Using machine learning methods to avoid the pitfall of cognates and false friends in Spanish-Portuguese word pairs

Lianet Sepúlveda Torres[1], Sandra Maria Aluísio[1]

[1] Núcleo Interinstitucional de Linguística Computacional/Interinstitutional Center for Research and Development in Computational Linguistics ICMC-University of São Paulo São Carlos – SP – Brazil

lisepul@icmc.usp.br, sandra@icmc.usp.br

Abstract. *The fact that 85% of the Portuguese lexicon contains Spanish cognates and that the linguistic structures of both languages are highly coincident is believed to be an advantage for the Spanish speaker who learns Portuguese. However, these similarities have some negative aspects in the learning of Portuguese, such as, the pitfall of false friends, since about 20% of cognates are false. The aim of this article is to identify cognates and false friends between Spanish and Portuguese automatically to build dictionaries of these words. One of the uses for these dictionaries is to support scientific writing tools, which can help lower barriers for Spanish speakers when they write in Portuguese.*

1. Introduction

As a result of the globalization process in society, learning a new language became a requirement for all researchers, especially when they intend to study or work abroad. In general, this new language is English, but this is not always the case in Latin America, for example. This phenomenon can be observed in Brazilian universities, where the number of foreign students is high and has increased over the past years, with a majority of native Spanish-speaking students. This article addresses problems these native Spanish speakers face when writing their theses and dissertations in Portuguese, a requirement in Brazilian university graduate programs.

Since Portuguese and Spanish are the closest romance languages [Henriques 2000], several investigations have come to the conclusion that Spanish speakers have different characteristics in relation to other Portuguese language learners [Mohr 2007]. In the literature, the similarity between these two languages is considered as a positive element that often becomes an obstacle, because similarity and closeness frequently conceal differences and hinder learners from mastering the Portuguese, keeping interferences from their native Spanish both when speaking and writing in Portuguese [Santos 1999, Gomes 2002].

The fact that 85% of the Portuguese lexicon contain Spanish cognates and that the linguistic structures of both languages are highly coincident is believed to be an advantage for the Spanish speaker who learns Portuguese [Henriques 2000, Santos 1999]. The awareness that there are words in two languages that share similar meaning, spelling and pronunciation – the so-called cognates – can indeed be used as a tool to understand a second language. However, these similarities have some negative aspects in the learning of Portuguese, such as: (i) the learner is prevented from perceiving in which language he/she is communicating; (ii) the pitfall of false friends, since about 20% of cognates are false [Henriques 2000, Santos 1999]; (iii) interlanguage fossilization in the beginning of the learning process, as a result of

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 67–76,
Cuiabá, MT, Brazil, October 24–26, 2011. ©2011 Sociedade Brasileira de Computação

mutual understanding among interlocutors; (iv) different pronunciation patterns in these two languages, given the small number of vocalic phonemes in Spanish, the fact that there are no open or nasal vowels in Spanish, as well as all open diphthongs of Portuguese; and (v) total acceptance of Spanish speakers' interlanguage by Portuguese native speakers, which makes learners feel satisfied with their level of proficiency [Santos 1999].

The aim of this article is to identify cognates and false friends between Spanish and Portuguese automatically to build dictionaries of these words. One of the uses for these dictionaries is to support scientific writing tools, which can help lower barriers for Spanish speakers when they write in Portuguese, in view of the fact that in the process of learning a new language, writing and speaking (production skills) are more difficult than reading and listening (reception skills). Although there exists a few Spanish-Portuguese cognates and false friends dictionaries, they are not in digital format, are expensive and usually restricted to general language.

Here, we consider Frunza and Inkpen's (2009) definition of cognates and false friends, in which: **Cognates**: are pairs of words that are perceived as similar and are mutual translations. Pronunciation can be identical or not, for example: *amor-amor* and *jefe* (in Spanish) - *chefe* (in Portuguese). Frunza and Inkpen's (2009) also consider as cognates, word pairs that are orthographically identical or have slightly different spellings. **False Friends**: are pairs of words in two languages that are perceived as similar, but have different meanings depending on the context, for example: *aula* (class, in Portuguese) - *aula* (classroom, in Spanish).

In this study, we propose to identify cognates and false friends based on orthographic and phonetic similarity measures that are going to be used as classification features in machine learning (ML) algorithms. Also, we introduce a semantic feature as a classification attribute, with the objective of improving algorithm's performance. We evaluated the impact of selecting different features, the averaging of them and combining them with several classification algorithms to identify cognates and false friends. Our data set is composed of word pairs divided into three classes: Cognates, False Friends and Unrelated. Included in the latter are the words that are mutual translations, but do not have orthographic or phonetic similarity.

In Section 2, we present works related to the identification and disambiguation of cognates and false friends. In Section 3, we introduce our approach to the identification of cognates and false friends for the language pair Portuguese-Spanish. In Section 4, the results of attribute selection and classification experiments are shown and Section 5 is the conclusion of our study and future work we envisage.

2. Related Work

Different natural language processing applications concentrate on cognate identification. Some of these applications include: sentence alignment [Simard and Isabelle 1992, Melamed 1999], parallel text alignment [Gomes 2009], inducing bilingual lexicons [Mann and Yarowsky 2001], and identification of confusable drug names [Kondrak and Dorr 2004].

There are several approaches to identify cognates between language pairs [Simard and Isabelle 1992, Kondrak 2001, Kondrak 2005, Inkpen's and Frunza 2005, Frunza and Inkpen's 2009]. Usually, the methods to compute similarity among words are divided into orthographic and phonetic. Some of the methods in the first group are EDIT distance, LCSR (longest common subsequence ratio) and measures based on the number of n-grams that are shared by words. Other frequently employed measure is the binary identity function. The most well-known phonetic approaches are Soundex and Editex, which try to take advantage of individual features to determine similarity between words [Wesley and Kondrak 2005].

Simard and Isabelle (1992) use cognates to align sentences in bitexts. These authors worked with the language pair French-English and considered as cognates those words in which the first four characters were identical. Kondrak (2001) starts from the premise that cognate words show phonetic and semantic similarity. Following this idea, he introduces a procedure for identifying related words estimating the semantic similarities of glosses. Kondrak identifies the keyword of a gloss and, employing WordNet [Fellbaum 1998], searches for synonyms among glosses and, then, similarities between words. In addition, he uses orthographic and phonetic similarity measures in his study. In [Kondrak 2005], the focus is on cognate identification based on orthographic similarity measures, with the aim of favoring word alignment. The author evaluates different similarity measures, especially those based on computing the longest common sequence between words (LCSR). Besides, he evaluates other orthographic similarity measures, such as PREFIX, DICE coefficient, and IDENT.

Frunza and Inkpen's (2009) offer an approach to identify cognates and false friends estimating different distance measures between word pairs. These authors propose to employ (ML) techniques to sort cognates from false friends. Their study is on the language pair French-Portuguese and considers orthographic and phonetic similarity measures as classification features. The measures used are IDENT, PREFIX, DICE, TRIGRAM, XDICE, XXDICE, LCSR, NED, SOUNDEX, in addition to other measures considered as generalizations of the metrics LCSR and NED, defined in [Kondrak and Dorr 2004]. Frunza and Inkpen's (2009) propose to combine several techniques to disambiguate cognates and false friends. According to these authors, there are several investigations dedicated to build cognate lists, but few studies center on the identification of false friends.

Here we expect our algorithm can sort unrelated words from cognates and false friends without much effort, but, given the orthographic similarity between cognates and false friends, our main problem is to identify if the orthographic measures employed to characterize data can distinguish these words.

3. Our approach to identifying Cognates and False Friends

3.1. Resources

This study's training set is made of 948 word pairs in Portuguese and Spanish. These pairs were classified manually in Cognates, Unrelated words, and False Friends for two annotators, the first native of Portuguese and the other native of Spanish. These data are illustrated in Table 1. The word pairs were selected from the following resources: online Spanish-Brazilian Portuguese dictionary[1]; online Spanish-Portuguese dictionary[2]; list of most frequent words in Portuguese and Spanish[3]; and online list of different words in Portuguese and Spanish[4].

[1] http://www.myzips.com/software/ACCESS-Wordlist-Spanish-Portuguese.phtml

[2] http://en.bab.la/dictionary/spanish-portuguese/

[3] http://www.wordfrequency.info/spanish_portuguese.asp

[4] http://rudhar.com/lingtics/ptesdiff.htm

Table 1. Data set (numbers in parenthesis represent the total of words with identical spelling).

Cognates	338 (122)
Unrelated words	238 (0)
False Friends	372 (199)
Total	948

As illustrated in Table 1, there is an unbalance in the training set favoring the classes Cognates/False Friends. There are no multiword expressions in the data set. In Table 1, it can be observed that there are 122 words classified as cognates with identical spelling and 199 words classified as False Friends showing the same feature, which increases the complexity of our classification problem.

3.2. Similarity Measures

The measures used in the task of classifying cognates, false friends and unrelated words are explained below.

IDENT is a measure employed as baseline. It returns 1 when words are identical, otherwise it returns 0. *PREFIX* calculates the length of the prefix common to two words divided by the length of the longest word. For example, the prefix common to the words *mejorar* and *melhorar* is 2 (the first two letters) divided by 8 (the longest length between these two words), that is, 0.25. *DICE* divides the amount of bigrams common to two words by the sum of bigrams in each word [Frunza and Inkpen's 2009].

$$\frac{2|\mathrm{bigrams}(x) \cap \mathrm{bigrams}(y)|}{|\mathrm{bigrams}(x) + \mathrm{bigrams}(y)|} \quad (1)$$

in which, bigrams(x) is the set of bigram characters in word x. For example, the DICE's coefficient for the words *mejorar* and *melhorar* = 8/13 = 0.61 (the shared bigrams are me-or-ra-ar).

TRIGRAM is defined in the same way as the DICE's coefficient, but it works with trigrams. *LCSR* (Longest Common Subsequence Ratio) is a measure largely employed to evaluate word similarity [Kondrak and Dorr 2004, Wesley and Kondrak 2005, Frunza and Inkpen's 2009]. It searches for the longest sequence common to two words and divides it by the length of the longest word [Melamed 1999]. For example, LCSR (*mejorar,melhorar*) = 6/8 = 0.75. *NED* (Normalized EDIT Distance) calculates the minimal number of necessary operations to transform a word into another. In its ordinary definition, each operation (insertion, elimination or substitution) has a fixed cost of 1. In the case of the normalized EDIT distance, the sum of all costs is divided by the length of the longest sequence.

SOUNDEX: The phonetic algorithm SOUNDEX was proposed for English, with the aim of indexing names by similar pronunciation. This algorithm transforms all letters of a word, following a numerical code, keeping the first letter of the word. After this transformation, the code is truncated, and the word is represented by four characters. In this article, we adapted the SOUNDEX algorithm for Portuguese and Spanish. To this end, it was necessary to take different transformation rules into account. In the construction of the algorithm in Portuguese, we considered the rules defined by [Chbane 1994]. The SOUNDEX for Spanish was implemented in python, following the instructions available in http://oraclenotepad.blogspot.com/2008/03/soundex-en-espaol.html. To estimate the similarity between the codes generated, we employed the NED.

3.3. Methods

Using the Weka package [Witten and Frank 2005], we evaluate the performance of several supervised algorithms to single out the proposed classes.

3.3.1. Instances and the Vector of Features

Machine Learning algorithms require a specific format to process data. In this study, the instances to be classified are word pairs in Spanish and Portuguese. Each pair is labeled according to one of the classes defined previously. The set of these instances forms the data set to be analyzed and described below. For each pair of word in Portuguese and Spanish, we estimated a total of seven similarity measures. These measures configure the vector that characterizes the word pairs. As an additional measure, we considered the average of these similarity measures; therefore, our data are described by a total of eight features. Thus, we represent our data set as a matrix in which each row contains word pairs and the columns include the proposed similarity measures. These similarity measures are the attributes employed by the classification algorithm to identify each one of the proposed classes. This data set is modified for the different experiments we conducted. Next, we discuss these experiments in detail.

3.4. Experiments

Each one of the experiments takes into account a variation of the data set, aiming to achieve a better classification performance. In each experiment, several classification algorithms are evaluated. One of our goals is to calculate the influence of the proposed similarity measures and the averaging on class differentiation.

Experiment 1: In this first experiment, the data set was divided into two classes: Cognates/False Friends and Unrelated. It is possible to combine the classes Cognates and False Friends, because the orthographic features of these two sets are very similar. In this experiment, it is easy to perceive an increase in class unbalance, but it is expected that this fact does not affect the performance of classifiers, given that orthographic similarities between unrelated words are more distant.

Experiment 2: After the classification performed in the first experiment, we intend to know whether the proposed classification algorithms can adequately divide the data set into two classes. This time, the set is composed of Cognates and False Friends. It is expected that this experiment produces the worst classification result, since the measures of orthographic similarity might not be able to perform the differentiation.

4. Experiment Results

4.1. Attribute Selection

Before evaluating the performance of classifiers in each experiment, we used an attribute selection technique to observe the importance of the proposed similarity measures for class identification. The algorithm chosen to rank the attributes was the InfoGainAttributeEval from Weka. This algorithm evaluates attributes by measuring their information gain with respect to the class [Lue et al. 2010]. In Figure 1, one can observe the behavior of the eight attributes in each experiment. The similarity measure LCSR is very well ranked in the two experiments. In Experiment 1, the lowest ranked attributes were incrementally eliminated until the attribute LCSR remained as the only classification feature. The algorithm Instance-based learning (IBK), with K=3, maintained a performance of 95.56%, using the LCSR as the unique classification attribute. Thus, it is possible to see why it is one of the most used attributes in the text alignment and words similarity studies. The EDIT measure and the average of all similarity measures proved to be decisive attributes in the classification of algorithms. As expected, we observed that the measure that provides less information is IDENT. According to the attribute selection algorithm, the similarity measures do not provide much information for the classification algorithm in Experiment 2. We carried out several tests with classification algorithms, eliminating less significant attributes, but our results changed only in the decimal places. For this reason, we are considering all attributes in the vector of features.

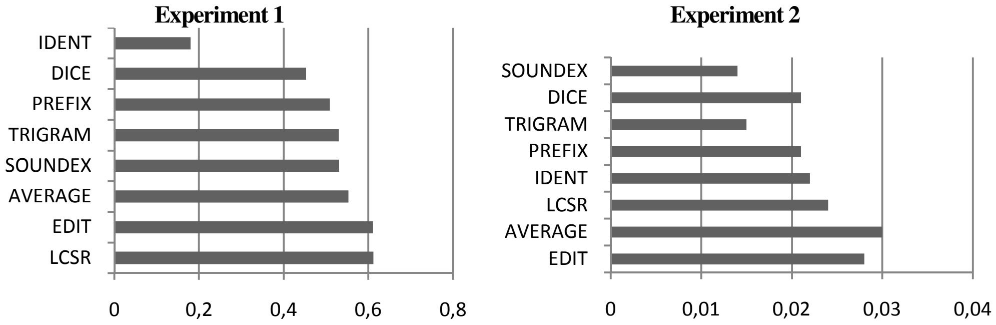


Figure 1. Results of the attribute selection technique for each experiment

4.2 Classification Algorithms

In the proposed experiments, we used the 10-fold cross-validation technique to sample data. Thus, the training set is divided into ten partitions in each run, using nine partitions for training and one for testing. From the Weka package we trained several classification algorithms: the probabilistic method Naïve Bayes, IBK, Multilayer Perceptron, Suport Vector Machine (SMO), Rule-based Method (JRip) and the J48 decision tree classifiers. Several modifications were made in these algorithms' parameters, producing a better performance for our data set. For example, in the case of IBK algorithm, we observed performance variations for different K (number of neighbors). In Experiment 1, the best result were with K=3, whereas in Experiment 2 K=9.

Table 2 illustrates the results obtained from each experiment after using the classification algorithms. The best performance of all was achieved by the algorithm IBK, and the worst was that of Naïve Bayes. It can be observed that the proposed classifiers were able to distinguish the classes Cognates/Unrelated with a high hit ratio, in spite of the unbalance in this experiment (Experiment 1). This result is related to the fact that unrelated words have low orthographic similarity, whereas cognates and false friends are very similar. In Experiments 2, it can be observed that the algorithms' performance dropped significantly, since the main bottleneck of our problem is the ability to distinguish cognates from false friends. This was an expected result, after the attribute selection shown in Section 4.1, in which we observed that the similarity measures would not provide much information for classification. We examine the misclassified pairs for the classifiers that performed better, IBK. The number of false positive in Experiments 2 is very high. The classifier makes many mistakes when differentiating the classes Cognates/False Friends.

Table 2. Performance of the classifiers tested in the proposed experiments

Classifiers	Experiment 1	Experiment 2
	Instances classified correctly	
Naive Bayes	93.8819%	57.1831%
IBK	**95.9916%**	**63.5211%**
SVM	95.8861%	61.6901%
MultilayerPerceptron	95.6751%	62.8169%
JRIP	95.2532%	61.9718%
J48	95.5696%	61.2676%

These mistakes were mostly caused by the high number of identical words in the set of Cognates and in the set of False Friend. In view of the unsatisfactory results obtained in Experiments 2, we proposed a third experiment, in which we consider new classifications attribute, based on the semantic features of words.

4.3 Experiment 3: the contribution of a new feature

In this experiment, a new attribute - translation - is added as a new data feature. An automatic search for the meanings of words in our data set was conducted in a Spanish-Portuguese dictionary generated with NATools[5]. It is expected that cognate words and unrelated words show as mutual translations in the dictionary, whereas false friends should return different meanings, depending on the context in which they are being used. This dictionary was built using a parallel text corpus extracted from the journal Pesquisa FAPESP[6]. Therefore, the words can have several meanings, which will depend on the likelihood of a certain word and its translation being in our corpus. We used the likelihood of the occurrence of the two words as a classification feature. If a word is in the dictionary it is likely to occur, otherwise the likelihood is zero. One of the deficiencies of this experiment is that many of the words in our list are not in the corpus. Table 3 shows the number of words that are in the dictionary.

Table 3. Number of words in the dictionary (number in parenthesis represents words from the data set that are not in the dictionary).

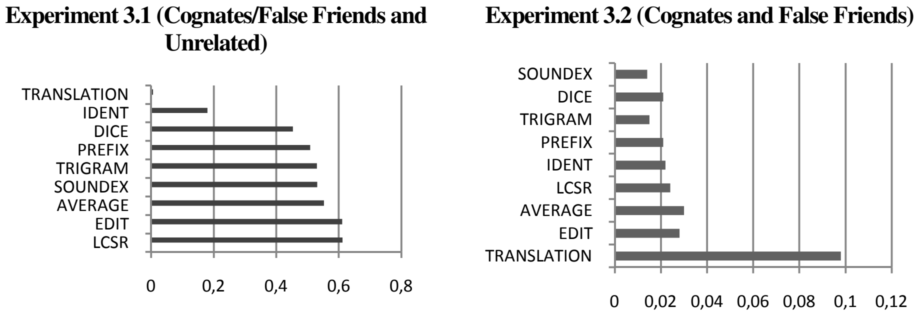
Cognates	338 (116)
Unrelated words	238 (95)
False Friends	372 (172)
Total	948

4.3.1. Results

Attribute Selection: The attribute selection was carried out for a data set divided into two classes (Cognates/False Friends and Unrelated) in Experiment 3.1; and for a data set divided into Cognates and False Friends, Experiment 3.2. In this way it is possible to demonstrate in Experiment 3.1 (Figure 2) that the worst ranked attribute was translation, whereas in Experiment 3.2 (Figure 2) it was best ranked. This is an interesting result, because it confirms that the introduction of semantic evidence in cognate detection helps to substantially increase the precision of cognate identification [Mulloni et al. 2007].

[5] NATools is a set of tools development to work with parallel corpus, which is freely available at http://natura.di.uminho.pt/natura/natura/ under the specifications of the GNU-General Public License

[6] URL of version online from the journal Pesquisa FAPESP: http://revistapesquisa.fapesp.br



Experiment 3.1 (Cognates/False Friends and Unrelated)

Experiment 3.2 (Cognates and False Friends)

Figure 2. Attribute selection in Experiment 3

Table 4 shows the performance of classification algorithms in Experiment 3.2. It can be observed that the inclusion of the new translation attribute improved the performance of all classifiers. Again, IBK performed better, while Naïve Bayes produced the worst result. We believe that if the number of words from the training set included in the dictionary was higher, the algorithms' performance would have improved as well. In Experiment 3.2 the rate of false positive is still high, but it improved significantly when compared with the rate of false positive obtained for Experiment 2. Mulloni et al. (2007) combined orthographic and semantic evidence on words of different languages to identify cognates and false friends. The precision of these results for English-Spanish's languages is an average 91.85%, but it is - important to note that the authors try to deal with the problem of the absence of words in the dictionary, one of the deficiencies of our experiment.

Table 4. Performance of classifiers tested in the experiment 3.2

Classifiers	Experiment 3.2
	Instances classified correctly
Naive Bayes	71.73%
IBK (K = 9)	**76.3713%**
SVM	72.1519%
MultilayerPerceptron	75.7384%
JRIP	74.6835%
J48	74.2616%

5. Conclusions and Future Work

This article is the starting point for researching the construction of several linguistic resources to be used in natural language processing tools, especially scientific writing tools for Spanish speakers who write in Portuguese. With this initial proposal, we tried to evaluate the performance of (ML) algorithms in the identification of cognates and false friends, based on orthographic similarity measures, confirming that these measures are not sufficient for this type of classification; however, when combined with other measures, they can achieve more satisfactory results. A result from this investigation was the development in python of the algorithm Soundex for Spanish and Portuguese, which is freely available. The fact that we worked with a Spanish-Portuguese dictionary to search for the most probable word meanings in improved the algorithm' performance, but it was not enough, since many of the words in our data set were not in the corpus employed. For this reason, one of the future studies we propose is to increase our study corpus and, consequently, our dictionary. Moreover, we propose to increase the number of words in the training set to build more robust classification models that make possible to classify new data.

Acknowledgments

We thank the Brazilian Science Foundation FAPESP for financial support.

References

Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. and Scuse, D. (2010) “WEKA Manual for Version 3-6-4”, The University of WAIKATO.

Chbane, T.D. (1994) “Desenvolvimento de sistemas para conversão de textos em fonemas no idioma português”, Master Degree thesis, Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil.

Fellbaum, C. (1998) “WordNet: An electronic lexical database”, MIT Press, Cambridge, Massachusetts.

Frunza, O. and Inkpen's D. (2009) “Identification and Disambiguation of Cognates, False Friends, and Partial Cognates Using Machine Learning Techniques”, International Journal of Linguistics, vol. 1, no. 1, p. 1-37, Otawa, Canada.

Gómes, V.F.P.G. (2002) “Característica da Interlíngua Oral de Estudantes de Letras/ Espanhol nos dois últimos semestres de estudo”, Congresso Brasileiro de Hispanistas.

Gomes, L. (2009) “Parallel Texts Alignment”, Master Degree thesis, Universidade Nova de Lisboa, Lisboa, Portugal.

Henriques, E.R. (2000) “Intercompreensão de Texto Escrito por Falantes Nativos de Português e de Espanhol”, DELTA, vol. 16, no. 2, p. 263-295.

Inkpen's, D. and Frunza, O. (2005) “Automatic Identification of Cognates and False Friends in French and English”. In: Proceedings of the Recent Advances in Natural Language Processing, (RANLP).

Kondrak, G. and Dorr, B.J. (2004) “Identification of confusable drug names: A new approach and evaluation methodology”, In: Proceedings of The Twentieth International Conference on Computational Linguistics, p. 95-958, (COLING 2004), Geneva, Switzerland.

Kondrak, G. (2001) “Identifying Cognates by Phonetic and Semantic Similarity”. In: Proceedings of The 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, p. 103-110.

Kondrak, G. (2005) “Cognates and Word Alignment in Bitexts”. In: Proceedings of The Tenth Machine Translation Summit, p. 305-312. (MT Summit X), Phuket, Thailand.

Lue, Y., Ben, K. and Mi, L. (2010) “Software metrics reduction for fault-proneness prediction of software modules”. In: Proceedings of The 2010 IFIP International Conference on the Network and Parallel Computing, p.432-441.

Mackay, W. and Kondrak, G. (2005) “Computing Word Similarity and Identifying Cognates with Pair Hidden Markov Models”. In: Proceedings of The Ninth Conference on Computational Natural Language Learning, p. 40-47. (CoNLL2005), Ann Arbor, Michigan.

Mann, G.S. and Yorowsky, D. (2001) “Multipath translation lexicon induction via bridge languages”. In: Proceedings of 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL’01), p.1-8.

Melamed, I.D. (1999) “Bitext maps and alignment via pattern recognition”, Computational Linguistics. 25, p.107-130.

Mohr, D. (2007) "Português para Hispanofalantes: Uma Alternativa para o Ensino de Gêneros Escritos", Encontro de Professores de Línguas Estrangeiras do Paraná Línguas: culturas, diversidade, integração (XV EPLE), p. 372-387.

Mulloni, A., Pekar, V., Mitkov, R. and Blagoev, D. (2007). "Semantic Evidence for Automatic Identification of Cognates". In: Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons, p. 49-54, Borovets, Bulgaria.

Santos, P. (1999) "O ensino de Português como Segunda Língua para Falantes de Espanhol: Teoria e Prática", Em CUNHA, M.J. e SANTOS, P. (orgs.) Ensino e Pesquisa em Português para Estrangeiros. Editora UnB, Brasília, Brazil.

Simard, M.F. and Isabelle, P. (1992) "Using cognates to align sentences in bilingual corpora". In: Proceedings of The 4th International Conference on Theoretical and Methodological Issues in Machine Translation, p. 67-81, Montreal, Canada.

Witten, I.H. and Frank, E. (2005) "Data Mining: Practical machine learning tools and techniques" 2nd Edition. Morgan Kaufmann, San Francisco.