexical functions in this paper, it is worth mentioning that the collocations in these final corpora are labeled using 60 LFs, which may be useful to evaluate extraction and classification strategies (Wanner et al., 2016; Rodríguez-Fernández et al., 2016; Kolesnikova and Gelbukh, 2018).

4 Evaluation

This section describes the experiments carried out to evaluate the performance of the different AMs in our gold-standard corpora.

4.1 Data

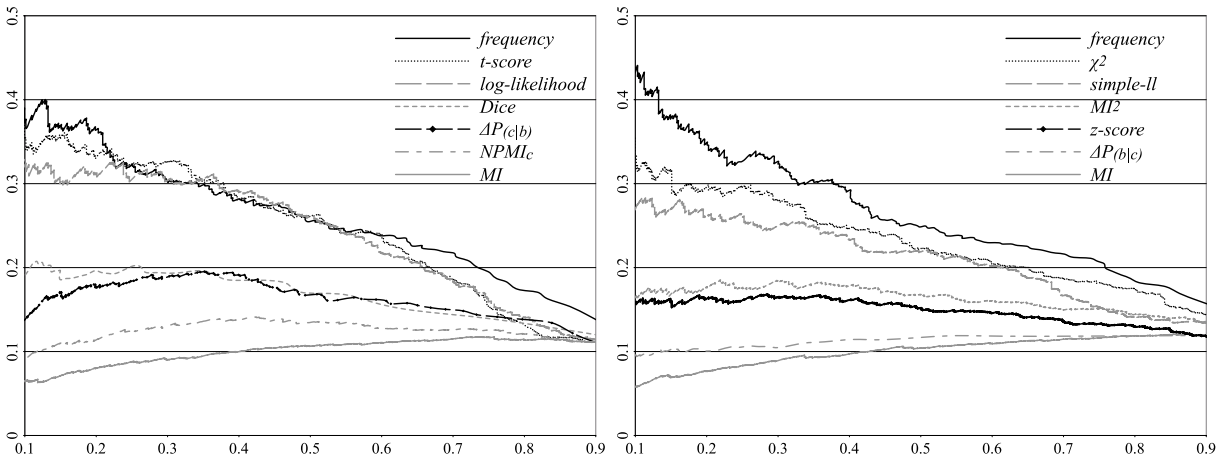
From the above presented gold-standard corpora we used the $12.5k$ dependency triples of the three syntactic patterns as testing data to assess the impact of different AMs. The labeled collocations were used as true positives and the rest of the examples as true negatives. Since the size of our data ($\approx$ $155k$) is not enough to extract statistical data for the computation of suitable association values, we compiled three corpora (one per language) with about 100 million tokens each. These reference corpora were used to obtain the statistical values of the 12.5 triples. With a view to obtain comparable results, we created these corpora in an analogous way. Thus, each of them contains 50 million tokens from Wikipedia, 20 million from the Europarl corpus (Koehn, 2005), 10 million from OpenSubtitles (Lison and Tiedemann, 2016), and a set of 20 million tokens formed by news, web pages, and small corpora from the Universal Dependencies 2018 and PARSEME 1.1 shared tasks (Zeman et al., 2018; Ramisch et al., 2018). The texts were tokenized, PoS-tagged and lemmatized by LinguaKit (Gamallo et al., 2018), and parsed by UDPipe, a state-of-the-art dependency parser based on neural networks (Straka and Straková, 2017). We used the Universal Dependencies formalism, which yielded the best results in a similar comparison (Uhrig et al., 2018), training the models with the 2.3 version of the UD treebanks (Nivre et al., 2018).

4.2 Experiments

Besides raw frequency, we have evaluated eleven association measures which have been used for both dependency and ngram-based collocation extraction. As symmetric measures we used simple-ll, t-score, z-score, (pointwise) mutual information (MI), MI^2, Dice, log-likelihood, and χ^2 (Evert, 2008; Pecina, 2010). Also, we have included two directional AMs which have been proposed to model the asymmetry of collocations (see Section 3.1): $DeltaP$ (Gries, 2013) in both directions ($\Delta\text{P}_{(base|collocate)}$, and $\Delta\text{P}_{(collocate|base)}$), and $NPMI_c$ (Carlini et al., 2014). See Tables 3 and 4 in Appendix A for the equations.

For each language and collocation pattern, we computed precision and recall values for every AM and plotted them into two dimensional *precision–recall* (PR) curves. PR curves allow us to compare the performance of the different measures, by looking at those curves closer to the top-right corner. Figure 1 includes two examples of different PR curves in English and Portuguese (where x-axis is recall and y-axis precision). These graphics are useful to rapidly observe those measures that are clearly better than others (i.e., they have higher precision in most recall values), but the visualization may be ambiguous if the curves cross each other along the plot (in those AMs which are better than others only in specific recall intervals).

To provide comparable results for the different scenarios we computed two single values in each experiment: *area under curve* (AUC), which measures the area below each PR curve (Davis



(a) Micro-average PR curves in the English corpus.

(b) Micro-average PR curves in the Portuguese corpus.

Figure 1: Precision-recall (y- and x-axis) curves for different AMs in English and Portuguese. Except for the best and worse measures (*frequency* and *MI*), Figures 1a and 1b include different AMs to facilitate the visualization.

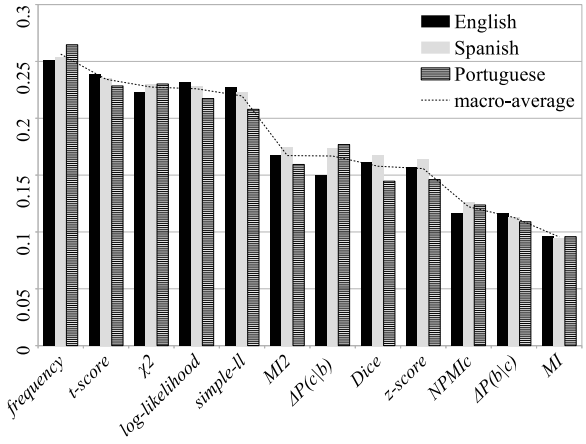


Figure 2: *Area under curve* results (micro-average) for each language.

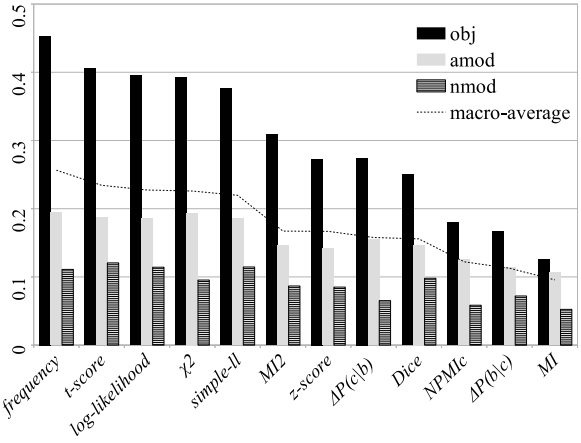


Figure 3: *Area under curve* results (macro-average using the data of the three languages) for each dependency relation.

and Goadrich, 2006), and *mean average precision* (MAP), which represents the mean of the precision in each recall value (Pecina and Schlesinger, 2006; Pecina, 2010). Following Pecina we computed MAP in the recall interval $\langle 0.1, 0.9 \rangle$.

First, we will show the micro-average results of each AM per language, followed by the results for each dependency relation. Finally, we will also present the AUC and MAP values for each language and relation.

4.3 Results

As mentioned, Figure 1 contains the *precision-recall* curves for different AMs in English (1a) and Portuguese (1b). To guarantee a proper visualization we included only seven AMs in each plot: in both cases we drew the best and worse measures (frequency and MI, respectively), and five different curves for English and Portuguese.

In both cases, the best results were obtained by those AMs which promote recurrent combinations, such as the raw frequency or t-score, and the lowest ones by mutual information (which tends to assign high values to low-frequency candidates). A deeper analysis of the curves in the three languages allowed us to define three groups of AMs: (i) those with better PR curves, including frequency, t-score, log-likelihood, χ^2, and simple-ll; (ii) another set with intermediate values (MI^2, z-score, $\Delta P_{(c|b)}$, and Dice); and (iii) three measures which produced lower results in most cases: $NPMI_c$, $\Delta P_{(b|c)}$, and MI. Even if this classification varies in some scenarios, the average results in

the three languages seem to confirm this tendency.

In Figure 2 it can be seen the micro-average AUC results for each language, with a clearer and more comparable visualization of the behaviour of the evaluated AMs in each language.

The next experiment was carried out to compare the performance of the AMs in each dependency relation. Figure 3 contains the AUC results for each pattern. On the one hand, these values show that the results are quite different in each dependency relation. In this respect, *verb-object* combinations were those more accurately extracted, while the quality of the *nmod* extractions were much lower than both *obj* and *amod* (with intermediate results). On the other hand, Figure 3 also shows that the previously referred groups of AMs follow the same tendency in this evaluation as well.

Finally, Tables 1 and 2 display the AUC and MAP values for each relation and language. Overall, the AUC results follow the above mentioned tendencies. For *verb-object* collocations raw frequency obtained the best results in the three languages, followed by χ^2 (in Portuguese and Spanish), and by t-score in English. In *amod* and *nmod* patterns, however, there are some differences in those measures with higher results. Thus, *amod* candidates were better ranked by simple-ll in English, while χ^2 was the best AM in the Portuguese data. For *nmod*, frequency, t-score, and simple-ll obtained the best numbers in English, Portuguese, and Spanish, respectively. Apart from these variations in the highest-ranked measures, it is worth mentioning that the Dice coefficient had better results than χ^2 in the classification of *nmod* combinations in Portuguese and Spanish.

The mean average precision (Table 2) produced a different ranking of the AMs with regard to the previous evaluation. First, it is important to note that Dice has the best macro-average MAP values, achieving the first position in three scenarios (*obj* collocations in Portuguese and Spanish, and *nmod* in Spanish). Raw frequency performed better than Dice in five cases, but with low results in Portuguese with the *obj* dataset. Even though the best measures are basically the same (with the exception of Dice), the intermediate and lowest results differ from the AUC values: z-score gets worse macro-average numbers than the other measures, while MI obtains better results, specially in *obj* in Portuguese and Spanish. Regarding the directional measures, both ΔP variants get lower results, while $NPMI_c$ shows a good performance when compared to its AUC numbers.

5 Discussion

To make a better interpretation of the previous figures, this section discusses the most interesting results provided by the performed experiments.

First, when compared to the values provided by other studies, the low precision obtained in our experiments is striking. As an example, MAP values in Pecina (2010) surpass 0.65, while the best average results in our tests were of about 0.30. In this regard, we consider that the different concept of collocation of both studies lead to critical differences in the results, as pointed in Section 2. For instance, the annotators of the referred study (Pecina, 2010) considered collocations some expressions not covered by our data, such as idioms, phrasal verbs, or terms, with more than 20% of true collocations (*versus* the 11% of our corpus). Other papers which use diverse corpora also obtain different results depending on the gold-standard (Evert et al., 2017). Apart from that, and as Figure 2 and Tables 1 and 2 refer, there are evident differences among the collocation patterns, so direct comparisons should take this fact into account. Specially in *nmod*, the low results might be due to our restrictive annotation guidelines, which caused that only 5.5% of the candidates were labeled as collocations (*versus* 14.6% and 12.6% in *obj* and *amod*, respectively). As pointed out in Section 3.3, the inter-annotator agreement in this particular relation was also the lowest one.

Despite the divergences among the dependency relations, the average results per language did not show evident variations with respect to the evaluated AMs. Small differences occur, however, inside each of the three mentioned groups. For instance, log-likelihood and MI^2 work better, respectively, than χ^2 and $\Delta P_{(c|b)}$ in English, while Portuguese had the opposite tendencies in both cases.

With regard to the directional measures, our experiments showed that, in spite of the asymmetric structure of collocations, symmetric measures produced, on average, better results. The low values of $\Delta P_{(b|c)}$ are somehow expected because this AM encodes the directionality from the collocate to the base, and not the other way around as the theoretical descriptions of collocations propose. However,

	English			Portuguese			Spanish			*macro*
	obj	*amod*	*nmod*	*obj*	*amod*	*nmod*	*obj*	*amod*	*nmod*	*avg*
frequency	**0.470**	0.215	**0.094**	**0.424**	0.198	0.123	**0.507**	**0.172**	0.128	**0.259**
t-score	0.415	0.228	0.092	0.373	0.171	**0.134**	0.461	0.168	0.146	0.243
log-likelihood	0.403	0.228	0.084	0.357	0.172	0.122	0.465	0.160	0.149	0.238
χ^2	0.372	0.216	0.071	0.374	**0.202**	0.103	0.466	0.167	0.120	0.232
simple-ll	0.385	**0.230**	0.085	0.334	0.170	0.121	0.455	0.159	**0.150**	0.232
MI^2	0.307	0.184	0.048	0.270	0.129	0.098	0.411	0.125	0.125	0.189
z-score	0.266	0.183	0.046	0.239	0.123	0.096	0.374	0.122	0.125	0.175
Dice	0.242	0.199	0.060	0.213	0.127	0.105	0.356	0.122	0.138	0.174
$\Delta P_{(c\|b)}$	0.230	0.180	0.027	0.288	0.144	0.082	0.313	0.147	0.090	0.167
$NPMI_c$	0.148	0.160	0.023	0.189	0.111	0.071	0.212	0.108	0.088	0.123
$\Delta P_{(b\|c)}$	0.151	0.151	0.038	0.153	0.107	0.078	0.229	0.086	0.102	0.121
MI	0.105	0.141	0.020	0.129	0.098	0.063	0.151	0.080	0.081	0.096

Table 1: Area Under Curve (AUC) results for each language and collocation pattern, sorted by macro-average. Numbers in bold are the best results of each column.

	English			Portuguese			Spanish			*macro*
	obj	*amod*	*nmod*	*obj*	*amod*	*nmod*	*obj*	*amod*	*nmod*	*avg*
Dice	0.318	0.217	0.078	**0.478**	0.130	**0.131**	**0.540**	0.134	0.157	**0.243**
frequency	**0.496**	0.230	**0.098**	0.245	**0.225**	0.112	0.406	**0.188**	**0.163**	0.240
χ^2	0.310	0.215	0.047	0.300	0.194	0.110	0.403	0.160	0.131	0.208
t-score	0.273	**0.234**	0.057	0.328	0.168	0.100	0.414	0.164	0.114	0.206
log-likelihood	0.280	0.219	0.048	0.334	0.174	0.084	0.352	0.148	0.090	0.192
MI^2	0.246	0.190	0.042	0.256	0.136	0.096	0.358	0.126	0.119	0.174
$NPMI_c$	0.153	0.170	0.025	0.280	0.117	0.110	0.390	0.113	0.131	0.165
simple-ll	0.273	0.220	0.049	0.142	0.173	0.073	0.295	0.147	0.113	0.165
MI	0.110	0.149	0.021	0.305	0.100	0.129	0.391	0.083	0.138	0.158
$\Delta P_{(c\|b)}$	0.275	0.187	0.030	0.193	0.157	0.075	0.217	0.171	0.094	0.156
$\Delta P_{(b\|c)}$	0.190	0.145	0.046	0.218	0.096	0.094	0.292	0.073	0.120	0.142
z-score	0.206	0.189	0.039	0.130	0.128	0.068	0.157	0.123	0.087	0.125

Table 2: Mean Average Precision (MAP) results for each language and collocation pattern, sorted by macro-average. MAP values were computed in the recall interval 0.1–0.9. Numbers in bold are the best results of each column.

both $\Delta P_{(c|b)}$ and $NPMI_c$ achieved low results when compared to symmetric measures such as t-score or log-likelihood. With respect to $NPMI_c$, it is worth pointing that, as in Carlini et al. (2014), its results are better than MI (but lower, however, than its variant MI^2). A qualitative analysis of the data shows, for instance, that $\Delta P_{(c|b)}$ promotes non collocations such as *separate [a] property* (*obj*), *contemporary house* (*amod*), or *freedom interval* (*nmod*), while ranks very low combinations such as the *verb-object* collocations *pay [a] tribute* or *make [a] mistake*, the *adjective-noun wide variety*, or the *nmod* example *cup [of] coffee*.

Apart from the previous observations, the most relevant result when compared to similar evaluations is the impact of raw frequency in the ranking of candidate collocations. In Krenn and Evert (2001) the best values were achieved in most cases by t-score and by frequency, but in other studies such as Evert et al. (2017), frequency-based extractions had only better results than MI (and than ΔP variants in some cases). In our data, however, candidates ranked by frequency were those with the best results (except in a few cases, see Tables 1 and 2). In this respect, Figure 4 shows an overview of the frequency *versus* MI distribution of both collocations and non collocations in our gold-standard. This graph indicates that most collocations have less than $3,000$ occurrences in our reference corpora. More interesting, frequen-

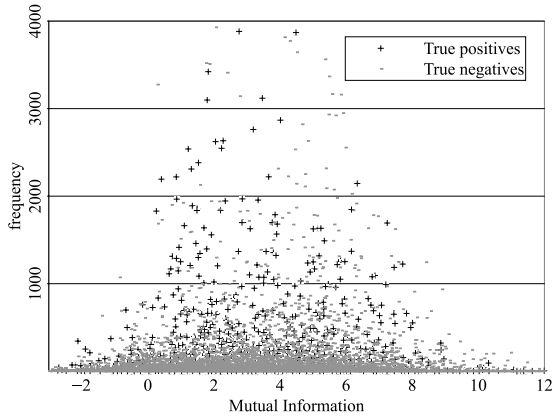


Figure 4: Frequency *versus* mutual information of both collocations and non collocations in the three languages. Statistics are computed using the data from the reference corpora.

cies higher than ≈ 200 are those with a better ratio between true positives and true negatives, while in low frequent combinations most candidates were not considered collocations. When looking at the data, one can see that frequent (non collocational) combinations such as *find [a] way* or *do [a] thing* appear on the top positions, while less recurrent collocations (e.g., *give [a] shrug*, or *crude intensity*) are not promoted due to their low frequency.

In order to delve deeper in the frequency impact we selected, from each reference corpus, the 10 most frequent combinations of each syntactic pattern and classified them into four categories: free combinations, collocations, idioms, or other. In *verb-object* and *adjective-noun* patterns, most candidates were classified as collocations (80% in *obj*, and ≈40% in *amod*, which presents a wider distribution). However, 60% of nominal compounds were classified as free combinations, 30% as idioms, and only 10% as collocations. This brief evaluation shows that, even if the use of raw frequency is not enough to carry out an accurate extraction of collocations, these data are useful to identify some types of collocations. Despite frequency-based extraction (and also other measures which promote recurrent candidates) identifies recurrent free combinations and ignores true but infrequent collocations, those top-ranked candidates are not especially noisy (except for *nmod*), so they may be a good starting point for further annotation. In this respect, it is important to note, as referred by studies such as Lin (1999), Seretan and Wehrli (2006), or Evert (2008), that these observations are especially relevant for dependency-based collocation extraction, which only selects as candidates those pairs of lemmas with a particular POS-tag which are related by specific syntactic relations. Other extraction strategies, using for instance ngrams of tokens, are more sensible to the effects of different AMs, because they often do not include the referred morphosyntactic and syntactic constraints which reduce the the noisy data. It is worth mentioning, however, that other experiments where co-occurrence frequency had a decisive impact also made use of syntactic information by means of constrained lexico-syntactic patterns (Krenn and Evert, 2001; Evert and Krenn, 2001).

In sum, the analyses carried out in this paper point out that frequency information plays an important role in dependency-based collocation extraction. Nevertheless, the results also showed that the precision of both frequency and other AMs is not enough to automate the identification of collocations, so other strategies should be utilized. Finally, our evaluation have also shown that most AMs behave similar in the three evaluated languages, but also that each syntactic pattern reacts differently to the various AMs. In this regard, it would be interesting to apply specific AMs for different relations and frequency folds, aimed at identifying low-frequency cases (Evert and Krenn, 2001). Apart from that, combining different AMs (Pecina and Schlesinger, 2006; Pecina, 2010), and using semantic compositionality to identify idioms and other non collocation candidates might be useful to improve the unsupervised extraction of collocations from corpora (Cordeiro et al., 2019).

6 Conclusions and Further Work

In this paper we have performed an evaluation of the impact of different statistical measures on the automatic extraction of dependency-based collocations in different languages. To carry out these experiments, we annotated gold standard corpora containing $1,394$ unique collocations in English, Portuguese, and Spanish. The annotation was done by means of syntactic dependencies, and each collocation was enriched with its lexical function in the Meaning-Text Theory.

We have compared 12 statistical measures, both symmetric and directional, which have provided interesting results. First, it has been shown that the average performance of each association measure is similar in the three evaluated languages. Second, each dependency-based pattern (specially

nmod combinations) reacts differently with respect to the various measures. Third, the 12 measures can be grouped in three different clusters regarding their behavior in the precision-recall curves. Fourth, in spite of the asymmetric structure of collocations, symmetric measures achieve better results than the directional ones. And finally, the results of our experiments indicate that, in syntax-based collocation extraction, raw frequency performs as well as the best AMs.

The results also confirm that single association measures are not enough to successfully extract collocations from corpora, so further work can be focused on the combinations of statistical information from different measures. In this respect, distributional approaches that automatically classify MWEs regarding its compositionality may be also useful to filter out non collocational expressions from the extracted candidates.

Acknowledgments

This research was supported by a 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation), by Ministerio de Economía, Industria y Competitividad (project with reference FFI2016-78299-P), and by the Galician Government (Xunta de Galicia grant ED431B-2017/01). Marcos Garcia has been funded by a Juan de la Cierva-incorporación grant (IJCI-2016-29598), and Marcos García-Salido by a post-doctoral grant from Xunta de Galicia (ED481D 2017/009).

References

M. Benson. 1989. The Structure of the Collocational Dictionary. *International Journal of Lexicography*, 2(1):1–14.

Morton Benson. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23–34.

Roberto Carlini, Joan Codina-Filba, and Leo Wanner. 2014. Improving collocation correction by ranking suggestions using linguistic knowledge. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 1–12, Uppsala. LiU Electronic Press.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84. The COLING 2016 Organizing Committee.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Mark Davies and Joseph L Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An international handbook*, volume 2, pages 1212–1248. Mouton de Gruyter, Berlin.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France. Association for Computational Linguistics.

Stefan Evert, Peter Uhrig, Sabine Bartsch, and Thomas Proisl. 2017. E-VIEW-affilation–A large-scale evaluation study of association measures for collocation identification. In *Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch*, pages 531–549.

John R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 1–32.

Pablo Gamallo, Marcos Garcia, César Pineiro, Rodrigo Martinez-Castaño, and Juan C Pichel. 2018. LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244. IEEE.

Marcos Garcia, Marcos García-Salido, Susana Sotelo, Estela Mosqueira, and Margarita Alonso-Ramos. 2019. *Pay attention when you pay the bills*. a multilingual corpus with dependency-based and semantic annotation of collocations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence. Association for Computational Linguistics.

Alexander Gelbukh and Olga Kolesnikova. 2012. *Semantic Analysis of Verbal Collocations with Lexical Functions*, volume 414 of *Studies in Computational Intelligence*. Springer.

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

Franz Josef Hausmann. 1989. Le dictionnaire de collocations. *Wörterbücher, Dictionaries, Dictionnaires*, 1:1010–1019.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Olga Kolesnikova and Alexander Gelbukh. 2018. Binary and multi-class classification of lexical functions in spanish verb-noun collocations. In *Advances in Computational Intelligence*, pages 3–14, Cham. Springer International Publishing.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France. Association for Computational Linguistics.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, Maryland, USA. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Igor Mel'čuk. 1995. Phrasemes in language and phraseology in linguistics. In Martin Everaert, Erik-Jan van der Linden, André Schenk, and Rob Schreu, editors, *Idioms: Structural and psychological perspectives*, chapter 8, pages 167–232. Hillsdale: Lawrence Erlbaum Associates.

Igor Mel'čuk. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Language Companion Series*, pages 37–102. John Benjamins Publishing.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony Paul Cowie, editor, *Phraseology. Theory, analysis and applications*, pages 23–53. Clarendon Press, Oxford.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, volume 9041 of *Lecture Notes in Computer Science*, pages 3–16. Springer.

Joakim Nivre et al. 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 651–658, Sydney, Australia. Association for Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. Association for Computational Linguistics.

Sara Rodríguez-Fernández, Luis Espinosa Anke, Roberto Carlini, and Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 499–505. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2276/2010 of *CICLing '02*, pages 1–15, London, UK. Springer-Verlag.

Violeta Seretan. 2011. *Syntax-based collocation extraction*, volume 44 of *Text, Speech and Language Technology*. Springer Science & Business Media.

Violeta Seretan and Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 953–960, Sydney, Australia. Association for Computational Linguistics.

John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press, Oxford.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–178.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Aristomenis Thanopoulos, Nikos Fakotakis, and George Kokkinakis. 2002. Comparative evaluation of collocation extraction metrics. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Peter Uhrig, Stefan Evert, and Thomas Proisl. 2018. Collocation Candidate Extraction from Dependency-Annotated Corpora: Exploring Differences across Parsers and Dependency Annotation Schemes. In *Lexical Collocation Analysis*, Quantitative Methods in the Humanities and Social Sciences, pages 111–140. Springer.

Leo Wanner. 1996. *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Corpus Linguistics*. John Benjamins Publishing.

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

Leo Wanner, Gabriela Ferraro, and Pol Moreno. 2016. Towards Distributional Semantics-Based Classification of Collocations for Collocation Dictionaries. *International Journal of Lexicography*, 30(2):167–186.

Eric Wehrli and Luka Nerima. 2018. *Anaphora resolution, collocations and translation*, volume 341 of *Current Issues in Linguistic Theory*, pages 244–256. John Benjamins.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

A Appendices

	collocate	¬collocate	
base	O	b	$= B$
¬base	c	D	$= d_2$
	$= C$	$= d_1$	$= N$

Table 3: Contingency table for base–collocate combinations. Occurrences are computed only in dependencies with the target syntactic relation.

simple-ll	$2(O \cdot log\frac{O}{E} - (O - E))$
t-score	$\frac{O-E}{\sqrt{O}}$
z-score	$\frac{O-E}{\sqrt{E}}$
MI	$log_2\frac{O}{E}$
MI^2	$log_2\frac{O^2}{E}$
Dice	$\frac{2 \cdot O}{B+C}$
log-likelihood	$2\sum_{ij} O_{ij} log\frac{O_{ij}}{E_{ij}}$
χ^2	$\sum_{ij}\frac{(O_{ij}-E_{ij})}{E_{ij}}$
$NPMI_c^\star$	$\frac{MI}{-log\frac{C}{N}}$
$\Delta P_{(c\|b)}$	$\frac{O}{B} - \frac{c}{d_2}$
$\Delta P_{(b\|c)}$	$\frac{O}{C} - \frac{b}{d_1}$

Table 4: Association measures compared in this this paper. E means expected frequency ($E = \frac{BC}{N}$).

*Following Carlini et al. (2014) $NPMI_c$ was computed using the natural logarithm instead of log_2.