Identifying Portuguese Multiword Expressions using Different Classification Algorithms - A Comparative Analysis

Alexsandro Fonseca
University of Quebec in Montreal

201 President Kennedy,
Montreal, QC, Canada

affonseca@gmail.com

Fatiha Sadat
University of Quebec in Montreal

201 President Kennedy,
Montreal, QC, Canada

sadat.fatiha@uqam.ca

Alexandre Blondin Massé
University of Quebec in Chicoutimi

555, boul. de l'Univ.
Chicoutimi, QC, G7H 2B1

alexandre.blondin.masse@gmail.com

Abstract

This paper presents a comparative analysis based on different classification algorithms and tools for the identification of Portuguese multiword expressions. Our focus is on two-word expressions formed by nouns, adjectives and verbs. The candidates are selected on the basis of the frequency of the bigrams; then on the basis of the grammatical class of each bigram's constituent words. This analysis compares the performance of three different multi-layer perceptron training functions in the task of extracting different patterns of multiword expressions, using and comparing nine different classification algorithms, including decision trees, multilayer perceptron and SVM. Moreover, this analysis compares two different tools, Text-NSP and Termostat for the identification of multiword expressions using different association measures.

1 Introduction

The exact definition of a multiword expression (MWE) is a challenging task and it varies from author to author. For example, Moon (1998) says: "… there is no unified phenomenon to describe but rather a complex of features that interact in various, often untidy, ways and represent a broad continuum between non-compositional (or idiomatic) and compositional groups of words." Moreover, this phenomenon receives different names in the literature (Proost, 2005): phraseological units, fixed expressions, word combinations, phrasemes, etc.

In this study, we consider MWE in a similar way Mel'čuk (1998) defines a phraseme: a phrase which is not free, i.e. the expression's signifier and/or signified are not unrestrictedly and regularly constructed.

A phrase P is unrestrictedly constructed when the rules applied to construct P are not mandatory. For example, instead of the phrase: "doing a research" it is possible to say "performing a research", "executing a research" i.e., this expression is not fixed. However, in a sign like "No smoking", it is not common to see variants like "Smoking prohibited" or "Do not smoke", although those are grammatically correct variants which express the same meaning. Then, "No smoking" is a phraseme (MWE), because it is not unrestrictedly constructed.

A phrase P is regularly constructed when the words forming it are combined following the general rules of the grammar and its sense can be derived exclusively from the sense of its constituent words. The phrase: "he died yesterday", is regularly constructed because it follows the rules of the grammar and its sense follows from the sense the words forming it. However, the expression "kicked the bucket" is not regularly constructed, in relation to its meaning (the combination of words follows the rules of the grammar), because its sense, "died", cannot be derived from the sense of its constituent words. On the other hand, the expression “passing by” is not regularly constructed because it does not follow the general rules of the grammar.

According to Mel’čuk (1998), it is possible to divide the phrasemes (MWEs) in two groups: pragmatemes and semantic phrasemes. As pragmatemes, we can have:

- Expressions in which both the signified and the signifier are not unrestrictedly constructed (although they are regularly constructed), e.g. “all you can eat”, or
- Expressions in which only the signified is not unrestrictedly constructed. For example, in a library it is possible to have signs like “Please be quiet”, “No talking please”, etc. In this case, the signifier (the form, the words forming the expression) is more or less free; however, the sense is always the same.

In semantic phrasemes, the signified is free (it is constructed unrestrictedly; however, it is not constructed regularly) and the signifier is not free. We can have three types of semantic phrasemes:

- Idioms: the sense of the expression goes beyond the sense of its constituent words, and does not include their senses. Examples: “of course”, “(to) pull (someone’s) leg”, “(to) spill the beans”;
- Collocations: the sense of the expression includes the sense of one of its constituent words, say, w1. The other word is freely chosen and w1 is chosen contingent to it. Collocations can be (Manning and Schütze, 1999): light verbs constructions (e.g. make a call, take a decision), verb particle constructions (e.g. to switch on, to pass by), proper names (e.g. San Francisco, Bill Gates) and terminological expressions, i.e., multiword terms (e.g. gross domestic product, light year).
- Quasi-phrasemes or quasi-idioms: the signified of the expression contains the signified of its constituent words; however, it also contains a signified that goes beyond the signified of the isolated words (e.g. (to) start a family, bed and breakfast).

For a more complete explanation about pragmatemes and semantic phrasemes, refer to (Mel’čuk, 1998) or to (Morgan, 1978). For a more detailed linguistic description on the properties of MWEs, see (Baldwin and Kim, 2010).

In this paper, we assume as MWE any kind of phraseme. However, we are interested in the study of Portuguese two-word expressions formed mostly by nouns, adjectives and verbs. For this reason, since most of pragmatemes and idioms are formed by more than two words, basically our focus is on quasi-phrasemes and collocations (mostly light verbs constructions, proper names and multiword terms (MWT)).

The literature on MWE extraction describes different methods for the identification or extraction of MWEs. Many of them rely on association measures, such as Dice’s coefficient (Smadja, 1996) or mutual information (Church and Hanks, 1990). A complete explanation on the use of this association measures on the task of extraction MWEs from text can be found in (Manning and Schütze, 1999). The main idea behind such measures is that the higher the association among the words that appear together in a text, the higher the probability that they constitute a single semantic unit.

There are other methods, which use linguistic information or hybrid approaches that combine statistical measures with the linguistic information, such as the grammatical class of each word, the sense of the expression or the syntactic regularities. Yet others are based on classification algorithms, popular in machine learning systems.

In this study we performed two types of comparison. In the first one, we compared the performance of nine different classification algorithms in the task of identifying MWEs. In the second, we compared two different tools, Text-NSP and Termostat, using different association measures, in the task of extracting MWEs from text. Although our focus is in general MWE, the current study could also be applied to corpus in a specific area for the extraction of multiword terms (MWT).

2 Related Work

Baptista (1994) presents a linguistic study about the nominal expressions formed by more than two words in Portuguese. From a set of 10,000 expressions, he created a typology of nominal MWEs. He found that 70% of the nominal MWEs follow only five different patterns (A = adjective, N = noun, V = verb and P = preposition): A-N, N-A, N-P-N, N-N and V-N. He analyses the syntactic proprieties of each of these groups, focusing his attention on the patterns N-A and N-P-N, which he considers less rigid and more difficult to treat automatically. Finally, he integrates the MWEs' morphological information to an electronic dictionary.

Antunes and Mendes (2013) propose the creation of a MWE typology that includes its semantic, syntactic and pragmatic properties, aiming the annotation of a MWE lexicon using this typology information. They divide the MWEs in three groups, from a semantic standpoint: expressions with compositional meaning, e.g. "banana bread", expression with partial idiomatic meaning, e.g. "vontade de ferro" (iron will) and expressions with total idiomatical meaning (or with no compositionality), e.g. "spill the beans". Within each of these three groups, the expressions are subdivided according to their grammatical categories and lexical and syntactical fixedness.

After a survey and a comparison on different association measures, algorithms and tools used on the identification of MWEs, Portela (2011) presents a study on the identification of Portuguese MWEs following two patterns, N-A and N-P-N, using different association measures. After the extraction of candidates, syntactic criteria are applied to them, to verify their fixedness and determine if a candidate is a MWE. Examples of syntactic criteria applied to bigrams following the pattern N-A and N-P-N:

- Loss of adjective's predicative characteristic: when the adjective comes after the noun and it can be paraphrased by a copulative verb (e.g. verb "to be") + the same adjective, keeping the same sense, the adjective has a predicative function. For example, in the expression: "homem cansado" (tired man, lit. man tired), it is possible to substitute "cansado" for "que estava cansado" (that was tired), and the adjective's predicative characteristic is maintained. However, in the expression "sorriso amarelo" (false, not natural smile, lit. smile yellow), if we substitute the expression for "sorriso que é amarelo", (smile that is yellow), the predicative characteristic is not maintained, because the original sense is lost. This loss of predicative characteristic shows that the expression is fixed, and it is evidence that the expression is a MWE.
- Insertion of elements in the expression (N-P-N): consider the expression "livro de bolso" (pocket book, lit. book of pocket). It is not possible to freely insert a modifier, for example "*livro do Paulo de bolso" (lit. book of Paulo of Pocket). In this example, the modifier can be inserted only at the end of the expression: "livro de bolso do Paulo". This kind of fixedness is evidence that the expression is a MWE.

3 Methodology

We restricted the present study on the extraction of two-word MWEs. For their data, for example, Piao et al. (2003) found that 81.88% of the recognized MWEs were bigrams.

The current study uses CETENFolha (Corpus de Extractos de Textos Electrónicos/NILC Folha de São Paulo) as a Brazilian Portuguese corpus, available on the Linguateca Portuguesa website, which is part of a project on the automatic processing of the Portuguese language (Kinoshita et al., 2006). CETENFolha is composed by excerpts from the Brazilian newspaper "Folha de São Paulo", and contains over 24 million words. At the current stage, we use a small fraction of the corpus, comprising 3,409 excerpts of text (about 250,000 words). Each excerpt corresponds to individual news covering different areas. The number 3,409 represents 1% of the number of excerpts composing the corpus.

We performed different types of evaluation. First, we generated a reference file containing the most frequent MWEs in the corpus and we compared nine different classification algorithms against this reference in the task of identifying Portuguese MWEs. Second, we tested a multilayer perceptron using three different training functions in the task of classifying MWEs in different patterns. We also extracted automatically the 2,000 most frequent bigrams from the entire corpus and we identified, by hand, which ones are MWEs, and we classified them in patterns. Finally, we used two different tools for the identification of MWEs: Text-NSP (Banerjee and Pedersen, 2003) and Termostat (Drouin, 2003). For these tools, we are interested in two types of evaluation. In the first evaluation, we used our reference list to automatically compare the best candidates obtained by each tool against this reference. In the second evaluation, we manually counted the number of MWEs, among a list of the 500-best candidates ranked by one of the association measures, log-likelihood, and we calculated the precision for each tool.

3.1 Reference File Creation

Before the indexation, some pre-processing methods on the corpus were performed, such as lemmatization and elimination of stop words (articles, prepositions, conjunctions). In this study, we are mostly interested in analyzing MWEs formed by nouns, adjectives and verbs. And since those stop words are very common in Portuguese, their elimination reduces considerably the number of MWE candidates that would not be relevant to this study. In this case, some common Portuguese MWEs are not considered, especially the ones following the pattern noun-preposition-noun, e.g. "teia de aranha" (cobweb), or the pattern preposition-noun, e.g. "às vezes" (sometimes).

We obtained 49,589 bigrams and we established a frequency of 3 as a threshold. We selected 1,170 bigrams that appeared more than 3 times in our corpus' excerpts as our MWE candidates, and by hand we recognized 447 of them as Portuguese MWEs, and we considered those 447 MWEs as our reference file.

It is important to note that our reference file does not contain all the two-word MWEs in the corpus' excerpt, since we generated more than 49,000 bigrams, and we could not evaluate all of them by hand. Furthermore, the corpus is formed by newspaper texts, treating different subjects, thus it is more difficult to create a closed set of all possible two-word MWEs. Therefore, our evaluation in the present study is based on a comparison of how many of the most frequent two-word MWEs in our corpus are ranked as *n*-best candidates by some of the association measures implemented by each tool.

3.2 Comparison of Different Classification Algorithms

First, we computed the frequency of each of those 1,170 bigrams and the frequency of its constituent words. Then, we classified by hand each of the words according to their grammatical class: 1 for nouns, 2 for adjectives, 3 for verbs, 4 for other classes (adverbs, pronouns and numbers) and 5 for proper names. We decided not to use a POS-tagger to guarantee the correct grammatical class assignment to each word. This gave us 25 patterns of bigrams: N-N (noun-noun), N-A (noun-adjective), N-V (noun-verb), V-N, PN-PN (proper name-proper name), etc.

Second, we created a matrix of 1,170 lines and five columns. For each line, the first column represents the frequency of a bigram in the excerpt of text, the second column represents the frequency of the first bigram's word, the third column represents the frequency of the second bigram's word, the fourth column represents the grammatical class of the first bigram's word and the fifth column represents the grammatical class of the second bigram's word. This matrix was used to evaluate the precision and recall of nine different classification algorithms: decision tree, random forest, ada boost (using decision stamp as classifier), bagging (using fast decision tree learner as classifier), KNN (K nearest neighbors), SVM, multilayer perceptron, naïve Bayesian net and Bayesian net.

3.3 Bigrams Pattern Classification

We chose one of the algorithms with the best performance (multi-layer perceptron) and we evaluated it using three different training functions, Bayesian regulation back propagation (br), Levenberg-Marquardt (lm) and scaled conjugate gradient (scg), and we compared their performance in the classification of different patterns of bigrams as MWE. The data used for the classification is formatted in the same way as in the Subsection 3.2. However, for this comparison, we used only the patterns that gave 10 or more samples of MWE, for example, the patterns: N-A, N-N and N-PN.

3.4 The Text-NSP Tool

Text-NSP is a tool used in the task of MWE extraction from texts (Banerjee and Pedersen, 2003). In order to use Text-NSP tool, we do not provide a file containing the POS patterns of the bigrams that we would like to extract as MWE candidates. Therefore, before applying this tool, the only pre-processing task we performed with the source corpus, was removing the XML tags they contained. The next step was to define a stop words list file, since we were interested in finding MWEs following the bigram's patterns formed only by nouns, adjectives, verbs and others classes (adverbs, pronouns and numbers), e.g. N-N, N-A, N-V, O-N.

We ran the program using the "count.pl" script, giving the stop words file and the corpus files as parameters, and 2 as n-gram value, which refers to our aim to generate only bigrams.

The output file is a list of all bigrams in the corpus, and each line contains a bigram, the frequency of the bigram, and the frequency of each of the two words forming the bigram.

Using the output file and the "statistics.pl" script, we generated the candidates' files ranked by four different association measures: Dice's coefficient (dice), log-likelihood (ll), pointwise mutual information (pmi) and Student's t-test (t). Then we transformed each of the candidate files to the XML format used by MWEtoolkit (Ramisch, 2012) and used MWEtoolkit's scripts to create files with the *n*-best candidates (*n* = 50, 100, 500, 1000 and 3000) and compare each candidate file against the reference file.

3.5 The Termostat Tool

Termostat (Drouin, 2003) is a tool developed for an automatic extraction of terms. It can be currently used with five different languages: English, French, Italian, Portuguese and Spanish. It generates statistics for simple and complex expressions. Since in this study we are interested in MWE, we extracted only the complex expressions.

As for Text-NSP, Termostat requires the elimination of the XML tags the corpus contained; which was the only pre-processing step of the corpus.

After the analysis of the corpus, the system generated the lists of expressions ranked by four association measures: log-likelihood (ll), chi-squared (χ^2), log-odds ratio (lor) and the "spécificité" measure (Lafon, 1980) (sp).

Then we proceeded as for Text-NSP: we created files with the *n*-best candidates, ranked by the four association measures and compared each candidate file against the reference file.

3.6 Comparison between the 500-best Candidates of each Tool

Using the association measure that is implemented by both tools, the log-likelihood, we analyzed the 500-best candidates ranked by this association measure using each tool. We selected by hand the MWEs among those candidates and we calculated the precision of each tool, for the *n*-best first candidates (*n* = 50, 100, 150…500).

4 Evaluations

4.1 Comparison of Different Classification Algorithms

First, we had to proceed to an indirect estimative of the recall. We found 49,589 bigrams in the selected excerpts of texts, and the manual evaluation of each one, in order to decide which one is a MWE, would take too much time. So, we estimated the amount of MWEs for the total 49,589 bigrams as in (Piao et al., 2003). Using 100 excerpts of text we generated all the bigrams, with all frequencies. We obtained 1,715 bigrams.

Then, we found by hand 136 MWEs, which tells us that about 7.93% of the bigrams are MWEs. Considering that the corpus is homogeneous, we can extrapolate and say that about 7.93% of the 49,589 bigrams in our total excerpts are MWEs, which gives 3,932 MWEs. Since we found 447 MWEs after applying the filter of frequency (> 3), our base recall is 11.37% (447/3,932). We used this base recall as a multiplying factor for the recall given by each classification algorithm.

We used our generated data to test nine different classification algorithms: decision tree, random forest, ada boost, bagging, KNN (K nearest neighbors), SVM, multilayer perceptron, naïve Bayesian net and Bayesian net. The main parameters used with each algorithm are listed below.

Decision tree: C4.5 algorithm (Quinlan, 1993) with confidence factor = 0.25.

Random Forest (Breiman, 2001): number of trees = 10; max depth = 0; seed = 1.

Ada Boost (Freund and Schapire, 1996): classifier = decision stamp; weight threshold = 100; iterations = 10; seed = 1.

Bagging (Breiman, 1996): classifier = fast decision tree learner (min. number = 2; min. variance = 0.001; number of folds = 3; seed = 1; max. depth = -1); bag size percent = 100; seed = 1; number of execution slots = 1; iterations = 10.

KNN (Aha and Kibler, 1991): K = 3; window size = 0; search algorithm = linear NN search (distance function = Euclidian distance).

SVM (Chang and Lin, 2001): cache size = 40; cost = 1; degree = 3; eps = 0.001; loss = 0.1; kernel type = radial basis function; nu = 0.5; seed = 1.

Multilayer perceptron: learning rate = 0.3; momentum = 0.2; training time = 500; validation threshold = 500; seed = 0;

Bayesian net: search algorithm = k2 (Cooper and Herskovits, 1992); estimator = simple estimator (alpha = 0.5).

The results are summarized in Table 1, where Recall-1 is the recall given by each algorithm based on the 447 MWEs found among the MWE candidates and Recall-2 is Recall-1 multiplied by 0.1137 (base recall, as previously calculated), which gives an estimative of the recall for the entire corpus.

As we see in Table 1, the values of precision are very similar for all the algorithms, varying between 0.830 (random forest) and 0.857 (bagging), with the exception of SVM, which gave a precision of 0.738. The recall-1values were between 0.831 and 0.857 (0.655 for SVM) and the recall-2 between 9.4% and 9.7% (7.4% for SVM).

We observe that we obtained good precision and weak recall. This is due, as observed by Piao et al. (2003), to the fact that the extraction of the MWE candidates is based only on the frequency of the bigrams, and only after the extraction of these candidates we applied the linguistic information (classification in grammatical classes).

However, we must consider that, although we extracted only about 11% of the MWEs, these 11% are the most frequent and they represent about 46% of all the MWEs in the corpus, if we sum up the frequency of each MWE. Together, the 447 MWEs found appear 4,824 times in our corpus' excerpt, while the remaining 3,485 (from a predicted 3,932 MWEs in the corpus' excerpt) appear 5,576 times. In absolute terms we have: 4,824 / (4,824+5,576) = 0.46.

Algorithm	TP Rate	FP Rate	Precision	Recall	Recall-2
Decision tree	0.853	0.158	0.854	0.853	0.097
Random forest	0.831	0.194	0.830	0.831	0.094
Ada boost	0.837	0.196	0.836	0.837	0.095
Bagging	0.857	0.163	0.857	0.857	0.097
KNN – k = 3	0.846	0.171	0.846	0.846	0.096
SVM	0.655	0.553	0.738	0.655	0.074
M. perceptron	0.852	0.174	0.851	0.852	0.097
Naïve B. net	0.836	0.170	0.839	0.836	0.095
Bayesian net	0.842	0.170	0.843	0.842	0.096
Base recall	**0.1137**				

Table 1: True-positive rate, false-positive rate, precision and recall for nine classification algorithms.

4.2 Bigrams Patterns Classification

We obtained eight patterns that together represent 59% of the candidate bigrams (689/1,170) and 94% of the MWEs that appear three or more times in the corpus (420/447). The rest of the bigrams' patterns (41%) rarely formed MWE (only 6% of the total MWEs). Table 2 shows the results. "N" stands for "Noun", "A" for adjective, "O" for other classes (adverbs, pronouns and numbers) and "PN" for "proper names".

Analyzing the table, we had best results with the patterns N-A (e.g. "comissão técnica", "banco central", "imposto único") and PN-PN ("Fidel Castro", "José Sarney", "Max Mosley"). The function lm gave the best value for the F1 measure (0.912) for the pattern N-A, and the function scg gave the best value for the pattern PN-PN (0.931).

In general, we had the weakest results with the patterns O-N, e.g. "terceiro mundo", (third world) and A-PN, e.g. "Nova York", "Santa Catarina". Using the training functions "lm" and "scg", none of the 10 MWEs belonging to the pattern O-O, e.g. "até agora" (until now), "além disso" (moreover, lit. beyond this) was recognized, and none of the 46 MWEs belonging to the pattern O-N was recognized, when using the training function "scg".

The last line of each table presents the total values for the eight patterns, for the three learning functions. We had the best precision and recall using the "lm" function.

			br			**lm**			**scg**		
Pattern	**Bigrams**	**MWE**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
N-A	229	193	0.867	0.912	0.889	0.845	0.990	0.912	0.850	0.969	0.906
O-N	164	46	0.378	0.304	0.337	0.647	0.239	0.349	0.720	0.000	0.000
PN-PN	117	101	0.862	0.931	0.895	0.863	1.000	0.927	0.871	1.000	0.931
A-N	53	21	0.813	0.619	0.703	0.810	0.810	0.810	0.630	0.810	0.708
O-O	46	10	0.357	0.500	0.417	0.000	0.000	0.000	0.783	0.000	0.000
N-PN	34	16	0.438	0.438	0.438	0.688	0.688	0.688	0.222	0.125	0.160
N-N	31	20	0.647	0.550	0.595	0.696	0.800	0.744	0.692	0.900	0.783
A-PN	15	13	0.750	0.231	0.353	0.500	0.154	0.235	0.667	0.154	0.250
All Pat.	689	420	0.776	0.769	0.773	0.819	0.831	0.825	0.815	0.779	0.797

Table 2: Multi-layer perceptron precision, recall and *F*-measure in the classification of the most common bigram's patterns using different training functions: Bayesian regulation back-propagation (br), Levenberg-Marquardt (lm) and scaled conjugate gradient (scg).

Using Text-NSP tool, we extracted from the entire corpus all the bigrams (including the ones formed by stop words) and we analyzed by hand the 2,000 most frequent bigrams. We found 165 two-word MWEs formed by nouns, adjectives, verbs and other classes (adverbs, pronouns and numerals) and we classified them according to their pattern. Table 3 shows the number of MWEs and their total frequency in the corpus, classified by patterns. The words belonging to the classes of adverb, pronoun and numeral were classified as "O" (other classes).

The much smaller proportion of bigrams recognized as MWEs (165/2000) in comparison to the previous analysis (447/1,170) is explained by the fact that in the previous analysis we had eliminated the stop words before generating the bigrams, and now all the bigrams were generated. This created many bigrams composed by prepositions or conjunctions that do not form MWE, for example: "de um", "de uma", "de São", "que os", "diz que", "do que", "em que".

We note that the five most common patterns are the same as found before, in the small excerpt of text, with the pattern N-A giving the greatest number of expressions, e.g. "ano passado" (last year, lit. year last), "Banco Central" (Central Bank, lit. bank central), "norte americano" (north American), "seleção brasileira" (Brazilian team, lit. selection Brazilian), "equipe econômica" (economic team, lit. team economic). In terms of frequency, the MWEs following the pattern N-A represent about 38% of the most frequent two-word MWEs found in the corpus.

It is important to observe that, although we are not differentiating Brazilian and Portuguese MWEs in this study, the recognized MWEs follow the Brazilian orthography (e.g. "equipe econômica" vs "equipa econômica", "seleção brasileira" vs "selecção brasileira"), since we used a Brazilian Portuguese corpus.

Pattern	**MWE**	**Frequency**
N-A	58	101,442
O-N	27	29,697
PN-PN	24	39,270
O-O	23	13,923
A-N	13	51,460
N-N	12	21,559
A-O	2	1,975
A-PN	2	2,115
V-N	2	2,263
N-PN	1	2,589
N-V	1	1,423
Total	165	267,716

Table 3: Frequency of the most common MWEs patterns extracted from the entire corpus

4.3 Text-NSP

Before applying this tool, the only pre-processing performed in the corpus was to remove the XML tags. The next step was to define a stop words list file like in Subsections 4.1 and 4.2.

We ran the program using the script "count.pl", giving as parameter the stop word file and the corpus file, and 2 as n-gram value, meaning that we wanted to generate only bigrams.

The exit file is a list of all bigrams in the corpus' excerpt, and each line contains a bigram, the frequency of the bigram, and the frequency of each of the two words forming the bigram.

Using the output file and the script "statistics.pl" we generated the candidates' files ranked by the four association measures listed in Subsection 3.4. Then we transformed each of the candidates' files to the XML format used by the MWEtoolkit and we used the MWEtoolkit's scripts to create files with the *n*-best candidates and to evaluate each of the files against our reference file. Table 4a shows the results of this evaluation.

The results show that for values of *n* = 50, 100 and 500 we had the best results using the log-likelihood measure and for *n* = 1000 and 3000, Student's t-test gave the best results.

Table 4b shows the precision, recall and *F*-measure that we obtained using the log-likelihood measure. We had very good values of precision using the Text-NSP using this measure. For example, from the 50 best ranked candidates by this measure, 31 were MWEs present in our reference list.

4.4 Termostat

Termostat generated n-grams following eleven POS patterns, all of them are nominal ones: N-N, N-A, N-P-N, N-N-N, N-P-N-A, N-N-N-N, N-V-N, N-N-N-N-N, N-A-A, N-N-A and N-A-N. In total, 4,284 n-grams were generated, and we selected only the bigrams (N-N and N-A), which gave 3,458 bigrams (81% of all n-grams). The last five patterns listed above produced less than ten candidates each one and the patterns N-P-N-A, N-N-N-N produced less than 30 candidates each one.

Those 3,458 candidates were ranked according to the four association measures listed in Subsection 3.5. Then we compared the *n*-best candidates against our reference file. The results are in Table 5a. Table 5b shows the precision, recall and *F*-measure that we obtained using the log-likelihood measure.

Looking at Table 5a, we notice that we had best performance with χ^2 for the 50 and 100 best candidates and for the 500, 1000 and 3000 best candidates we had better results using the ll measure.

Comparing with Text-NSP, Termostat had best performance for the first 50 and 100 candidates. However, Text-NSP outperformed for *n* = 500, 1000 and 3000, when using the ll measure and Student's t-test.

	dice	**ll**	**pmi**	**t**
50	7	31	0	23
100	7	64	0	39
500	8	241	1	180
1000	11	314	4	331
3000	69	375	11	392

(a)

ll	**TP**	**Prec.**	**Recall**	**F1**
50	31	0.62	0.07	0.12
100	64	0.64	0.14	0.23
500	241	0.48	0.54	0.51
1000	314	0.31	0.70	0.43
3000	375	0.13	0.84	0.22

(b)

Table 4: Text-NSP: Number of MWEs among the first *n*-best candidates, ranked by four association measures (a) and precision, recall and *F*-measure for the log-likelihood measure (b).

	χ^2	**ll**	**lor**	**sp**
50	42	38	32	38
100	72	68	66	68
500	153	162	117	159
1000	181	197	127	192
3000	198	211	143	208

(a)

ll	**TP**	**Prec.**	**Recall**	**F1**
50	38	0.76	0.09	0.15
100	68	0.68	0.15	0.25
500	162	0.32	0.36	0.34
1000	197	0.20	0.44	0.27
3000	211	0.07	0.47	0.12

(b)

Table 5: Termostat: Number of MWEs among the first *n*-best candidates, ranked by four association measures (a) and precision, recall and *F*-measure for the log-likelihood measure (b).

4.5 Comparing the 500-best candidates of each tool

We analyzed by hand the 500-best candidates obtained using Text-NSP and Termostat, ranked by the log-likelihood association measure, to decide which ones are MWEs. Table 6 shows the precision given by each tool, for the first *n* candidates, *n* = 50, 100, 150…500.

With Termostat, we had the best precision for all values of *n* candidates, going from 86% for the first 50 candidates to 82% for the first 500 candidates. Using Text-NSP, the precision starts with 82% for the first best 50 candidates and decreases to 72% for the first 500-best candidates.

As in the tests performed in Subsection 4.2, the most common patterns of MWE found by both tools were noun-adjective, e.g. "Congresso Nacional", "emenda constitucional", "deputado federal" and proper name-proper name, e.g. "Fernando Collor", "Getúlio Vargas", "Itamar Franco".

***n* first cand.**	**Text-NSP**	**Termostat**
50	0.82	0.86
100	0.82	0.85
150	0.83	0.86
200	0.79	0.84
250	0.76	0.84
300	0.75	0.84
350	0.74	0.83
400	0.74	0.82
450	0.73	0.81
500	0.72	0.82

Table 6: Text-NSP and Termostat precision for the first *n* best candidates, using log likelihood association measure.

5 Conclusions and Future Work

In this paper, we presented a comparative study on different classification algorithms and tools for the identification of Portuguese multiword expressions, using information about the frequency, the grammatical classes of the words and bigrams and different association measures.

In what concerns the classification algorithms, bagging, decision trees and multi-layer perceptron had a slightly better precision. Using multi-layer perceptron with three different training functions, we identified the part-of-speech patterns that are best classified as two-word MWEs. Using the function Levenberg-Marquardt we had better results in classifying the pattern noun-adjective (the most common in our corpus) and we were more successful in classifying MWEs following the pattern "proper name-proper name" using the function scaled conjugate gradient.

With the objective of making an estimative on the part-of-speech patterns followed by the most frequent two-word MWEs in the corpus, we applied Text-NSP to the extraction of the 2,000 most frequent bigrams and we identified and classified the MWEs, according to their part-of-speech patterns. As a result, we found that the patterns "noun-adjective" and "proper name-proper name" are the most common two-word MWE patterns in the corpus. We also found that verbs do not form a great variety of two-word MWE in Portuguese.

The comparison between tools for the automatic identification of MWEs showed that Termostat had better precision than Text-NSP when applied to a small number of candidates (50 and 100). When the number of candidates increases, Text-NSP had better precision using log-likelihood measure and Student's t-test association measures.

As future work, we intend to apply the same tools, especially Termostat, to a specific domain corpus, in order to compare their performance in the identification of Portuguese multiword terms, not limiting the study to bigrams, but also analyzing n-grams in general.

References

Aha, D. and Kibler, D. (1991). Instance-based learning algorithms. *In: Machine Learning*. 6:37-66.

Antunes, S. and Mendes, A. (2013). MWE in Portuguese - Proposal for a Typology for Annotation in Running Text. *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013)*, pp. 87–92, Atlanta, Georgia.

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. Nitin Indurkhya and Fred J. Damerau (eds.), *In: Handbook of Natural Language Processing, Second Ed.* Chapman & Hall/CRC, London, UK., pp. 267-292.

Banerjee, S and Pedersen, T. (2003). The Design, Implementation, and Use of the Ngram Statistic Package. *In: Proceedings of Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 370-381, Mexico City. http://search.cpan.org/~tpederse/Text-NSP/

Baptista, J. (1994). Estabelecimento e Formalização de Classes de Nomes Compostos. Master Thesis. Faculdade de Letras, Universidade de Lisboa, 145 pp.

Breiman, L. (2001). Random Forests. *In: Machine Learning*. 45(1):5-32.

Breiman , L. (1996). Bagging predictors. *In: Machine Learning*. 24(2):123-140.

Chang, Chih-Chung and Lin, Chih-Jen (2001). LIBSVM - A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

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

Cooper, G. and Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *In: Machine Learning*. 9(4):309-347.

Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage, *In: Terminology,* 9(1): 99-117. - http://termostat.ling.umontreal.ca/

Freund, Y. and Schapire, R. E (1996). Experiments with a new boosting algorithm. *In: Thirteenth International Conference on Machine Learning, San Francisco*, pp. 148-156.

Kinoshita, J., Nascimento Salvador, L.D., Dantas de Menezes, C., E. (2006). CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus. *In: Proceedings of Fifth International Conference on Language Resources and Evaluation*, pp. 2190-2193.

Lafon, P. (1980). Sur la Variabilité de la Fréquence des Formes dans un Corpus. *In: MOTS*, no 1, pp. 128-165.

Manning, C. D. and Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press, 1999, 680 pp.

Mel'čuk, I. (1998). Collocations and Lexical Functions. *In: A.P. Cowie (ed.), Phraseology. Theory, Analysis, and Applications*, 1998, Oxford: Clarendon Press, pp. 23-53.

Moon, R. E. (1998). Fixed Expressions and Idioms in English: A Corpus Based Approach. Oxford: Clarendon Press, 356 pp.

Morgan, J. L. (1978). Two Types of Convention in Indirect Speech acts. *In: P. Cole (ed.), Syntax and Semantics, v.9. Pragmatics* (New York etc.: Academic Press), pp. 261-80.

Piao, S., Rayson, P., Archer, D., Wilson, A., and McEnery, T. (2003). Extracting Multiword Expressions with a Semantic Tagger. *In: Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, at *ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics*, pp. 49-56, Sapporo, Japan.

Portela, R. J. R. (2011). Identificação Automática de Nomes Compostos. Instituto Superior Técnico, Universidade Técnica de Lisboa. Master Thesis. November 2011, Lisbon, Portugal, 104 pp.

Proost, K. (2007). Conceptual Structure in Lexical Items: The Lexicalisation of Communication Concepts in English, German and Dutch. John Benjamins Pub. Co, 304 pp.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 303 pp.

Ramisch, C. (2012). A Generic and Open Framework for MWE Treatment – From Acquisition to Applications - Ph.D. Thesis, Universidade Federal do Rio Grande do Sul - UFRGS, Brazil, 248 pp. http://mwetoolkit.sourceforge.net/PHITE.php?sitesig=MWE

Smadja, F. A. (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Association for Computational Linguistics*, 22 (1):1-38.