Generating a Lexicon of Errors in Portuguese to Support an Error Identification System for Spanish Native Learners

Lianet Sepúlveda Torres, Magali Sanches Duran and Sandra Maria Aluísio
Núcleo Interinstitucional de Linguística Computacional (NILC)/Interinstitutional Center for
Research and Development in Computational Linguistics ICMC-University of São Paulo São Carlos-SP-Brazil
lisepul@icmc.usp.br,magali.duran@uol.com.br,sandra@icmc.usp.br

Abstract

Portuguese is a less resourced language in what concerns foreign language learning. Aiming to inform a module of a system designed to support scientific written production of Spanish native speakers learning Portuguese, we developed an approach to automatically generate a lexicon of wrong words, reproducing language transfer errors made by such foreign learners. Each item of the artificially generated lexicon contains, besides the wrong word, the respective Spanish and Portuguese correct words. The wrong word is used to identify the interlanguage error and the correct Spanish and Portuguese forms are used to generate the suggestions. Keeping control of the correct word forms, we can provide correction or, at least, useful suggestions for the learners. We propose to combine two automatic procedures to obtain the error correction: i) a similarity measure and ii) a translation algorithm based on aligned parallel corpus. The similarity-based method achieved a precision of 52%, whereas the alignment-based method achieved a precision of 90%. In this paper we focus only on interlanguage errors involving suffixes that have different forms in both languages. The approach, however, is very promising to tackle other types of errors, such as gender errors.

Keywords: artificial generation of errors, interlanguage errors, identification and correction of errors

1. Introduction

Any tool conceived to support writing activities depends on the correct identification of errors before providing suggestions or automatic correction. The automatic detection of errors requires a huge amount of knowledge since the nature of the error can be lexical, grammatical, semantic and/or discursive, which represents a challenge for developers of tools for natural language processing (Nagata et al., 2011).

Regarding foreign language learners, the process of error detection is more complex than for native speakers (Leacock et al., 2010). Language learners errors may be common errors made even by native speakers, or errors made by most non-native speakers, or errors made by a language specific native speakers group or, finally, be idiosyncratic, that is, errors not shared by other learners.

In general, the error analysis consists of manually annotating a learner corpus with a predefined set of error tags (Dahlmeier & Tou Ng, 2011; Genoves et al., 2007), however, it is not easy to categorize learners' errors. Dagneaux et al. (1998) and Tono (2003) show how a typology of errors may be used to annotate learners' corpora and to support automatic analysis of errors. However, it is difficult to gather learners' writing samples with original errors, because many of them are already edited by using language checkers available in the most popular text editors, thus masking the errors one would like to detect.

In addition, studies on detecting learner errors employ expensive computational tools, such as parsers and part-of-speech taggers to support the task. Nagata et al. (2011) argue these computational resources may add errors in the process of detecting errors, causing a drop in performance of the algorithms. In addition, the errors' detection approaches based on statistical methods require the learner corpus be annotated with a large number of different types of errors to maintain their performance (Leacock et al., 2010).

Summing up, on the one hand it is difficult to predict and to categorize learner's errors and it is very time consuming and costly to annotate learners corpora. On the other hand, the algorithms that detect learners' errors need a large and varied amount of errors examples. In such a scenario, therefore, the approach of automatically generating a lexicon of likely learners' errors is a bootstrap to the construction of language tools tailored for foreign learners.

In what concerns Portuguese, foreign learners lack language resources that support their learning and writing activities. Grammar and spelling checkers are available in the most popular word processors; however, they are designed to deal with typical native speaker's types of errors and are of no help to tackle errors typical of foreign learners.

Faced with this situation and motivated by the increasing interest of native Spanish speakers in learning Portuguese, Sepúlveda-Torres et al. (2014) is developing HABLA (Hispanic speakers purchasing a Base Academic Language), a system designed to support scientific written production of native Spanish learners of Portuguese, to complement the support provided by existing grammar and spelling checkers. The research reported herein is intended to inform a module of such system to deal with errors involving suffixes.

One of the HABLA functions is to detect and suggest corrections for lexical errors. Automatic error detection systems may produce two types of feedback: 1) to classify the input material as correct or incorrect and 2) to suggest the correct form. In what concerns lexical errors, the first feedback may be produced by verifying whether a word form belongs or not to a dictionary of word forms in the target language. The second type of feedback is addressed automatically using similarity measures, but such

approach assumes the writer made a spelling error and he/she is able to judge whether one of the proposed similar words is suitable to convey the intended meaning.

In foreign learners writing context, however, the learner almost always has a limited knowledge of the second language vocabulary and he/she is not able to recognize the similar words suggested and even less to choose one of them (Duran, 2008). Besides that, the error may be not a spelling error, but a trial to "guess" the second language word equivalent to the word in the native language. This behavior was observed in the *Espanhol-Acadêmico-Br* corpus of Spanish native learners of Portuguese compiled by Sepúlveda-Torres et al. (2014). As both languages are very close, learners are unwilling to consult a dictionary, because they know there is a high probability of "guessing" the right equivalent in Portuguese.

Spanish and Portuguese languages share a lot of cognates: 85%, according to Santos (1999). Part of them are true cognates, part are false. The problem of false cognate's identification has been addressed by Sepúlveda-Torres & Aluísio (2011). The analysis of the *Espanhol-Acadêmico-Br* learner corpus, however, led us to identify another type of "guessing" strategy employed by Spanish native speakers when they write in Portuguese: they learn the equivalent suffixes in both languages and produce a new word substituting the Spanish suffix for the Portuguese one. For example, words ending in "*-dad*", as "*felicidad*" produce words ending in "*-dade*" in Portuguese, as "*felicidade*" (happiness). This strategy is almost always well succeeded, but when it is not, the spelling checker may not be able to suggest the right word, as the wrong form produced belongs neither to Spanish nor to Portuguese language: it is a wrong word produced in the interlanguage by interference of the native language. An example of such situation is "*vecindad*", which the learner may use to produce "*vecindade*", a word that does not exist in Portuguese. The right equivalent, in this case, is "*vizinhança*" (neighborhood).

Therefore, the spelling checker may detect such kind of lexical error, but it is of no help in what concerns suggesting corrections. To tackle with this problem, we need to recognize the interlanguage words produced by learners. We automatically generated a lexicon of wrong words likely to be produced by Spanish native learners of Portuguese, using the same reasoning observed in the learner's corpus. In this way, we keep control of the correct word form, using it to provide the correction or, at least, useful suggestions for the learners.

The lexicon of Spanish-native-like errors in Portuguese we produced is composed by three forms: the Spanish form, the interlanguage (wrong form) and the Portuguese form. The interlanguage form is used to identify the error; the Spanish form (SF) is used to produce the suggestion: 'Do you intend to say "SF" in Portuguese?' 'The equivalent may be "PF"' (the Portuguese form).

The remainder of this paper is structured as follows. Section 2 briefly reviews related work regarding artificial generation of lexical errors. In Section 3 we present the resources and the procedures used to create the lexicon of interlanguage errors, which support the task of identification and correction of lexical errors. In Section 4 we show the results of our experiment, which are discussed in Section 5. Finally, in Section 6, we conclude with a summary of our findings and an outline of future work.

2. Related Work

Our research is related to others that use artificial generation of errors to bootstrap the lack of large learner corpora containing error annotation. NLP systems that deal with text correction need both positive and negative evidence (examples of well written texts and examples of errors), but negative evidence is useless if it does not represent plausible errors. Foster & Oistein (2009) stress that "artificial errors need to be tailored for the task at hand", otherwise the accuracy of classification methods may drop when applied to real learner texts. They present a tool, called GenERRate, which is used to produce different types of syntactically noisy data to classify English sentences in grammatical or ungrammatical.

The same approach has been used for Russian by Dickinson (2010), focusing on combinations of a stem and a suffix with the purpose of creating realistic data for machine learning systems. As the random combinations of a stem and a suffix can result in many unlikely errors, he guided the combinations, using a loose notion of likelihood to ensure that the errors fall into a reasonable distribution.

Such researches differ from ours in what concerns their purpose, as they intended to produce errors in context to provide negative evidence for machine learning. Our purpose, on the other hand, is to produce a lexicon to inform an error detection and correction system. Other difference is the fact that our approach is informed by real learner's errors from a specific native language group of learners (Spanish native speakers), whereas Foster & Oistein (2009) used a corpus containing several native-language learners and Dickinson (2010) does not mention to have been inspired in the analysis of learners' corpus.

3. Materials and Methods

The starting point of our research was the analysis of the *Espanhol-Acadêmico-Br* corpus. The *Espanhol-Acadêmico-Br* is a learner corpus, which consists of introductions of academic texts written in Portuguese by Spanish native speakers enrolled in the courses of Engineering, Physics, Chemistry, Mathematics, Computer Science and Architecture from University of São Paulo (USP) in São Carlos, Brazil. The corpus contains 13 texts, with a total of 617 sentences and 17,795 words. In the *Espanhol-Acadêmico-Br* corpus we found many types of errors, some of which were detected by spelling and grammar checkers tailored for Portuguese native speakers. Other errors, however, have not been corrected by such tools, mostly because they are errors never made by Portuguese native speakers, that is, they are errors specific of foreign learners of Portuguese. After that, we decided to address each type of error separately in order to

simplify the task and improve the spelling and grammar checker gradually. In this paper we focused on errors caused by the substitution of Spanish suffixes for Portuguese suffixes, one of the errors caused by the transfer of rules of the native language to the foreign language. Aiming to identify such interlanguage errors made by Spanish learners of Portuguese and to provide corrections for them, we adopted the steps presented in Figure 1, which shows the complete procedure to generate the lexicon of errors using equivalent suffixes.

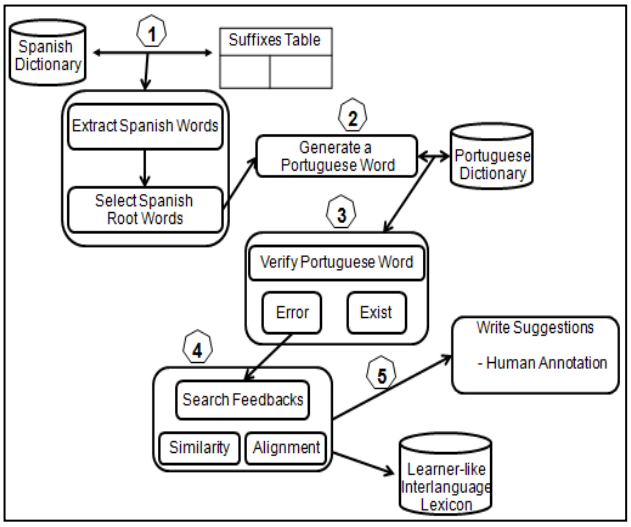

Figure 1: Procedure to create a learner-like interlanguage lexicon.

First we constructed a table of equivalent suffixes in both languages and selected those which were not cognates (Step 1). Then we used the Spanish dictionary[1] of *Universidad Carlos III de Madrid* (with 54,000 stems and their derived forms) to extract words containing such suffixes. Subsequently, we substituted the Spanish suffixes by the Portuguese ones, generating a learner-like interlanguage lexicon (Step 2).

The next step (3) was to check the existence of the generated words in a freely available Portuguese dictionary developed by NILC[2], with approximately 880.000 words (Muniz et al., 2005). Those generated words not listed in the Portuguese dictionary were labeled as "errors" and submitted to two automatic correction procedures: i) a similarity measure algorithm and ii) a translation algorithm based on aligned corpus being developed within HABLA project (Step 4).

The first correction procedure searches in the Portuguese dictionary for the most similar word to the wrong word, employing the Longest Common Subsequence Ratio (LCSR) similarity measures. We opted for LCSR measure because it is a measure largely employed to evaluate word similarity (Kondrak & Dorr 2004; Frunza & Inkpen's 2009) and in Sepúlveda-Torres & Aluísio (2011) it presented the best performance among several measures tested to identify cognates between Spanish and Portuguese. The second correction procedure searches the possible translation for the Spanish word (source of the errors) using the sentence aligned corpus *Revista Pesquisa FAPESP* (Aziz & Specia 2011). We used the statistical word aligner GIZA++ (Och & Ney 2000) to align the words of the corpus. For those words that occurred in the aligned corpus, we also provided manual translations and used these translations as a gold standard to evaluate both automatic procedures (Step 5).

4. Results

The Table 1 shows the results obtained with the artificial word generation. The first column contains the pair of Spanish-Portuguese suffixes; the second column presents the quantity of Spanish words extracted from the Spanish dictionary using each suffix as search parameter; the third column shows the total of words artificially generated which have been validated as belonging to the Portuguese lexicon and the fourth column, the total of words artificially generated that probably do not belong to the Portuguese lexicon. The later words are likely wrong words, but we can not categorically assure this since we used a lexicon to verify the existence of such words and we know any lexicon is a finite resource of a language (the fact that these words have not been attested in a dictionary does not mean they are any less correct).

Parallel Suffixes	**Generation Process**		
Spanish/ Portuguese	**Spanish Words**	**Portuguese Words**	**Likely Wrong Words**
-aje/-agem	231	62	169
-dad/-dade	956	500	456
-ción/-ção	2002	1108	894
-anza/-ança	71	22	49
-miento/-mento	1014	245	769
-tud/-tude	45	19	26
Total	**4319**	**1956**	**2363**

Table 1 - Results of artificial word generation process.

The similarity between Spanish and Portuguese languages is one of the motivations to create the method to generate the lexicon of interlanguage errors. To measure these similarities in the context of this paper we compared the Spanish and Portuguese words of the lexical resources used. For that, we compared the Spanish words (second column of Table 1) with the generated Portuguese words that belong to the Portuguese lexicon (the third column of Table 1) and the manual translations for the possible

[1] http://www.datsi.fi.upm.es/~coes/espell_leame.html
[2]http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html

wrong words. Evidence of these similarities is showed in Figure 2. As may be seen, 91.90% of the equivalent words in both languages have identical stems or their stems differ only for one letter. We also observed that in 3.91% of cases the derivation process in Spanish uses a suffix while the derivation process in Portuguese uses another one with the same function. Finally, in 4.41% of the cases, the stems of the Spanish words are different from their respective equivalent words in Portuguese in at least two letters.

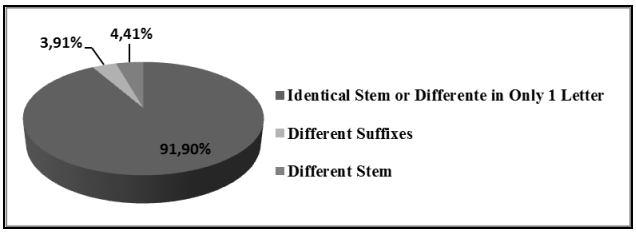

Figure 2: Comparison between Spanish and Portuguese words involved in the process of artificial generation of words for the lexicon of interlanguage errors.

As explained in the methodology, for the likely wrong words we applied two correction procedures. The first one, using similarity measures, suggested corrections for 93.86% of those words. The other, using an aligned corpus, suggested corrections only for the words that occur in the aligned corpus, which represent 28% of possible wrong words. For the likely wrong words that received correction suggestions from both procedures, we provided manual translations and used them to evaluate the suggestions. Table 2 shows the results of such evaluation. In the same table we compare the performance of the alignment-based method in two sceneries: considering all the words suggested and considering only words with more than four occurrences in the aligned corpus.

	frequency > 0	**frequency > 4**
	Precision	**Precision**
Similarity-based method	**~52%**	**~52%**
Alignment-based method	**~79%**	**~90%**

Table 2: Evaluation of the suggestions methods to correct the wrong words.

Figure 3 provides a more detailed comparison of both methods.

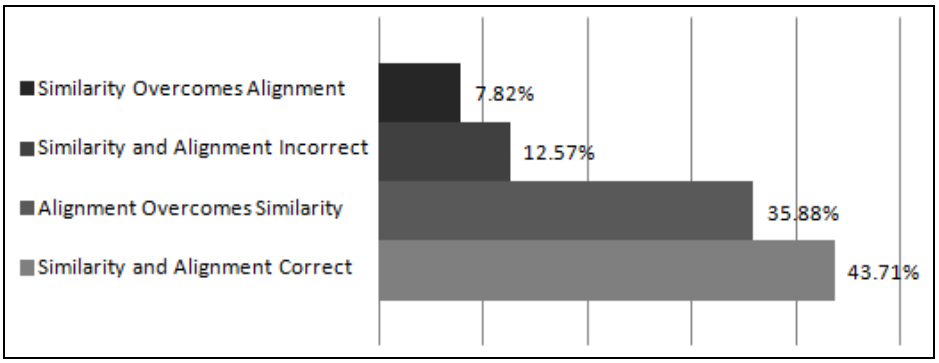

Figure 3: Comparison among methods based on similarity and on alignment.

5. Discussion

As may be inferred from the precision obtained using similarity measures (52%), the major part of likely wrong words are still very similar in both languages. This assumption confirms the result showed in Figure 2, in which 91.90% of the stems of the generated words are identical or differ only for one letter. Even in the words for which the correction using similarity measures failed, we observed similarities not captured by the method. Therefore, some recurrent changes in the stems of the words may be useful to train the algorithms of word similarity measures for the specific task of suggesting similar words in Portuguese taking Spanish words as base.

For example, the sequence "*cua*" and "*cue*" changes frequently into "*qua*" and "*que*", as in the pairs of equivalents "*cuanto-quanto*" (how), "*frecuencia-frequência*" (frequency) (the word similarity measures we used did not identify the similarity among them). Other recurrent similarities observed are the following changes from Spanish to Portuguese direction:

- S into SS: *escasez* / *escassez* (shortage); *resonancia* / *ressonância* (resonance), *esencia* / *essência* (essence);
- MN into M or N: *omnipotencia* / *onipotência* (omnipotence); *inmunodeficiencia* / *imunodeficiência* (immunodeficiency); *inminencia* / *iminência* (imminence), *somnolencia* / *sonolência* (sleepiness/somnolence);

- B into V or V into B: *absorbencia / absorvência* (absorbency); *inmovilidad / imobilidade* (immobility); *aprobación / aprovação* (approbation);
- CIA into ÇA: *sentencia / sentença* (sentence/ judgment); *diferencia / diferença* (difference), *herencia / herança* (heritage), *licencia / licença* (license);
- CT into T: *reluctancia / relutância* (reluctance); *actualidad / atualidade* ((in the) present); *electricidad / eletricidade* (electricity);
- DH into D: *adherencia / aderência* (adherence);
- UCIÓN into UIÇÃO: *contribución / contribuição* (contribution); *institución / instituição* (institution); *distribución / distribuição* (distribution);
- suppression of H: *inhibición / inibição* (inhibition); *rehabilitación / reabilitação* (rehabilitation) ; *deshidratación / desidratação* (dehydration).

Cases that will hardly be solved by the method based in similarity measures are those for which the Spanish derivation process uses a suffix and the Portuguese derivation process uses another one (3.91%). For example: "*lactancia-lactação*" (lactation); "*suciedad-sujeira*" (dirt) ; "*filmación-filmagem*" (filming).

Table 2 shows that the alignment-based method achieved a precision of 79%, considering all the words occurring in the corpus (frequency > 0), surpassing the other method in 27%. This result is even better if we consider only words with more than four occurrences in the corpus (frequency > 4), reaching 90% of precision. The alignment-based method, therefore, is strongly influenced by the frequency of the words in the corpus: as a high number of Spanish words occurs only once, in some cases the algorithm cannot identify the correct translations.

Figure 3 presents a comparison between the method based in similarity and the method based in parallel corpus alignment. This comparison shows that although the first method failed in several cases, it overcomes the method based in alignment in 7.82%. In general, this outperformance occurred because some Spanish words have low frequency in the aligned corpus. Even though the similarity-based method can not identify the correct translations whenever the word frequency is low, it overcomes the similarity-based method in 35.88%.

A shortcoming of our approach is that in 12.57% of the cases both methods failed. On the other hand, in 43.71% of the cases both methods succeeded. This is a promising result because it means we can combine both methods to improve the final translation for low frequency words in the aligned corpus.

An important feature of the methods proposed to provide correction for the generated wrong words is that both of them depend on other linguistic resources: the similarity-based method depends on a Portuguese dictionary and the alignment-based method depends on a parallel corpus. Then, to ensure an adequate performance of both approaches, it is necessary to have in the dictionary all the possible words to be suggested and the occurrence of all Spanish words (source of errors) in the parallel corpus. Approximately 72% of the total Spanish words extracted for this experiment from the Spanish dictionary do not appear in the aligned parallel corpus. This is a shortcoming of the present experiment. Actually, the lack of language resources is the main problem to create computational tools to support learning and writing activities for foreign learners of Portuguese.

6. Future Work

The lexicon generated in this research[3] will integrate the grammar and spelling checker of Portuguese designed for Spanish native speakers. We foresee also the opportunity to use this lexicon to customize the similarity measures used to suggest possible equivalents in Portuguese for the words produced by Spanish native learners. In this first investigation we focused only on suffixes that have different forms in both languages, however we intend to extend the same approach to observe the errors generated when the learner presumes that identical suffixes in both languages produce identical words and other types of errors, such as gender.

7. Acknowledgments

The authors are grateful to *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP) for supporting this research.

8. References

Aziz, W., Specia, L. (2011). Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. *In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*. STIL 2011, Cuiabá, MT, pp. 67--76.

Dagneaux, E., Denness S. and Granger, S. (1998). Computer-aided error analysis System. *System 26 (1998),* pp. 63--174.

Dahlmeier, D. and Tou Ng, H. (2011). Grammatical error correction with alternating structure optimization. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Volume 1, HLT '11, pp. 915--923, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dickinson, M. (2010). Generating learner-like morphological errors in Russian. *In Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010). Beijing, pp. 259--267.

Duran, M. S. (2008). *Parâmetros para a elaboração de dicionários bilíngues de apoio à codificação escrita em línguas estrangeiras* (PhD. Thesis in Linguistics). IBILCE, *Universidade Estadual Paulista*. Brazil. Available at:

[3]Available at HABLA by December 2014, in the ending of HABLA project: http://www.nilc.icmc.usp.br/habla/index.php

http://www.athena.biblioteca.unesp.br/exlibris/bd/brp/33004153069P5/2008/duran_ms_dr_sjrp.pdf.

Foster, J. and Oistein, A.E. (2009). GenERRate: generic errors for use in grammatical error detection. *In Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*. Boulder, Colorado, pp. 82--90.

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*, 1(1), Otawa, Canada, pp. 1--37.

Genoves, L., Lizotte, R., Schuster, E., Dayrell, C., and Aluísio, S. (2007). A two-tiered approach to detecting English article usage: an application in scientific paper writing tools. *In Recent Advances in Natural Language Processing*, Borovets.

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*, (COLING 2004), Geneva, Switzerland, pp. 95--958.

Muniz, M. C. M., Nunes, M. G. and Laporte, E. (2005). UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. *In Proceedings of the Workshop on Technology of Information and Human Language* (TIL), São Leopoldo (Brazil): Unisinos. Available at http://www.nilc.icmc.usp.br/til/til2005/arq0102.pdf.

Leacock, C., Chodorow, M., Gamon, M., and Tetreault, J. (2010). Automated grammatical error detection for language learners. Morgan and Claypool Publishers.

Och, F. and Ney, H. (2000). Improved statistical alignment models. *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440--447.

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.

Sepúlveda-Torres, L and Aluísio, S. (2011). Using machine learning methods to avoid the pitfall of cognates and false friends in Spanish-Portuguese word pairs. *In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 2011, Cuiabá/MT. Volume 1, pp. 67--76.

Sepúlveda-Torres, L. Rodrigues, R. and Aluísio, S. (2014) *Espanhol-Acadêmico-Br: A corpus of academic Portuguese learners produced by native speakers of Spanish*, In: Aluisio, S. M. and Tagnin. S. E. O. (eds.) *New languages technologies and linguistic research: a two-way road*, pp. 98--111.

Tono, Y. (2003). Learner corpora: design, development and applications. *In Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University: University Centre for Computer Corpus Research on Language, 2003, pp. 800--809.