A Lexicon of Verb and *–mente* Adverb Collocations in Portuguese Extraction from Corpora and Classification *

Lucas Nunes Vieira[1,2,4], Cláudio Diniz[2], Nuno Mamede[2,3], Jorge Baptista[2,4]

[1] Univ. Franche-Comté, Besançon, France
[2] L2F-Sopken Language Lab, INESC-ID Lisboa, Lisboa, Portugal
[3] Instituto Superior Técnico, Univ. Técnica de Lisboa, Lisboa, Portugal
[4] CECL/Univ. Algarve, Faro, Portugal
{lucasnvieira,cfpdiniz}@gmail.com, nuno.mamede@inesc-id.pt, jbaptis@ualg.pt

1. Introduction

Collocations started to be a target of research in the twentieth century after FIRTH (1957) coined the term and called attention to the fact that the way we combine words in natural language is far from being unconstrained. In the sense of Firth, a pair or group of words can be considered a collocation if the probability for their co-occurrence exceeds chance levels. For a long time this concept has prevailed in the literature as the rationale behind the task of collocation extraction. However, the more recent formulation of MEL'ČUK (2003) provides a more semantic-based view on the phenomenon that does not necessarily coincide with an attested high frequency of the word combinations. According to Mel'čuk, the meaning of certain words would dictate the adjacent use of others, forming groups or pairs of *base* words and *collocates*.

Concerning the linguistic pattern investigated in this study, namely pairs of verb and –*mente* ('-ly') ending adverbs, the verb would be the base of the combination, while the adverb would be the collocate. The strategy here adopted for the extraction of this pattern profits both from Firth's and Mel'čuk's formulations, since, at different stages, it relies both on frequency of distribution and on meaning-oriented human annotations.

The corpus used for the extraction of verb-adverb bigrams was the CETEMPúblico[1] (SANTOS & ROCHA, 2001) corpus of European Portuguese, consisting of 192M words of journalistic texts. This is, to the best of our knowledge, the largest freely distributed corpus of Portuguese. Albeit constituting just over 10% of all simple adverb occurrences in the corpus, adverbs ending in *–mente*, henceforth *Adv-mente*, represent in fact the majority of the simple-word lemmas of this grammatical class, based on data from the CETEMPúblico.

While a number of initiatives at collocation extraction have relied substantially on a search for adjacent words, as CHOUEKA (1988), newer studies suggest that methods involving some level of syntactical parsing might prove more precise for certain linguistic patterns (SERETAN, 2011). In view of this, we experiment with a syntax-based approach for the extraction of verb-adverb pairs from the corpus. This pattern could be deemed a challenging one in respect to the extraction task, since adverbs can occupy different positions in the sentence, being commonly associated with a rather loose mobility in the speech (BECHARA, 2003).

In the remainder of this paper, we describe the syntax-based approach adopted to extract collocations from the corpus (Section 2), explain the linguistically-motivated classification of collocation candidates (Section 3), and present an empiric evaluation of statistical association measures in identifying {*V, Adv-mente*} collocations (Section 4). We conclude by discussing the appropriateness of the methods experimented with in view of {*V, Adv-mente*} pairs and by proposing future work (Section 5).

* Research for this paper has been partially funded by Fundação para a Ciência e a Tecnologia/MCTES, Portugal, through the REAP.PT project under program Portugal-CMU (proj.ref. HuMach/0053/2008) and by the European Union, Education and Training, under the Erasmus Mundus Master Course Program (proj ref. EMMC/IMNLPHLT/2008-0083).

[1] http://www.linguateca.pt/CETEMPublico/ [Accessed 5 May 2012]

2. Collocation extraction

2.1 Syntactically parsing the corpus

In order to illustrate some of the pitfalls that the linguistic pattern investigated may pose to the task of collocation extraction, a potentially problematic context is presented below:

(1) *O aluno leu o livro atentamente e resumiu-o*
«The student read the book attentively and summarised it»

Even though the *Adv-mente* in this example co-occur with two verbs, *ler* «to read» and *resumir* «to summarise», it only holds a syntactical dependency with the verb *ler* «to read», forming the pair *ler atentamente* «to read attentively», which can be considered to have collocational status in Portuguese. However, there is a high probability that the pair *resumir atentamente* «to summarise attentively» would erroneously come up as a collocation candidate in a search for adjacent words, since the adverb co-occurs with the two verbs in considerable proximity in the same sentence.

In order to overcome similar problems, the strategy described in this paper included the syntactical parsing of the corpus envisaging a more precise extraction of verb-adverb pairs that actually hold a syntactical dependency between them. The STRING computer-based text processing chain (MAMEDE ET AL., 2012) has been used for that purpose. In broad terms, the chain comprises three main stages: pre-processing, disambiguation, and syntactic analysis, respectively. The pre-processing stage is responsible for text segmentation, part-of-speech (POS) tagging, and for the splitting of the input into sentences. In the last stage, the syntactic parsing of the text is performed by XIP (Xerox Incremental Parser) (AÏT-MOKHTAR et al., 2002), a rule-based finite-state parser that establishes syntactic dependencies between words. In this framework, the output provided by the system includes a dependency relation (called MOD[fier]), with the correct verb-adverb pair. An example of the output yielded for sentence (1) is provided below:

```
VDOMAIN(leu,leu)
VDOMAIN(resumiu-o,resumiu-o)
MOD_POST(leu,atentamente)
SUBJ_PRE(leu,aluno)
SUBJ_PRE_ANAPH0(resumiu-o,aluno)
CDIR_POST(resumiu-o,o)
CDIR_POST(leu,livro)
TOP{NP{O aluno} VF{leu} NP{o livro} ADVP{atentamente} e
VF{resumiu-o} NP{o}}
```

After processing the entire corpus, MOD dependency relations of this kind between verbs and *Adv-mente* were extracted from the output. The number of verb-adverb bigrams obtained was of 65,535, whose frequency in the corpus exceeds 290,000 occurrences altogether.

2.2 Filtering the extraction output

In order to narrow down the universe of potential cases of collocations, a number of filtering strategies were applied to the total set of bigrams extracted from the corpus so as to eliminate from the outset cases that did not present any potential of forming collocations. A frequency threshold of five ($f \geq 5$) was established for the consideration of pairs as collocation candidates. This is a threshold that can be deemed considerably low given the fact the total number of words in the corpus is 192M.

With regard to *Adv-mente*, we have augmented the adverbial classification carried out by FERNANDES (2011) for *Adv-mente* in Portuguese, which initially covered approximately 520 adverbs. This number has now been increased to nearly 1,000.

Having knowledge of the class or classes a given *Adv-mente* belongs to played an important role in filtering out adverb categories that do not hold a straight connection with the verb, which consequently impedes the formation of verb-adverb collocations. That would be the case of adverbs that play the single role of modifying a sentence, i.e. sentence-modifying *Adv-mente*, namely conjunctive adverbs (PC), disjunctive adverbs of style (PS), and disjunctive adverbs of attitude (PA) (MOLINIER & LEVRIER, 2000). Focus adverbs (MF), albeit being commonly integrated in the clause, were also filtered out due to their low potential of receiving collocation status, since their sole purpose in an utterance is to emphasise a sentence constituent.

Certain verbs with little semantic content were also filtered out at this stage. So-called *support verbs* (GROSS, 1981), such as *fazer* «to do», *dar* «to give», *ter* «to have», and *haver* «to exist»/«there is/are» were amongst the verbs to be discarded, as well as copula verbs such as *ser* «to be», *estar* «to be», *permanecer* «to remain», *ficar* «to stay», and *parecer* «to seem». However, if these verbs were part of a verb chain, i.e. a verb phrase forming a compound tense with a past participle form, the modifying relation would be established between the participle, as head of the phrase, and the adverb. Participles used as adjectives were also ignored.

Due to the high frequency with which the verbs just mentioned are used in the language, they were present in the vast majority of verb-adverb combinations extracted from the corpus. After the filtering process, the remaining number of verb-adverb different bigrams was 5,973, which was then considered the set of collocation candidates on which manual annotations would be made.

3. Manual classification of collocation candidates

3.1 Establishing empiric criteria

The criteria empirically devised to establish the collocational status of candidate pairs take into account the relationship of a given term with its possible collocates as well as the meaning of the words involved in the combination. An explanation of these criteria, along with illustrating examples, is presented bellow.

1. The adverb has a hyperbolic meaning in the combination, e.g.:

(2) *Ele **esperou eternamente** pelo telefonema*
«He waited eternally for the phone call»

2. The adverb holds a non-literal meaning in the combination, e.g.:

(3) *O time **venceu confortavelmente** a partida* «The team **won** the match **comfortably**»
≠ *O time estava confortável* «The team was comfortable»
(4) *Ele **deitou-se confortavelmente** na cama* «He **lay comfortably** in bed»
= *Ele estava confortável* «He was comfortable»

While in (4) the adverb *confortavelmente* «comfortably» holds its literal meaning, connected to the idea of physical comfort, in (3) it assumes a figurative meaning adopted to express the idea that the match was won *effortlessly* or by a large scoring difference. The non-literal meaning in this case attributes a unique character to this combination that could be potentially indicative of its collocational value in Portuguese. In (4), the adverb in the combination, a manner adverb with scope on the action itself and on the subject of the verb, can be paraphrased by its equivalent base adjective operating on the same subject. In (3) this transformation is not possible, which reinforces the non-literal meaning of the adverb in the context of this sentence.

In another example, the adverb modifies the verb by according a quantifying/intensive value to it, such as *perdidamente* «lost-ly», below:

(5) *Ele **apaixonou-se perdidamente** por ela* 'He **fell lost(ly)**[2] in love for her'
≠ *Ele estava perdido* 'He was lost'

[2] This is a case where equivalent collocations in English and Portuguese have adverbs that differ altogether, both morphologically and etymologically. *Perdidamente* «lost-ly» is an adverb for which there is no literal translation in English. There is also no equivalent verb for *apaixonar-se*, so that instead the support verb *to be in love*, in the inchoative variant *to fall in love,* was considered here (the inchoative variant was used for consistency with the aspectual value of the Portuguese verb). For this English expression, an adverbial collocation could be *madly*. Interestingly, the base adjective *perdido* also appears in a compound adjectival construction *estar perdido de amores por* «to be lost of/from love-pl for». A similar construction also exists with the adjectives *doido* and *louco* «mad»: *estar louco/doido de amores por* «to be mad of/from love_pl for».

In (5), the adverb *perdidamente* «lost-ly» is derived from the adjective *perdido* «lost», but its intensifying meaning in the combination can not be (regularly) derived from the meaning of the base adjective («one who does not know or is unable to find his/her whereabouts»).

3. The combination belongs to the specific vocabulary of a scientific or technical area of expertise, e.g.:

(6) *Ele **respondeu civilmente** pelo crime que cometeu*
«He **responded civically** for the crime he committed»

In (6), the *{V, Adv-mente}* pair is part of the vocabulary commonly used in the domain of law, which accounts for the fixedness of the expression in Portuguese.

4. Synonymic relations between adverbs are broken in the collocational context, e.g.:

(7) *Ela chorava **copiosamente*** «She cried **copiously**»
(8) *?Ela chorava **abundantemente*** «She cried **abundantly**»

Even though the adverbs *copiosamente* «copiously» and *abundantemente* «abundantly», in (7) and (8) respectively, could be considered synonymous, only the adverb *copiosamente* holds a collocational value in this context, since the use of *abundantemente* renders the construction unnatural in Portuguese. We thus say that the synonymic relation between these adverbs is broken. In (8), the adverb would be substituted by *hysterically* or *uncontrollably* in equivalent collocations in English.

5. In a collocation context, the adverb holding collocational status cannot be combined with the antonymous of the verb in question, e.g.:

(9) *O time **venceu** a partida confortavelmente* «The team **won** the match comfortably»
(10) **O time **perdeu** a partida confortavelmente* «The team comfortably **lost** the match»

While the *{V, Adv-mente}* combination in (9) can be considered a collocation in Portuguese, the antonymous of the verb seems to impede a coherent construction in (10), which can be used as an index of the collocational value of the pair in (9). Naturally, this criterion only holds true if an antonymous form of the verb exists in the language. Equally noteworthy is the fact that the simple use of negation does not function as a deciding parameter, as both the collocation status and coherence of the combination would be maintained in this case:

(11) *O time **não** venceu a partida confortavelmente*
«The team **did not** win the match comfortably»

6. The adverb can be combined with often only one subset of the possible meanings of the verb, e.g.:

(12) *A secretária **reproduziu fielmente** os documentos*
«The secretary **reproduced** the documents **faithfully**»
(13) **Coelhos **reproduzem-se fielmente*** «Rabbits **reproduce faithfully**»

While the adverb *fielmente* «faithfully» can be combined with the verb *reproduzir* «to reproduce» in (12), the combination is not possible in (13), as the verb in this sentence, albeit being homonymous to the one in (12), consists in fact of a different lexical item with a different syntactic construction.

Based on these criteria, the set of candidate bigrams were classified as to their collocation status, receiving either the tag of collocation or the tag of non-collocation. Table 3.1 shows the number of candidate bigrams per frequency range in the corpus, and the number of cases among them that have been classified as collocations.

Freq. Range	# Candidates	# Collocations
> 100	65	39
100 - 10	2700	334
5 - 10	3208	128

Table 3.1 Number of candidates and of collocations per frequency

3.2 Assessing native speakers' intuitions

In order to test the intuition of native speakers of Portuguese with regard to the collocational status of the linguistic pattern investigated, a sample classification task was carried out with 21 subjects, of which 13 were native speakers of European Portuguese and 8 of Brazilian Portuguese. The dataset to be classified was composed of 30 collocation candidates randomly selected, 15 having been previously classified as collocations, and 15 as non-collocations. The candidate pairs were presented to the subjects in the contexts where they actually occurred in the corpus, with the {*V, Adv-mente*} pairs highlighted in each sentence. Prior to making a decision on the status of the pairs, the subjects were asked to attentively consider the set of guiding criteria presented in Section 3.1.

Cohen's κ statistic chance-corrected inter-annotator agreement (COHEN, 1960) was calculated for classifications made on the entire set of 30 pairs randomly selected for the experiment. Results are presented in Table 3.2. Results based solely on the 15 pairs that had been previously classified as collocations are presented in Table 3.3.

κ for 30 randomly selected candidates	
Percent of overall agreement	0.57
Fixed-marginal kappa	0.06

Table 3.2 κ for 30 randomly selected pairs of collocation candidates

κ for 15 cases of collocation in the sample	
Percent of overall agreement	0.62
Fixed-marginal kappa	0.10

Table 3.3 κ for 15 pairs among random selection previously classified as collocations

Cohen's κ values can vary from -1.0 to 1.0, where 0 would represent chance agreement. The κ results for the entire set of randomly selected collocation candidates and just for the cases previously classified as collocations were of 0.06 and 0.10 respectively, which can be considered to stand in the range of slight agreement according to the interpretation scale proposed by LANDIS & KOCH (1977).

Even though these results are above what could be considered agreement by chance, they can arguably be deemed low. The most likely reason for this lies in the fact that the sample used in the experiment was too small, requiring an extremely high raw agreement percentage in order for the κ value to reach higher levels of significance. Because of this, κ values achieved in the experiment do not allow for definitive conclusions to be taken with respect to the agreement of the recruited subjects on the collocational status of word pairs. The limited size of the sample was due to the foreseen resistance that a larger set would most likely find among potential voluntary annotators, and to the risk of losing consistency if a larger list of examples had been presented to them.

The poor κ values could also be considered indicative of the elusive nature of the concept of collocation. In this respect, it can be seen in Tables 3.2 and 3.3 that the agreement achieved among cases that had been previously classified as collocations was higher than the overall agreement. This denotes that identifying negative cases poses more difficulty than identifying positive ones, which only confirms that the limit between both is far from being clear-cut. Considering just the positive cases, it can be noticed that a raw agreement of 62% has been reached, which, despite the low κ value, could be considered indicative in some degree of the collocational phenomenon addressed in this paper. One of the factors that denote this phenomenon is the premise that certain lexical combinations, in certain language producing contexts, tend to be given preference over other combinations that can be deemed nearly semantically equivalent. That would be the case of example (7) above, where to express the idea of crying in excess, the Portuguese speaker seems to give preference to the adverb *copiosamente* «copiously» or even *convulsivamente* «convulsively», as opposed to other adverbs of very similar semantic content, such as *abundantemente* «abundantly» and *excessivamente* «excessively». This is one of the reasons to consider the former pairs as collocations, which does not seem to be the case in regard to the latter ones.

4. Correlation of association measures with human classification

A number of statistical association measures have already been tested for capturing the linguistic phenomenon of collocations. PECINA (2010) provides an extensive account in this respect, remarking the particularly good performance of Unigram Subtuples (UnigSub) (PECINA, 2010) and Mutual Information (MI) (FANO, 1961) for large-sized corpora. SERETAN (2011), in turn, mentions the appropriateness of Log-likelihood Ratio (LLR) (DUNNING, 1993) for capturing low-frequency word combinations. In this Section, the manual classification of the collocation candidates will be contrasted with association measures that have received significant attention in previous research. The aim of this comparison is to unveil the measures that are most sensitive to the collocational pattern investigated.

The following measures were chosen for the experiment: *t*-test, Pearson's chi-square (χ^2), Mutual Information (MI), Log-likelihood Ratio (LLR), Dice Coefficient (Dice), and Unigram Subtuples (UnigSub).

The set composed of 5,973 collocation candidates, already classified as to their collocational status, was stratified into three subsets according to the frequency of the bigrams in the corpus. The subsets correspond exactly to the three frequency ranges presented in Table 3.1. The groups were denominated S1, S2, S3, in decreasing order of frequency, respectively. The *t*-test and χ^2 are both measures that have pre-established statistical significance thresholds for the analysis of results. The performance of these two measures was analysed in terms of precision, recall, and *F*-measure, taking into account a threshold value of 2.576 for the *t*-test, and 3.841 for χ^2, values that correspond to a confidence level of $\alpha = 0.005$ and $\alpha = 0.05$, respectively, which have been previously adopted in similar contexts aimed at identifying collocations (MANNING & SCHÜTZE, 1999 : 153 ; 159). Results of these two measures for S1, S2, and S3 separately, as well as for the set altogether, are shown in Table 4.1

	***t*-test**			**χ^2**		
	Precision	**Recall**	***F*-measure**	**Precision**	**Recall**	***F*-measure**
S1	0.603	0.974	0.745	0.609	1	0.757
S2	0.129	0.937	0.227	0.123	0.964	0.218
S3	0.082	0.460	0.140	0.041	1	0.079
All	0.128	0.818	0.222	0.084	0.976	0.156

Table 4.1. *t*-test and χ^2 results on collocation candidates

Figures in Table 4.1 clearly denote that the *t*-test and χ^2 significance threshold values fell far short of identifying the collocation pattern investigated. The reason behind the poor performance of these measures is most likely connected to the low frequency of the linguistic phenomenon dealt with, a fact that has already been reported in the literature with regard to the *t*-test (DUNNING, 1993). It can also be observed in Table 4.1 that the higher the frequency of the collocation candidates in the corpus, the more satisfactory the performance of the *t*-test and χ^2 is in identifying the phenomenon. The *F*-measure of both tests increases from S3 to S1. Even though a decision can always be made with regard to a threshold value to be applied to the results of statistical association tests, the other measures calculated for the collocation candidates extracted from the corpus – namely MI, LLR, Dice, and UnigSub – do not have a pre-established threshold for filtering results. The correlation of these measures with the manual classification of collocation candidates was assessed based on the Pearson product-moment correlation coefficient (*r*) (PEARSON, 1896), which measures the linear relationship between two variables – in this case, the referred measures and the classification of bigrams as (non-)collocations. Pearson's *r* values for the aforementioned measures, considering S1, S2, and S3 and the set altogether, are presented in Table 4.2.

	Pearson Correlation Coefficient (*r*)						
	t-test	χ^2	MI	LLR	Dice	UnigSub	# Instances
S1	0.0321	0.2358	*0.4562*	*0.3610*	*0.3831*	*0.3469*	65
S2	0.0759	0.0633	0.2876	0.2403	0.1711	*0.3816*	2700
S3	0.1126	0.0447	*0.3137*	*0.3312*	0.1144	0.1707	3208
All	0.1519	0.0528	*0.3093*	*0.3109*	0.2287	*0.3453*	5,973

Table 4.2. Pearson results for *t*-test, χ^2, MI, LLR, Dice, and UnigSub
for considering the classification of collocation candidates

Values for *r* can range from -1.0 to 1.0. According to COHEN (1988), an *r* of .10 could be considered to have a small *effect size* (ES), while an *r* of ± .30 would have a medium ES, and an *r* equal to or above .50 ($r \geq .50$), a large ES. In other words, the furthest the *r* value is from zero, the stronger the relationship between the two variables analysed should be.

In Table 4.2, it can be observed that the four association measures presented a medium ES for S1, the subset with frequent collocation candidates in the corpus. Concerning S3, the *r* value of Dice and UnigSub presented a considerably small ES, which stood at approximately 0.1 for both measures. The small ES of *r* for Dice and UnigSub seems to suggest that these two measures are not appropriate to capture the collocation pattern investigated when it occurs infrequently. LLR, on the other hand, has maintained *r* values from 0.24 to 0.36 across the three subsets. This corroborates findings of previous research that affirm this measure could be deemed reliable for the task of collocation extraction in general (SERETAN, 2011), since it would be sensitive to both high and low-frequency phenomena (DUNNING, 1993 : 62). MI showed a similar trend in this respect, with *r* values ranging from 0.28 to 0.45, where the lowest value corresponds to S2, the subset including pairs of medium frequency in the corpus. Considering the entire set of collocation candidates, UnigSub, LLR, and MI were, in descending order, the measures that presented the highest correlation with the human classification on the collocation status of the pairs. The *t* and χ^2 tests presented a notably low correlation with the classification, which seems to confirm the poor Precision, Recall and *F*-measure results of these two measures, as shown in Table 4.1.

5. Conclusion and future work

In this paper, we described a syntax-based approach to collocation extraction from corpora, and assessed the performance of traditional association measures in identifying the collocation pattern {*V, Adv-mente*} in Portuguese, based on a reference manual classification carried out in accordance with a set of empirically devised linguistic criteria.

Having information on collocational patterns is of high importance to fields such as NLP and Second Language Learning, since conforming to these patterns is of interest both to humans and to computer systems aimed at processing natural language. Pairs such as *apaixonar-se perdidamente* «fall in love madly» for example, due to their morphological/etymological difference between English and Portuguese, might pose a challenge to Machine Translation (MT) engines for instance, which at a number of times fail to provide correct translations for the linguistic pattern here addressed, as it has been shown in VIEIRA (2012).

The identification of collocations conforming to this pattern has proven to be challenging both for statistical association measures and for native speakers of the language. The relatively low frequency of the phenomenon seems to contribute to that difficulty, especially in regard to the statistical classification. Pre-established threshold values proposed for the *t* and χ^2 tests have shown to be particularly unsatisfactory in that respect. The correlation of other measures with the reference classification was considerably more promising.

As future work, we intend to use results of these measures as training data to build an automatic collocation classifier for the {*V, Adv-mente*} pattern in Portuguese using machine learning techniques. This is a strategy that would arguably profit more significantly from the results of these measures, as it disregards any decision based on critical values, but rather takes all results into account as being potentially useful for the classification task. As further assessments to be carried with native speakers, we intend to repeat the experiment previously described making use of a different questioning strategy, asking respondents to rate candidate pairs of similar meaning as opposed to deciding on the collocation status of a single pair.

References

AÏT-MOKHTAR Salah ; CHANOD Jean-Pierre ; ROUX Claude (2002), Robustness Beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*, 8. New York: Cambridge University Press, pp. 121–144.

BECHARA Evanildo (2003) *Moderna gramática portuguesa*. 37 ed. Rio de Janeiro: Lucerna.

CHOUEKA Yaacov (1988) Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in a Large Corpus. In: *Proceedings of RIAO '88*, pp. 609–623.

COHEN Jacob (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, pp. 37-46.

COHEN Jacob (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

DUNNING Ted (1993), Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1), pp. 61-74.

FANO Robert (1961) Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge, MA.

FERNANDES Gaia (2011) *Classification and Word Sense Disambiguation: The case of -mente Ending Adverbs in Brazilian Portuguese.* M.A. Thesis, Universidade do Algarve/ Universitat Autònoma de Barcelona.

FIRTH John Rupert (1957) A Synopsis of Linguistic Theory 1930-55. In: Firth, J. R. et al. *Studies in Linguistic Analysis*. Special volume of the Philological Society. Oxford: Blackwell.

GROSS Maurice (1981) Les bases empiriques de la notion de prédicat sémantique. *Langages* 63, pp. 7–52.

LANDIS J Richard ; KOCH Gary G (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, pp. 159-174.

MAMEDE Nuno ; BAPTISTA Jorge ; DINIZ Cláudio ; CABARRÃO Vera. (2012). STRING: A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. Demo PROPOR 2012. Available at <http://www.propor2012.org/demos/DemoSTRING.pdf> [Accessed 14 May 2012].

MANNING Christopher ; SCHÜTZE Hinrich (1999) *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

MEL'ČUK Igor (2003) Collocations, definition, rôle et utilité. In: GROSSMANN Francis ; TUTIN Agnès (eds.), *Les collocations, analyse et traitement,* Amsterdam: De Werelt, pp. 23-31.

MOLINIER Christian ; LEVRIER Françoise (2000) *Grammaire des adverbes. Description des formes en -ment.* Genève: Droz.

PEARSON Karl (1896) Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London* 187, pp. 253-318.

PECINA Pavel (2010) Lexical Association Measures and Collocation Extraction. *Language Resources and Evaluation* 44(1), pp. 137-158.

SANTOS Diana ; ROCHA Paulo (2001) Evaluating CETEMPúblico, a free resource for Portuguese. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, ACL'2001*, Toulouse, 9-11 July 2001, pp.442-449.

SERETAN Violeta (2011) *Syntax-Based Collocation Extraction. Text, Speech and Language Technology*. Dordrecht: Springer.

VIEIRA Lucas Nunes (2012) PT-EN Collocation Equivalents and Machine Translation Evaluation. *BULAG Natural Language Processing and Human Language Technology*, forthcoming.