Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 433 – 441

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Outlier Detection in Automatic Collocation Extraction

Octavio Santana Suárez[a], Isabel Sánchez-Berriel[b]*, José Pérez Aguiar[a], Virginia Gutiérrez Rodríguez[b]

[a]*Data Group and Computational Linguistic (GEDLC), University of Las Palmas de Gran Canaria, Edificio Departamental de Informática y Matemáticas, Campus Universitario de Tafira, Las Palmas de Gran Canaria, 35017, Spain*
[b]*Department of Computer and Systems, La Laguna University, Facultad de Matemáticas Campus de Anchieta,La Laguna, 38200*

Abstract

In this paper we have analysed different association measures between words, generally used for the automatic extraction of collocations in textual corpus. Specifically, they have been considered: relative frequency, mutual information, z-score, t-score and Dunning's test. The volume of handled corpus (300000000 words) requires reviewing of the usual approach to this matter, so a solution that is based on methods used to detect statistical outliers is proposed. It is evident from the results that a lot of free combinations extracted with collocations coming from the comparison of words with very different frequencies of use. For this reason, they are applied considering that each word generates a different sample, instead of generating rankings which come from corpus considered as a single sample. The experiment is also performed on a corpus with a much smaller amount of words and the results are reported so contrasted with those obtained with the full corpus. The conclusions and contributions arising give response automatic extraction of collocations from a textual corpus regardless its volume.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: collocations, association measures, outliers.

* Corresponding author. Tel.: +34-922319449; fax: +34-.
E-mail address: isanchez@ull.edu.es

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.463

1. Introduction

The term collocation in this work relates to combinations of words used recursively in a language. This definition, from the linguistic point of view is simple, but the phenomenon focuses on recurrence, which allows automatic extraction of them by processing textual corpus. Examples of Spanish collocations are: *tener apetito* (to have an appetite), *afrontar riesgos* (to take risks), *competir duro* (to compete hard), *conversación animada* (animated conversation)… In this problem should be considered the formal flexibility of the elements in the collocation, since they allowed changing grammatical category, adjectival modification, transformation passive, nominalization,… For this reason, we approach the problem from combinations of canonical forms, rather than graphic words: "*el trasplante de órganos*" (organ transplant,), and "*trasplantó el órgano*" (He transplanted the organ) are considered instances of the same collocation(Koike, 2001). However, the characteristic that distinguishes them from other combinations is the preference, as the speakers of the language could choose another combination to convey the meaning intended, but have mostly chosen to use collocations.

The main problem addressed is the automatic extraction of collocations by processing corpus evaluating association measures or collocational indicators that capture the relationship established between the base and the collocative through the use made of both elements in the corpus, individually and together.

Specifically, we consider the corpus as a sample of the use of the language use, any combination is expected to appear in it by chance, i.e. the general case that is considered in the production of combinations is free combinations.

The statistical concept of independence is an ideal tool to determine when two phenomena have occurred together by chance, that is independently. If instead of this, there has been motivation, i.e. the fact that some of them happen in some way influences the possibilities to originate the other. In terms of probabilities, this stated in the statistical law:

$P(x, y) = P(x) * P(y)$

In this paper, "*x*" means the word x appears in the corpus, same for "*y*" and "(*x, y*)" represents the co-occurrence of the word x and word y.

Actually, a textual corpus is a sample of the use of language, so instead of working with probabilities, calculations are done on estimates it through the observed frequencies. In this case we will refer the number of occurrences of the combination and the words x and y individually.

Nomenclature

f(x,y)	*x, y* co-ocurrence frequency
f(x)	word x frequency
f(y)	word y frequency
$\bar{f}$	arithmetic mean of frequencies
s	standard deviation
D-test	Dunning's test

1.1. Association Measures

Based on statistical independence have arisen various proposals to measure the association between two words that appear together in the corpus and use it to make ranking scores that allow to order combinations as collocations candidates. The most simple association measure is the relative frequency, also so called frequency of appearance of *x* with *y* (Koike, 2001). If this value is high, it means that if *x* appears it's probably that also appears *y*, it doesn't meant that we obtain the same value that appearance *y* with *x* frequency.

$$relfreq_x = \frac{freq(x,y)}{freq(x)}$$
$$relfreq_y = \frac{freq(x,y)}{freq(y)}$$

According to the approach taken to solve the problem we find measures based on information theory: it measures how much information provides the occurrence of one item of the combination on the occurrence of the other (Church, Hanks, 1999). In this group are the mutual information and its variants.

$$I(x,y) = \log_2 \frac{p(x,y)}{p(x) * p(y)}$$

On the other hand, there is the group of association measures relating to statistical tests, which are solved in terms of the distribution that is presupposed to the combinations. They try to answer the question: are the data extracted from corpus consistent with the hypothesis of combinations independence?. Z-score (Pearce, 2002), t-score (Evert, 2005), and *D*-test (Manning, Schütz, 1999) are in this group. The expression of each one of them that we used can be found in equations bellow

$$Z\text{-score} = \frac{f(x,y) - \bar{f}(x,y)}{\sqrt{\bar{f}(x,y) * \left(1 - p(y)\right)}}$$
$$\bar{f}(x,y) = p(y) * f(x) * |D|$$
$$p(y) = \frac{f(y)}{N - f(x)}$$

where *p(y)* is the probability to occur *y*, in a different position in the corpus, *N* is the amount of words in the corpus and *D* the number of possibilities in which *y* it could appear around *x*, it's with twice the collocational span used in the concordance lines.

$$t\text{-score} = \frac{f(x,y) - \frac{f(x) * f(y)}{N}}{\sqrt{f(x,y)}}$$
$$D\text{-test} = \log L(f(x,y), f(x), p) + \log L(f(y) - f(x,y), N - f(x), p)$$
$$- \log L(f(x,y), f(x), p_1) - \log L(f(y) - f(x,y), N - f(x), p_2)$$
$$\log L(k,n,p) = k \log p + (n-k)\log(1-p)$$
$$p = \frac{f(y)}{N}, p_1 = \frac{f(x,y)}{f(x)}, p_2 = \frac{f(y) - f(x,y)}{N - f(x)}$$

2. Experiments

Several association measures have been used for the automatic extraction of collocations from the corpus of GEDLC, a collection of plain text, without any linguistic information in which 300000000 of words have been accounted. It contains a collection of approximately 11000 texts. Among other genres this corpus includes works of literature, classical and contemporary, universal and Spanish poetry and prose, drama, fiction, essays, speeches and newspaper articles, in short, a large sample of Spanish.

Due to the characteristic of formal flexibility of collocations, actually the corpus has processed doing counts of canonical forms, rather than graphic words, for example, if we find "*zanjaron la polémica*", "*zanjaré la polémica*"

will be accounted as combination of canonical forms: "*zanjar-polémica*" (to resolve-polemic). Since texts are not labeled the GEDLC's lemmatizer was used to obtain all canonical forms which could come from each word processed (Santana, 2007). This tool was used through a web service; it is able to recognize 151103 canonical forms, so it does not limit the coverage of the process. On the other hand, we have taken into account the grammar structures: *noun + verb*, *adjective + noun* and *verb + adverb* considering the order doesn't matter. It is assumed that two canonical forms are part of a combination in the corpus if they are within the same sentence and have a spacing of at most 10 words on right or left. Furthermore, a catalog of empty words (or stop list) is rejected for processing including: articles, prepositions, conjunctions, interjections, determinants numerals, verb "*ser*", "*estar*" (to be), etc.

The extracted data were recorded on a database containing 14475136 combinations whose frequency is greater or equal to 3. That information is complemented by a collection of about 1800 combinations compiled from examples in works about Spanish collocations and other 28000 combinations that appear in REDES (Bosque, 2004), both groups are used to test the ability of association measures to detect collocations.

Empirically it was determined that failure to reach at least 0.0001 in the relative frequency or 1.5 for mutual information should not be taken into consideration. On the other hand, seemed sufficient indication for collocations have at least 10 samples of the combination and relative frequency that exceeds the value of 0.05 or mutual information greater than 6. In the group of doubtful cases free combinations are found.

Also was evaluated Z-score, t-score and D-test, the ranges of variation show that Z-score and t-score have no differences between control sets, this result seems useful to automate the process of cataloguing combinations as collocations according to the interval that fix the test sets.

Table 1. Range of variation of the statistics in the test sets vs. Corpus

	Z-score	t-score	D-test
Collected	[-55,41 695,99]	[-0,66 127,55]	[83,53 1904484]
C. D.	[-72,64 695,99]	[-2,38 197,37]	[1712,65 19664670]
Corpus	[-172,32 5539,22]	[-20,31 268,26]	[83,53 19044484,55]

A direct relation between t-score values and the co-ocurrence has been detected. Even more, in all the cases it's evident the influence of the frequency of use of the elements making up the combination with the punctuation obtained by this one: minimal values of Z-score for verbs as: *hacer, tener, formar, decir*, … which are commonly used in Spanish, maximum values of t-score when both elements of the combination have a high frequency of occurrence, or D-test t where we find the highest positions in the ranking corresponding to the so-called functional collocations.

We evaluated the data using a significantly less extensive corpus, compiling a collection of Galdós's novels, with a total of 2299920 words. In this case, they were analysed 41302 combinations *noun + verb* that they were found. The ranges of variation of all indicators are substantially modified, proving that the threshold values should be altered as you change the size. There demonstrates again the influence of the number of samples that is possible to gather of a combination on the values that mark the limit between what it is or not collocation. It would be necessary to revise the cutoff values, including the minimum number of samples that it is necessary to fix. One way to solve it would be to review the data manually, and to determine it empirically, as was done in the previous section. Another method implies using a training set, in our case the collected collocations or REDES, with the disadvantage of having such resources and that the corpus processed contains enough samples of they.

Table 2. Range of variation of indicators in the Galdos`s corpus.

	Relative Freq.	Mutual Inf.	Z-score	t-score	D-test
Collected	[0,0049 0,7846]	[0,28 13,13]	[-57,07 13,96]	[-46,72 5,93]	[-26,71 46634]
REDES	[0,0028 1]	[0,5 13,26]	[-61,11 10,98]	[-39,79 14,54]	
Corpus	[0,00092 1]	[-0,99 18,19]	[-136 76,48]	[-116,19 14,54]	

It's not very useful to compare association measures on the entire corpus, since its behaviour is determined by the number of samples of the combination and co-ocurrence frequency of their components, i.e. the use which is given to involved words. It is not possible to compare the value of an indicator based on the frequency between verbs as *dar* with frequency 1118012, *desempeñar* (18903), or *traspirar* (100).

3. Collocations as outliers

In view of the results obtained it exists a clear need to modify the strategy based on cutoff thresholds to delineate the border between collocations and free combinations. The ultimate goal is to obtain a tool to automate as much as possible the extraction of collocations in a corpus, with independence of its size. This strategy is based on the preference's property developed in the collocation concept, according to which it seems logical that, fixed a collocate, the analysis of phraseological characteristics is made between the samples obtained from it, and not on the whole amount of combinations extracted from the corpus. Thus the comparison is avoided between different words from the point of view of its frequency of use, whose differences increase as the corpus is considered to be more extensive.

Under these hypotheses, it was tried to construct an indicator that detects those combinations which use stands out with regard to the habitual thing, in contrast to the underlying idea in the statisticians analyzed in the previous sections, in which the decision is made in terms of rejecting the independence in the use of the collocates. In this sense it is considered that the relative frequencies will be essential, since they provided information about the use of a particular combination in relation to what has been used a certain word in the corpus on which work. Our hypothesis is based in that atypical values of this one will provide indications of the phenomenon of the preference. Using this concept allows us to compare the data in relative terms the use gave in the corpus to a particular word. For instance, will be detected what nouns are preferred using with a particular verb with independence of how it has been used in the reference corpus.

Following this methodology it was supposed that every word has its own sample and consider what cases are atypical in it, against considering the corpus as a single sample where to reject the independence of the collocates. This also allows us to abstract the problem of establishing cut-off values depend on the size of the corpus or to establish rankings that compare data that directly depend on how common is the use of the words involved in the combination.

We focus from now on experimentation with techniques traditionally used for identifying outliers in samples. In statistical, outlier is the term used to refer to the sample data that appears to be inconsistent with the rest of the set, ie they are values that seem too big or too small compared to other observations (Aggarwal, 2013). In this section are presented the analysis of two simple strategies widely used in other areas for the detection of these exceptional elements.

3.1. Method based on Chebyshev's inequality

Chebyshev inequality ensures that data from a sample verifies:

$$f(\{i: |x_i - \bar{x}| > ks\}) < \frac{1}{k^2}$$

Where $\bar{x}$ is the sample mean and s the standard deviation, considering atypical observations those with $|Z| > 3$, where:

$$Z = \frac{x_i - \bar{x}}{s}$$

Applied to our problem, the $x_i$values are relative frequencies for a fixed word, and $\bar{x}, s$ their arithmetic mean and standard deviation respectively, thus the statistical expression for collocations:

$$ZChebyshev(x_i, word) = \frac{relfreq(x_i, word) - \overline{relfreq}(word)}{s_{relfreq}(word)}$$

The combinations $(x_i, word)$ such that $ZChebyshev(x_i, word) > 3$ will be accepted as collocations.

Tables 12 and 13 summarizes the results of applying this strategy to *Noun + Verb* combinations to complete corpus and Galdo's Corpus, showing that the requisite that was demanded from the combinations to consider them to be collocations was excessively restrictive.

Table 3. Results of ZChebyshev in full corpus.

	Range	ZChebyshev ⩾ 3
Collected	[1.26E-5 11,82]	139
C.D.	[1.769E-5 14,1]	591
Corpus	[1.769E-5 19,75]	18387

Table 4. Results of ZChebyshev in Galdós's corpus.

	Range	ZChebyshev ⩾ 3
Collected	[0 4.4]	3
C.D.	[0 10.6]	18
Corpus	[0 13.19]	807

3.2. Hampel's method

One source of error in methods of identifying the outliers is precisely the use of anomalous data in the calculation of the mean and variance employed to evaluate the statistics (Leys et al., 2013). These out of range values alter the arithmetic mean as estimate of what to expect in terms of standard conditions for the variable. The statistical MEDA based on the median of the data is robust against atypical observations, so more effective methods to identify outliers are based on it, although the calculations are more complex. The second proposal is to consider the MAD method, based on MEDA from relative frequency samples associated with each word (Manoj Senthamarai, 2013).

$$MAD = \frac{|x_i - median(X))|}{MEDA}$$

Being:

$$MEDA = median\{|x_1 - median(X)|, |x_2 - median(X)|, \cdots, |x_n - median(X)|\}$$

$X = \{x_1, x_2, \cdots, x_n\}$ is the variable sample. *MAD* verifies that 50% of the sample data are in the interval:

$$[\bar{x} - MEDA, \quad \bar{x} + MEDA]$$

Applying this property and empirical evidence is admitted that the sample data with $|MAD| \geq 4{,}5$ can be considered outliers. Adapting this method to our problem leads us to consider:

$$MADF(x_i, word) = \frac{|relfreq(x_i, word) - median(relfreq(word))|}{MEDAF(word)}$$

$$MEDAF(Word) = median\{|relfreq(x_i, word) - median(word)|: (x_i, word) \in \text{Corpus}\}$$

Where:

$$median(relfreq(word)) = median\{relfreq(x_i, word): (x_i, word) \in \}$$

In this case the coverage rises notably, and we can perform calculations where the *MEDAF* is a non-zero.

Table 5. Maximum values for MADF: Corpus

			MADF				*MADF*
abrir-puerta	puerta	abrir	4847.37	abrir-ojo	ojo	abrir	2538.62
cerrar-puerta	puerta	cerrar	3393.14	Estados Unidos	estado	unir	2525.42
encoger de hombros	hombro	encoger	3224.00	pronunciar-palabra	palabra	pronunciar	2512.25
muchas veces	mucho	vezar	2923.77	dar-vuelta	vuelto	dar	2359.37
decir-señora	señora	decir	2918.89	dar-vuelta	vuelta	dar	2336.43
decir-señor	señor	decir	2918.89	punto de vista	visto	puntar	2269.12
saber-cómo	cómo	saber	2872.83	punto de vista	vista	puntar	2233.25
cerrar-ojo	ojo	cerrar	2640.28	partir-mayor	mayores	partir	2213.50
saber-bien	bien	saber	2571.00	partir-mayor	mayor	partir	2213.50
saber-bien	bienes	saber	2571.00	partir-mayor	mayora	partir	2213.50
saber-bien	bien	saber	2552.83	partir-mayor			

Table 6. Maximum values for MADF: Galdós's Corpus

		MADF			*MADF*
ojo	cerrar	172,99	palabra	pronunciar	110,00
puerta	abrir	160,33	bien	saber	107,66
decir	querer	135,66	bien	saber	107,66
cómo	saber	134,00	bienes	saber	107,66
puerta	cerrar	133,99	ojo	clavar	99,00
señor	decir	122,39	ojo	abrir	98,33
señora	decir	122,39	mano	coger	94,99
cosa	decir	121,40	cómo	ser	94,49
cartas	escribir	113,99	manos	coger	93,99
carta	escribir	113,99	rato	largar	93,99

In Table 7 are shown ranges achieved in each group according to the grammatical structure of the combination. Similar ranges were seen in groups *Noun + Adjective* and *Verb + Adverb*, which emphasize the much greater amplitude in the *Verb + Noun* case. Regarding coverage test sets, in all structures rises drastically compared to the results obtained with other criteria (Fig.).

Table 7.Ranges of *MADF* by grammatical structure

	MADF
Verb + Noun	[0 4309.99]
Noun + Adjective	[0 2531.47]
Verb + Adverb	[0 2332.56]

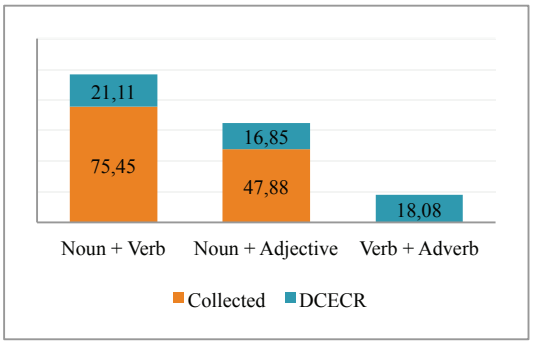

Fig. 1. Recall of MADF in test sets

Under our assumptions, the valuation requires to fix one of those collocates that has been used for the evaluation of MADF. Table 8 presents results for *dar* in both corpora, obtaining similar results. Also in Table 9 are Association measures for some collected collocations, all of them are duplicated because relative frequencies, *ZChebyshev* and *MADF* was calculated over word *x* and word *y* from the co-ocurrence. Collocations as "*tener-ardor"* or "*condenar-puerta"* have a low level of Mutual Information or Z-score but they are outliers, when they are fixed: *ardor* and *condenar*, respectively. Thus, we can say, that verb *tener* is preferred for *ardor*, or the noun *puerta* is preferred for verb *condenar*. In another hand, "*campaña-electoral*" or "*repoblación-forestal*" have high scores with mutual information and Z-score also they are atypical values.

Table 8. Maximum values when is fixed verb "*dar*"

		MADF			*MADF*
vuelto	dar	2359.37	vuelta	dar	85.50
vuelta	dar	2336.43	vuelto	dar	85.50
pasa	dar	1331.12	paso	dar	69.25
paso	dar	1331.12	pasa	dar	69.25
dios	dar	1109.81	gana	dar	47.25
diosa	dar	1109.81	dios	dar	44.00
hombre	dar	1026.31	diosa	dar	44.00
vido	dar	1003.49	mano	dar	41.25
vida	dar	997.68	manos	dar	41.00
gracia	dar	993.49	cuenta	dar	38.50
	CORPUS			GALDOS'S	

Table 9. Association measures for some collected collocations

x	*y*	*relfreqx*	*relfreqy*	*Mutual Inf.*	*Zscore*	*ZChebyshev*	*MADF*
optimismo	tener	0.044	5.01E-05	2.89	-9.79	0.16	1.23
optimismo	tener	0.044	5.01E-05	2.89	-9.79	6.74	85.99
ardor	tener	0.0343	0.0001	2.51	-23.75	0.05	7.028
ardor	tener	0.0343	0.0001	2.51	-23.75	7.09	103.33
puerta	condenar	0.0006	0.0035	2.38	-16.80	0.00	25.39
puerta	condenar	0.0006	0.0035	2.38	-16.80	0.843	7.79
absoluto	silencio	0.0294	0.0155	6.65	66.52	5.354	134.90
absoluto	silencio	0.0294	0.0155	6.65	66.52	8.41	150.00

desazón	producir	0.0351	0.0005	6.66	13.68	0.08	3.90
desazón	producir	0.0351	0.0005	6.66	13.68	3.48	53.00
futuro	negro	0.0021	0.0010	2.91	-9.32	0.03	6.45
futuro	negro	0.0021	0.0010	2.91	-9.32	0.22	9.24
amor	profesar	0.0034	0.0838	6.84	54.06	1.08	36.25
amor	profesar	0.0034	0.0838	6.84	54.06	12.73	247.33
campaña	electoral	0.0223	0.0476	9.65	123.23	4.80	64.49
campaña	electoral	0.0223	0.0476	9.65	123.23	5.23	77.59
asado	suculento	0.0027	0.0096	9.65	19.59	0.12	5.99
asado	suculento	0.0027	0.0096	9.65	19.59	0.23	0.74
forestal	repoblación	0.0465	0.1382	15.45	310.69	3.30	38.99
forestal	repoblación	0.0465	0.1382	15.45	310.69	5.21	39.00

4. Conclusions

This paper presents an analysis of the exploitation of a large corpus of Spanish based on the words frequencies and co-occurrence frequencies. It has reviewed the performance of different association measures used automatic extraction of collocations in texts. They are compared two statistical techniques used to solve the problem of outlier detection. The proposal aims to identify, when a word is fixed, with which other can be established that the use is out of the ordinary in their field. Have been evaluated two possibilities that we call: ZChebyshev and *MADF*. The first case, based on the mean and standard deviation, it was discarded by their limited coverage. *MADF* statistic, however, based on the median is revealed as a reliable indicator for the automatic extraction of collocations, since it allows to fully automating the process without having to use a manual review of a subsample or training sets. Using *MADF* the vast amount of free combinations that were incorporated in our catalogs are filtered and it's possible to detect which word in the combination has the preference for another one.

References

Aggarwal, C. (2013). *Outlier Analysis*. New York: Springer.

Bosque, I. (2001) Sobre el concepto de colocación y sus límites. *Lingüística Española Actual XXIII/1*, 9 - 40.

Bosque, I. (2004). *REDES Diccionario combinatorio del español contemporáneo*. Madrid: Ediciones SM.

Church, K. W., and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics, 16 (1)*, 22 - 29.

Evert, S. (2005). *The Statistics of Word Co-occurrences. Word Pairs and Collocations*. PhD Dissertation, Stuttgart University.

Koike, K. (2001). *Colocaciones léxicas en el español actual*. Madrid: Universidad de Alcalá.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting Outliers: Do Not Use Standard Deviation around the Mean, Use Deviation around the Median. *Journal of Experimental Social Pyschology, 49*, 764 – 766.

Manning, C.,and Schütze, D. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manoj K., and Senthamarai Kannan, K. (2013). Comparison of Methods for Detecting Outliers. *International Journal of Scientific and Engineering Research, 4 (9)*, 709 - 714.

Pearce-Lazard, D. (2002). A Comparative Evaluation of Collocation Extraction Techniques. *Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Santana, O., Pérez, J., et al. (Eds.) (2007). Development of Support Services for Linguistic Research over the Internet TIN2004-03988. *National Program in Computer Technology (TIN)*, 167 - 174.