Espanhol-Acadêmico-Br:UM CÓRPUS DE APRENDIZES DE PORTUGUÊS ACADÊMICO PRODUZIDO POR NATIVOS DE ESPANHOL

1.Introdução

O português é a quinta-língua mais falada no mundo com aproximadamente 272,9 milhões de falantes (CARVALHO et al., 2010). Como consequência do crescimento da economia brasileira e do aumento da presença de multinacionais no Brasil, o número de estrangeiros interessados em aprender o português tem aumentado na última década, sendo maioria os hispanofalantes. Esse crescimento se faz notar também pelo número de inscritos no exame CELPE-Bras (Exame de Proficiência da Língua Portuguesa) que saltou de 1,155 para 6,139 (FOREQUE, 2011).

Este cenário traz um aumento tanto na quantidade de pesquisas que buscam métodos mais eficientes de ensino e de aprendizagem do português por hispanofalantes (EVERS et al, 2011; CARVALHO et al., 2010; HENRIQUES, 2004; GRANNIER e CARVALHO, 2001), como em iniciativas para a criação de novos recursos, como o projeto *Portuguese for Spanish Speakers*, desenvolvido na Universidade do Arizona (CARVALHO et al., 2010).

Vários autores (DA SILVA, 2010; MOHR, 2007; GOMES, 2002; FURTOSO e GIMENEZ, 2000; SANTOS, 1999) acreditam que, devido à similaridade entre o português e o espanhol, os hispanofalantes possuem características diferenciadas em relação aos demais aprendizes, por isso as instituições oferecem cursos especializados a eles. Tal semelhança entre os idiomas apresenta-se tanto como um elemento positivo, quanto como um obstáculo, já que muitas vezes oculta as diferenças e impede o domínio, mantendo, na fala e na escrita em língua portuguesa, interferências da língua espanhola (GOMES, 2002; FURTOSO e GIMENEZ, 2000). Contudo, segundo SCARAMUCCI e RODRIGUES (2004), as produções realizadas por hispanofalantes também apresentam outros problemas, comuns a estrangeiros falantes de outras línguas e a próprios brasileiros.

GRANNIER e CARVALHO (2001) realizaram um levantamento dos erros cometidos por hispanofalantes ao escreverem em português. Esses erros foram encontrados em 15 provas, de gêneros diversos, do exame CELPE-Bras e classificados de acordo com os critérios linguísticos da Tabela 1.

Tabela 1: Categoria dos erros identificados por GRANNIER e CARVALHO (2001).

Erros lexicais	Relacionados com a seleção lexical, seja na sua forma ou na sua adequação semântica.

Erros morfossintáticos	Erros que afetam a estrutura interna da palavra e seus vínculos com outras palavras (Exemplo: erros de regência e de gênero).
Erros sintáticos	Relacionados com erros de ordem dos constituintes ou dos conectivos
Erros léxico-sintático-semânticos	Erros que incidem em dois ou mais itens lexicais em pelo menos uma das duas línguas e envolvem diferenças de recorte semântico e/ou uso em variadas estruturas sintáticas.
Inadequações	De registro e de estruturas (Exemplo: Uso de pronome átono ou sentenças relativas)

Por outro lado, em uma perspectiva pedagógica, DURÃO (1999) desenvolveu um estudo relacionado ao aprendizado do português por hispanofalantes, que se baseou na análise de textos escritos por 24 alunos do primeiro ano do Curso de Português oferecido na Universidade de Valhadolide, Espanha. Todos os alunos, com nível de proficiência entre básico e intermediário, eram nativos de países de língua espanhola. Na Tabela 2 observam-se as categorias mais gerais, os erros e suas respectivas porcentagens.

Tabela 2: Categoria de erros identificados por DURÃO (1999)

Categorias	Erros	Percentagem
Erros fonológicos e gráficos	595	51%
Erros léxicos	386	33,10%
Erros gramaticais	186	15,90%

Ainda que a classificação proposta por GRANNIER e CARVALHO (2001) seja mais geral que a de DURÃO (1999), ambas identificam problemas a partir de dados extraídos de textos de temas gerais, produzidos por aprendizes. No entanto, os dois estudos não tratam questões relacionadas à coesão, coerência, adequação ao gênero e estruturação dos textos, além de não formalizarem a criação do córpus, segundo princípios da Linguística de Córpus, e nem fundamentarem a tipologia de erros.

Córpus de aprendizes – coleções de textos produzidos por estrangeiros – são bastante comuns para a língua inglesa, apresentando os problemas reais que estrangeiros apresentam quando estudam o inglês (cf. por exemplo, o projeto Br-ICLE (http://www2.lael.pucsp.br/corpora/bricle/), com o subcórpus em português do projeto ICLE (http://www.uclouvain.be/en-cecl-icle.html). Por outro lado, a criação de córpus de aprendizes para o português do Brasil, como por exemplo o projeto COMAprend (TAGNIN, 2006), ainda é incipiente e este número é ainda menor se considerarmos um córpus específico de hispanofalantes. Como exemplos destes últimos, temos o de DA SILVA (2010) que criou um córpus com as produções orais de seis hispanofalantes aprendizes de português; e EVERS et al. (2011) que compilou um córpus de aprendizes de 16 textos, de gêneros diversos, produzidos em uma escola particular especializada em português como língua adicional, com um total de 8.873 palavras. Outras línguas

diferentes do inglês, como o tcheco e maltês, também estão começando a criar estes recursos (HANA et al., 2012; ROSNER et al., 2012).

O presente trabalho insere-se neste contexto e tem por objetivo formalizar uma tipologia dos principais erros de hispanofalantes matriculados em programas de pós-graduação, produzindo teses e dissertações. Neste trabalho, apresentamos a compilação inicial do córpus com textos de alunos dos programas dos Institutos de Física, Química, Ciências Matemáticas e de Computação, Escola de Engenharia e Arquitetura e Urbanismo da USP de São Carlos. Como trabalhos futuros, pretendemos estender a compilação do córpus até ele se tornar um recurso formal capaz de apresentar os problemas reais que aprendizes de português cometem quando escrevem textos acadêmicos, que requerem uma linguagem mais técnica e formal.

2. Anotação do córpus de aprendizes do português

O córpus contem 13 introduções, produzidas por diferentes alunos da pós-graduação, nos exames de qualificação e defesa, sendo composto por 617 sentenças e 17795 palavras. A Tabela 3 mostra as estatísticas do córpus, assim como as áreas de pesquisa.

Tabela 3: Estatística do córpus de escrita acadêmica por aprendizes do português

	Córpus de aprendizes		
Área de Pesquisa	**Introduções**	**Total de sentenças**	**Total de palavras**
Hidráulica e Saneamento	I1	25	827
Matemática	I2	22	699
Engenharia Civil	I3,I4,I5,I8,I9,I13	246	7599
Ciência da Computação	I6,I7,I10,I11,I12	324	8670
Total	**13**	**617**	**17795**

Após a compilação do córpus, foi realizado um questionário para se conhecer as principais dificuldades que os hispanofalantes, engajados em programas de pós-graduação na USP São Carlos, apresentam quando escrevem textos acadêmicos em português. O questionário, composto por 17 perguntas relacionadas aos problemas identificados por GRANNIER e CARVALHO (2001) e DURÃO (1999) e com aspectos associados à proficiência da língua, foi respondido por 66 pessoas. A Figura 1 mostra os erros incluídos no questionário, assim como seu nível de dificuldade, segundo os aprendizes de português.

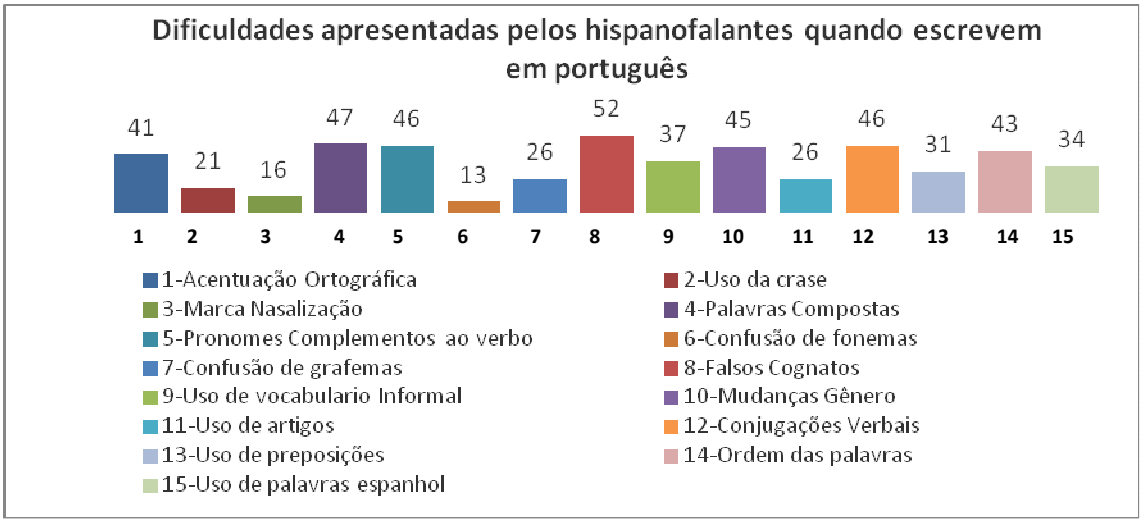

Figura 1: Quantificação dos erros apresentados no questionário.

Todos os problemas da Figura 1 foram apontados como dificuldades na escrita. Os mais preocupantes foram: falsos cognatos, palavras compostas, pronomes complemento, conjugações verbais, mudanças de gênero, ordem das palavras nas sentenças, acentuação ortográfica, uso de vocabulário informal, uso de palavras em espanhol nos textos e uso de preposições. Os restantes podem ser tratados pelos corretores ortográficos mais utilizados em processadores de textos.

Após quantificar as perguntas do questionário, o córpus de introduções foi anotado com base nos erros considerados no questionário e nas anotações realizadas em outros estudos. Dois anotadores analisaram cada erro separadamente e, conforme novos tipos de equívocos foram sendo identificados, novas categorias foram consideradas e caracterizadas por um número, incluindo o tipo de erro e uma sugestão para erradicá-lo. Alguns erros podem ser incluídos em várias categorias, o que representou um problema na anotação, mas optou-se por classificá-los sempre na categoria considerada mais "grave" ou relevante para o aprendizado. Outra possibilidade de anotação seria utilizar multirrótulos e pode ser considerada futuramente.

No começo da anotação, esperava-se que o córpus apresentasse poucos erros, pois as redações foram escritas de forma digital, utilizando corretores ortográficos e/ou gramaticais, e produzidas para exames de qualificação/defesa de mestrado/doutorado, que exigem níveis avançados de proficiência da língua portuguesa. No entanto, depois da anotação, identificaram-se erros que corretores ortográficos e/ou gramaticais tratam e outros que ainda não foram atendidos; em sua maioria porque são erros específicos de aprendizes de uma língua.

Um levantamento desenvolvido por DURAN (2008) comprovou que os corretores mais difundidos não foram construídos para identificar erros produzidos por aprendizes. Seguindo esta linha, este estudo visa: (i) apontar categorias de erros que não são cobertos por ferramentas computacionais; (ii) identificar lacunas ou deficiências destes sistemas e (iii) selecionar os problemas que podem ser automatizados, tendo como base o córpus de aprendizes.

A Tabela 4 mostra a tipologia de erros construída com base no córpus de aprendizes. Cada categoria apresenta um exemplo retirado do córpus para ilustrar o problema tratado; o número da categoria que apresenta o erro e a sua correção aparecem em negrito. A Figura 2 ilustra um gráfico com as ocorrências de cada categoria no córpus.

Tabela 4: Tipologia de erros construída após a análise do conjunto inicial de textos.

1- Erro de Ortografia (Acentuação ortográfica equivocada; Marca equivoca de nasalização; Confusão de fonemas; Confusão de grafemas para o mesmo fonema) Ex: ... tem a tarefa de analisar um **paragrafo (parágrafo/1)** ou uma sentença de texto ...
2- Ausência ou uso inadequado da crase Ex: ... um processo altamente complexo quanto ao aspecto microbiológico e extremamente sensível **as (às/2)** condições ambientais...
3- Uso do hífen (separação e união de palavras) Ex: ... para tornar o Brasil **auto-suficiente (autossuficiente/3)** na produção de petróleo e gás natural.
4- Colocação pronominal (Posição dos pronomes oblíquos) Ex: **Podese (pode-se/4)** perceber a grande quantidade de interações ...
5- Uso de falsos cognatos Ex: **-** Também, métodos de **patrões (padrões/5)** foram empregados através do uso de templates ...
6 -Uso de um vocabulário informal para o gênero científico Ex: ... essas referências são para mencionar **uns poucos (alguns/6)** artigos em diferentes contextos.
7- Mudanças de gênero Ex: Denomina-se estocástica devido à entrada ao sistema não ser conhecida e, portanto, suposta como um processo de **origem aleatório (origem aleatória/7)**.
8- Uso incorreto de artigos Ex: Em particular, quanto mais alta é _ **(a/8)** taxa de compressão, ...
9- Conjugações verbais Ex: O Desenvolvimento Sustentável é definido como um modelo econômico, político, social, cultural e ambiental equilibrado, que **satisfaze (satisfaz/9)** as necessidades das gerações atuais. - Contudo, a extensão destes recursos não é tão significativa como se **desejasse (deseja/9)**.
10- Uso incorreto de preposições Ex: - O recurso está sendo desenvolvido **na base de (com base em/10)** um pequeno córpus, em comparação ao córpus original.
11- Problemas de ordem das palavras e expressões Ex: A identificação modal operacional **com só (só com/11)** medições das saídas é um atrativo
12- Uso de palavras ou expressões em espanhol e do espanhol no texto Ex: - Neste aspecto ainda não tem sido muito abordada a validez dessa **assunção (afirmação/12)** ... - ... não só usar características dos dados rotulados, **se não que também (mas que também/12)** aproveite atributos de dados não rotulados.
13- Concordância nominal **Ex: -** ... conjuntos de técnicas de Mineração de Dados e Textos têm sido **usadas (usados/13)** para auxiliar ...
14- Categorias difusas ((i) Sentenças longas com falta de conectivos e de sinais de pontuação ; (ii) Repetição de palavras (iii) Usos menos frequentes na língua. Ex: - **Umas das áreas (Uma das áreas /14)** de maior pesquisa no campo da análise modal, ... - Felizmente, verifica-se que gradualmente se vai fomentando a procura deste tipo de materiais, e maior é a tendência dos pesquisadores de estimular a busca de novas matérias-primas que sejam provenientes de fontes renováveis, menos poluentes, e locais, seja porque está a surgir uma mudança de mentalidade da sociedade, seja por uma questão de moda ou mesmo pela simples necessidade de mudança **(Sugestão: Uso de conetivos, Sentença longa, criar várias sentenças/14)**.
15- Mudança da função da palavra por desconhecimento de sua ortografia Ex: - A digestão anaeróbia sofre grande **influencia (influência/15)** do regime hidráulico. - As técnicas de **analise (análise/15)** de DNA tais como ....

Além do erro apontado no exemplo da categoria 12, está errada a palavra validez, que poderia ser corrigida com o uso de validade ou veracidade, mas foi anotado unicamente o erro relacionado com a categoria atual.

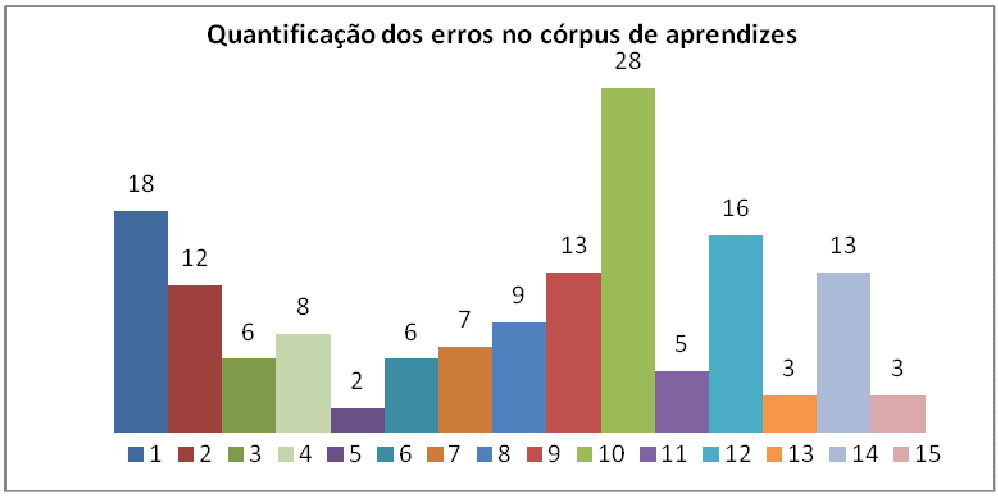

Figura 2: Quantificação dos erros presentes no córpus de aprendizes

Como observado na Figura 2, os erros mais recorrentes nem sempre correspondem aos indicados no questionário, como o uso de falsos cognatos, que foi o mais inquietante na enquete e o de menor ocorrência no córpus. Acredita-se que esse problema é mais comum na linguagem falada, enquanto na escrita não é muito frequente. SEPÚLVEDA-TORRES, L e ALUISIO, S.M. (2011), apresentam um estudo inicial para a criar de forma automática dicionários de cognatos e falsos cognatos das línguas português e espanhol. Esse tipo de recursos são muito úteis na aprendizagem de línguas próximas.

Os erros das categorias difusas mostram que os aprendizes apresentam problemas para compor ideias claras e concisas, construindo sentenças longas, com ausência de conectivos, de sinais de pontuação e com palavras repetidas. Algumas conjugações verbais não podem ser corrigidas por corretores ortográficos, como o exemplo na categoria *9*.

Expressões em espanhol, no entanto, aparecem algumas vezes e podem ser identificadas automaticamente. Além disso, o uso da crase é um dos problemas de maior frequência e existem regras da língua portuguesa que permitem sistematizá-lo.

A Figura 3 apresenta, por meio de gráficos, erros identificados em três introduções do córpus nas quais verifica-se problemas que são comuns entre os aprendizes e outros que são mais específicos de um aprendiz. Observa-se que o uso inadequado da crase (2) é comum nos três textos e nos textos 11 e 9 se repetem erros relacionados a conjugações verbais (9), uso incorreto de preposições (10) e uso de expressões em espanhol (12). A Figura 3.1 mostra uma introdução com um número

menor de erros, a Figura 3.2, uma com um número maior de erros que a primeira e a Figura 3.3 uma com o maior número de equívocos, totalizando 26 erros. Por meio dessa análise pode-se obter uma avaliação dos textos em função dos erros cometidos, ponderando-os a partir de sua gravidade e estabelecendo uma rubrica/métrica de avaliação.

Figura 3: Tipos de erros identificados em introduções individuais

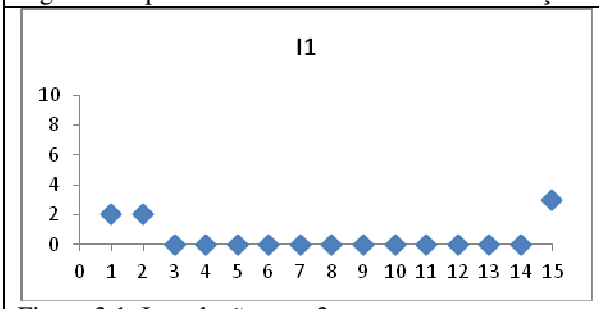

Figura 3.1: Introdução com 3 erros

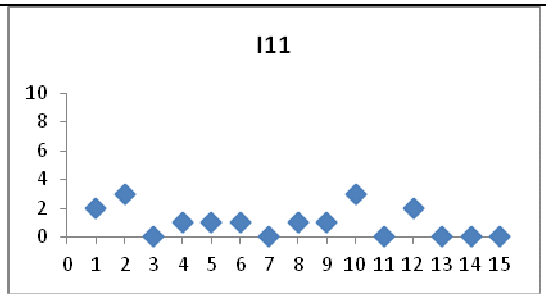

Figura 3.2: Introdução com 15 erros

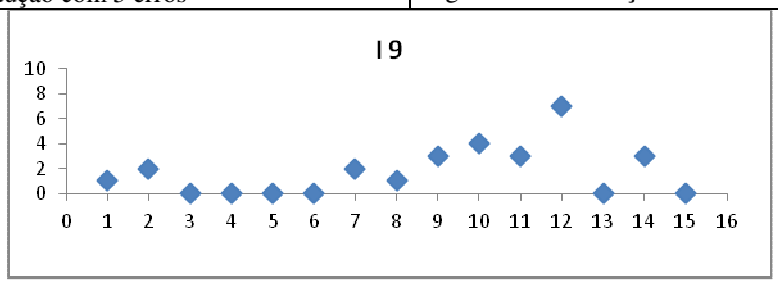

Figura 3.3: Introdução com 26 erros

3. Conclusões

Embora os hispanofalantes utilizem ferramentas que auxiliem a escrita de textos em português, existem alguns problemas que estas ferramentas não conseguem identificar. Desta forma, o presente estudo constitui uma base para a consolidação de um córpus que ilustre os problemas reais que nativos do espanhol apresentam quando escrevem em português. A análise destes problemas servirá para criar uma ferramenta automática que identifique estes erros, ofereça sugestões para melhorar a qualidade do texto e avalie a escrita do texto em função da ponderação dos erros. Além desta ferramenta, o córpus de aprendizes, que será acrescido de mais textos, será disponibilizado publicamente para pesquisas futuras.

Referências

CARVALHO, A., J.L, F., e DA SILVA, A. (2010). *Teaching portuguese to spanish speakers a case for trilingualism*. In Hispania, páginas 70–75.

DA SILVA, L. (2010). As formas de preenchimento do objeto direto na aprendizagem de português/LE por Argentinos. Universidade Federal de São Carlos. Centro de Educação e Ciências Humanas-Departamento de Letras. Cursos de Letras.

DURÃO, A. (1999). Análisis de errores e interlengua de brasileños aprendices de español y españoles aprendices de portugués. Editora UEL. Universidade Estadual de Londrina.

DURAN, M. (2008). Customização de corretores ortográficos para aprendizes de línguas estrangeiras. In Anais do VII Encontro de Linguística de Córpus.

EVERS, A., FINATTO, M., e PASQUALINI, B. (2011). Córpus de aprendizes de português como língua adicional (pla): Compilação inicial e primeiros resultados. In 18 InPLA - Intercâmbio de Pesquisas em Linguística Aplicada.

FOREQUE, F. (2011). Crescimento do brasil leva estrangeiros a aprenderem português. In Folha.com, 2011.

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.

FURTOSO, V. e GIMENEZ, T. (2000). Ensino e pesquisa em português para estrangeiros programa de ensino e pesquisa em português para falantes de outras línguas (peppfol). DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, 16:443 – 447.

GOMES, G. (2002). Características da interlíngua oral de estudantes de letras espanhol nos dois últimos semestres de estudo. In Congresso Brasileiro de Hispanistas, São Paulo, Brasil.

GRANNIER, D. e CARVALHO, A. (2001). Pontos críticos no ensino de português a falantes de espanhol - da observação do erro ao material didático. In Anais do IV Congresso da SIPLE, PUC Rio, Rio de Janeiro, Brasil.

HENRIQUES, R. (2004). Intercompreensão de Texto Escrito por Falantes Nativos de Português e de Espanhol. DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, 16:263–295.

HANA, J. e ROSEN, A. e STINDLOVA, P.J. (2012). *Building a learner córpus. Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12), p. 3228-3232. Istanbul, Turkey.

MOHR, D. (2007). Português para hispanofalantes: Uma alternativa para o ensino de gêneros escritos. In Professores de Línguas Estrangeiras do Paraná Línguas: culturas, diversidade, integração (XV EPLE), p. 372–387.

ROSNER, M. e GATT, A. e ATTARD, A. e JOACHIMSEN, J, (2012). *Incorporating an Error Córpus into a Spellchecker for Maltese. Proceedings of the Eight International*

Conference on Language Resources and Evaluation (LREC'12), p.743-750. Istanbul, Turkey.

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.

SCARAMUCCI, M. e RODRIGUES, M. (2004). Compreensão (oral e escrita) e produção escrita no exame celpe-bras: análise do desempenho de candidatos hispanofalantes. In In: SIMÕES, A. R. M. et al. (Org.Ed.). Português para falantes de espanhol: artigos selecionados escritos em português e inglês, 2004.

SEPÚLVEDA-TORRES, L e ALUISIO, S. M. (2011) *Using machine learning methods to avoid the pitfall of cognates and false friends in Spanish-Portuguese word pairs*. *In: The 8th Brazilian Symposium in Information and Human Language Technology*, 2011, Cuiabá/MT. v. 1. p. 67-76.

TAGNIN, E. (2006). *A multilingual learner córpus in brazil. Language and Computers*, 56(1):195–202.