Escrita científica em português por hispano
falantes: recursos linguísticos-computacionais
baseados em métodos de alinhamento de textos
paralelos

Lianet Sepúlveda Torres

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Lianet Sepúlveda Torres

**Escrita científica em português por hispano falantes:
recursos linguísticos-computacionais baseados em
métodos de alinhamento de textos paralelos**

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Ciências - Ciências de Computação e Matemática Computacional. VERSÃO REVISADA

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Sandra Maria Aluísio

**USP – São Carlos
Novembro de 2015**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

- S693e Sepúlveda Torres, Lianet
Escrita científica em português por hispano
falantes: recursos linguístico-computacionais
baseados em métodos de alinhamento de textos
paralelos / Lianet Sepúlveda Torres; orientadora
Sandra Maria Aluísio. -- São Carlos, 2015.
313 p.
- Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2015.
1. Córpus de aprendizes. 2. Léxicos bilíngues. 3.
Português como segunda língua. 4. Ferramentas de
suporte à escrita. I. Aluísio, Sandra Maria, orient.
II. Título.

Lianet Sepúlveda Torres

**Scientific writing in portuguese by hispanic speaking:
linguistics-computational resources based on alignments
methods of parallel text**

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Sandra Maria Aluísio

**USP – São Carlos
November 2015**

*A Lian por ser meu amor;
Aos meus pais pelo exemplo e a confiança.*

Agradecimentos

A Lian (Mi jevi) porque é meu amor, por ser minha força, pela paciência, pelo apoio, pelos momentos difíceis e pelas muitas alegrias.

Aos meus pais por manter uma família tão unida ainda na distância. Por me incentivar, apoiar e confiar em mim. A meu irmão que adoro, por sua generosidade e pelos anos não compartilhados.

À família que ganhei aqui no Brasil (Esperanza, Felipe, Orieta, Indara, Al-lita, Patricia e Luiz) pela constante preocupação, o apoio incondicional e por me escutar sempre. Adoro vocês!!!

À Profa. Sandra Maria Aluísio pelo apoio, a confiança e dedicação nestes anos de pesquisa.

Às eternas parceiras desta pesquisa Dra. Magali Sanches Duran, Roana Rodrigues e MSc. Aline Evers, sempre entusiastas e dispostas a colaborar.

Aos colegas do NILC que nestes anos me ajudaram profissionalmente, em especial a Lucia, Paula e Marcio.

À FAPESP pelo apoio financeiro, ao NILC e à USP pelas instalações.

Ao pessoal da secretaria, aos técnicos, por arrumar meus problemas urgentes na hora, às faxineiras, aos porteiros, a todos ..., meus agradecimentos por fazer que nossos dias sejam mais amenos.

À minha família de Cuba, pelo carinho e por estarem sempre na torcida. Em especial, a Elizabeth que sempre me apóia e me escuta. À família do Lian muito obrigada pela preocupação, em especial a Fina que sempre tem uma solução alternativa.

Um agradecimento especial para meus amigos, esses amigos de anos que estão sempre no meu coração e me incentivam cada dia.

Enfim, a cada pessoa que encontrei nestes anos, me apoiando, me criticando, me incentivando ...

Muito Obrigada.

*"Llegará un momento en que creas que todo ha terminado.
Ese será el principio".*
— *Epicuro de Samos (341 a.C. - 270 a.C.), filósofo griego*

Resumo

O número de estrangeiros interessados em aprender o português tem aumentado na última década, em consequência do crescimento da economia brasileira e do aumento da presença de multinacionais no Brasil. Esse fato se mostra pelo aumento do número de inscritos no exame de proficiência de português CELPE-Bras e de estudantes estrangeiros que ingressam nas universidades brasileiras. A maioria destes estudantes são de língua espanhola e precisam escrever seus textos acadêmicos em português. A proximidade das línguas portuguesa e espanhola apresenta-se tanto como um elemento positivo quanto como um obstáculo, pois oculta as diferenças e impede o domínio da língua portuguesa, mantendo, na fala e na escrita em português, interferências do espanhol. O maior número destas interferências acontece no nível lexical. Uma das alternativas para tratar os problemas em textos de aprendizes de uma língua é o emprego de ferramentas computacionais de pós-processamento e de suporte ao processo de escrita. No entanto, o número de recursos e ferramentas disponíveis para auxiliar a escrita de português como língua estrangeira é muito reduzido, diferentemente do cenário para a língua inglesa. Esta pesquisa propôs a criação de recursos e ferramentas de suporte à escrita no nível lexical como primeiro passo para a melhoria da qualidade linguística dos textos em português produzidos pelos nativos do espanhol. A Linguística de Córpus foi utilizada como metodologia para viabilizar a análise de erros de aprendizes. As ferramentas de auxílio utilizam léxicos bilíngues compilados por meio de técnicas de tradução, baseadas em alinhamento de córpus paralelos. Dado o número insuficiente de erros previamente anotados para suportar a detecção automática de erros, esta pesquisa propôs métodos baseados em modelo língua e na geração artificial de erros. A geração de erros artificiais se apresentou como um método eficiente para predizer erros lexicais dos aprendizes. As contribuições obtidas com a metodologia baseada em tradução automática para gerar auxílios à escrita entre línguas próximas, considerando a análise de erros lexicais extraídos de córpus de aprendizes, foco desta pesquisa, são: (i) do ponto de vista teórico, o levantamento e quantificação dos principais problemas causados pelas marcas do espanhol, deixadas nos textos acadêmicos em português escritos por nativos do espanhol; (ii) do ponto de vista de geração automática de recursos linguísticos, léxicos bilíngues de cognatos e falsos cognatos; léxico bilíngue de marcadores discursivos;

léxico de expressões formulaicas que aparecem nos textos científicos e léxico bilíngue de verbos relacionados com pesquisa científica em português e, (iii) do ponto de vista da criação de subsídios para a área de auxílio à escrita científica, o projeto e avaliação de auxílios para suportar a escrita científica em português por nativos do espanhol.

Palavras-chave: Córpus de aprendizes, Léxicos bilíngues, Português como segunda língua, Ferramentas de suporte à escrita.

Abstract

In the last decade, as a result of Brazilian economic growth and the increased presence of multinationals in the country, the interest of foreigners in learning Portuguese rose. This fact is also noted by the number of students enrolled in the Portuguese proficiency exam, CELPE-Bras and the number of foreigner students entering at the Brazilian Universities. Most of these students are Spanish speakers and need to write the dissertation or thesis in Portuguese. The similarity between Portuguese and Spanish 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. The largest number of this interference occurs at the lexical level. One alternative to deal with errors of second language learners is the use of computational post-edit tools and tools to support the writing process. However, the number of resources and tools available to help improve Portuguese writing as a foreign language is very small, unlike the scenario into English. This research proposed the creation of resources and writing support tools at the lexical level as a first step to improving the linguistic quality of the texts produced by Portuguese native Spanish. Corpus linguistics was used as a methodology to enable the learner's error analysis. The writing support tools use bilingual lexicons compiled through translation techniques based on alignment parallel corpus. Given the insufficient number of errors previously annotated to support automatic error detection, this research proposed methods based on language model and artificial generation of errors. The generation of artificial errors introduced himself as an efficient method for predicting lexical errors of learners. The contributions obtained with the methodology based on automatic translation to generate written supports between similar languages, considering lexical error, extracted by the analysis of learner's corpus, focus of this research are: (i) the theoretical point of view, the survey and quantification of the main problems caused by the Spanish marks, left in the academic texts written in Portuguese by native Spanish; (ii) the automatic generation of language resources point of view, bilingual lexicons of cognates and false cognates; bilingual lexicon of discourse markers; bilingual lexicon of formulaic expressions that appear in scientific texts and bilingual lexicon of verbs related to scientific research in Portuguese and, (iii)

the point of view of creating subsidies for the area of support scientific writing, design and evaluation of aid to support scientific writing in Portuguese by native Spanish speakers.

Key-words: Learner corpora, Bilingual lexicon, Portuguese as a second language, Writing support tools.

Lista de Figuras

1.1	Auxílios oferecidos pelos léxicos bilíngues	12
3.1	Alinhamento sentencial dos pares de resumos em português e espanhol, selecionados da Revista Latino Americana de Enfermagem.	48
3.2	Alinhamento lexical de pares de resumos em português e espanhol, selecionados da Revista Latino Americana de Enfermagem.	49
3.3	Modelo do canal ruidoso na tradução automática.	59
3.4	Exemplo de alinhamentos entre palavras e frases usados no treinamento do modelo de tradução.	60
3.5	Exemplo de agrupamento dos marcadores discursivos com a mesma função	88
5.1	Base de conhecimento do projeto HABLA	124
5.2	Análises para caracterizar a produção escrita dos hispano falantes	133
5.3	Layout de textos de nativos e aprendizes	138
5.4	Classificação manual de um texto.	139
5.5	Estrutura, Modelo de Classificação e <i>Class Matching</i> do conjunto de treinamento	140
5.6	Estrutura, atualização do modelo de classificação e <i>class matching</i> do conjunto de avaliação	141
5.7	Quantificação das respostas apresentadas no questionário	144
5.8	Quantificação dos erros dos aprendizes no cópua de Introduções	146
5.9	Tipologia de erros	147
5.10	Anotação de um texto com a ferramenta Brat	152
5.11	Anotação manual de trechos diferentes de textos	153
5.12	Anotação manual de trechos diferentes de textos	155
5.13	Total de erros por documento nos textos dos nativos e dos aprendizes	156
5.14	Erros cometidos pelos nativos e etiquetados na anotação manual	157
5.15	Erros cometidos pelos nativos e etiquetados na anotação manual	158
5.16	Erros cometidos pelos aprendizes e etiquetados na anotação manual	160
5.17	Erros cometidos pelos aprendizes e etiquetados na anotação manual	161

5.18 Erros cometidos pelos aprendizes e etiquetados na anotação manual	162
5.19 Nativos e Aprendizes: Distribuição da posição em que são usados os marcadores	164
5.20 Nativos e Aprendizes: Frequência de uso de marcadores por nativos e aprendizes	165
5.21 Erros cometidos pelos aprendizes e etiquetados no processo de anotação	166
5.22 Erros comuns e não comuns entre os nativos e os aprendizes nos textos anotados	166
6.1 Auxílios oferecidos pelos léxicos bilíngues	175
6.2 Estrutura da saída do procedimento para obter sinônimos entre itens lexicais.	187
6.3 Similaridade entre substantivos do espanhol e do português.	193
6.4 Procedimento para gerar o léxico bilíngue de cognatos e falsos cognatos.	195
6.5 Resultados da técnica de seleção de atributos para cada experimento	198
6.6 Resultados da técnica de seleção de atributos para cada experimento	200
6.7 Procedimento para gerar o léxico bilíngue de verbos.	203
6.8 Subtração da Tf-idf nos corpúsculos acadêmico e geral para os verbos em português.	209
6.9 Procedimento para gerar o léxico bilíngue de marcadores discursivos.	213
6.10 Procedimento para gerar erros artificiais de marcadores discursivos.	220
6.11 Procedimento para detectar e corrigir erros de marcadores discursivos por similaridade entre expressões.	223
6.12 Total de erros detectados e corrigidos pelos métodos propostos para detectar e corrigir erros de marcadores discursivos.	225
6.13 Procedimento para verificar manualmente a saída do sistema.	226
6.14 Procedimento para gerar o léxico bilíngue de expressões formulaicas.	231
6.15 Visão geral dos léxicos bilíngues incluídos em uma ferramenta de auxílio à escrita.	238
6.16 Tela com equivalências em português do marcador discursivo em espanhol "por último".	238
6.17 Tela com sinônimos do verbo "outorgar".	239
6.18 Tela com a detecção e correção do erro em que o item "em concreto" foi utilizado como marcador discursivo.	240
6.19 Tela com uma alerta que mostra a ocorrência de um falso cognato entre o espanhol e o português no texto.	240

Lista de Tabelas

1.1	Quantidade de estudantes e a porcentagem de estrangeiros, cuja língua nativa é o espanhol, nos programas de pós-graduação da USP de São Carlos.	2
2.1	Interpretação dos valores da medida <i>Kappa</i> (Landis e Koch, 1977)	24
2.2	Resumo dos corpú de aprendizes e suas características	27
2.3	Categorização dos erros que hispano falantes cometem quando escrevem em português (Grannier e Carvalho, 2001).	35
2.4	Continuação da categorização dos erros que hispano falantes cometem quando escrevem em português (Grannier e Carvalho, 2001).	36
2.5	Categorização dos erros que hispano falantes cometem quando escrevem em português (Durão, 1999).	37
2.6	Erros fonológicos e gráficos que hispano falantes cometem quando escrevem em português (Durão, 1999).	38
2.7	Erros lexicais que hispano falantes cometem quando escrevem em português (Durão, 1999).	40
2.8	Erros gramaticais que hispano falantes cometem quando escrevem em português (Durão, 1999).	40
4.1	Resumo dos trabalhos apresentados na Seção 4.2	113
4.2	Avaliação de sistemas de detecção e correção de erros gramaticais	118
4.3	Verificação das categorias de erros proposta por Gamon et al. (2008)	120
5.1	Revistas que disponibilizam textos ou resumos em português e espanhol	127
5.2	Estatísticas do corpú paralelo compilado no projeto HABLA . . .	127
5.3	Estatísticas do corpú de referência do projeto HABLA	129
5.4	Estatísticas do corpú Espanhol-Acadêmico-Br	131
5.5	Rubrica de avaliação da qualidade da escrita	136
5.6	Resultados numéricos da classificação do conjunto de treinamento e do conjunto de avaliação.	142

5.7 Resultados numéricos da classificação do conjunto de treinamento e do conjunto de avaliação com duas classes: Avançado e Iniciante.	143
5.8 Estatística do corpus de escrita acadêmica por aprendizes do português	145
5.9 Tipologia de Erros para anotar manualmente o corpus Espanhol-Acadêmico-Br	149
5.10 Tipologia de Erros para anotar manualmente o corpus Espanhol-Acadêmico-Br	150
5.11 Estatísticas do corpus anotado manualmente	151
5.12 Erros cometidos pelos nativos e etiquetados na anotação manual	157
5.13 Erros cometidos pelos aprendizes e etiquetados na anotação manual	160
6.1 Avaliação dos métodos de equivalências entre palavras	181
6.2 Avaliação da combinação dos métodos de equivalências entre palavras	182
6.3 Avaliação dos métodos de equivalências entre expressões formulaicas	184
6.4 Exemplo de erros artificiais da categoria Transferência de Sufixos	192
6.5 Amostra da geração de erros de sufixos	192
6.6 Conjunto de treinamento para separar cognatos de falsos cognatos (o número entre parêntesis indica o total de palavras com ortografia idêntica)	196
6.7 Desempenho dos classificadores testados para os experimentos propostos	199
6.8 Número de palavras no dicionário (número entre parênteses representa o número de palavras do conjunto de treinamento que não estavam no dicionário).	200
6.9 Desempenho dos classificadores testados no Experimento 3.2 . .	201
6.10 Ordem dos sinônimos do verbo "fixar"	210
6.11 Precisão na detecção e correção de erros com o léxico de erros artificiais.	228
6.12 Precisão na detecção e correção de erros com o método baseado no modelo de língua e na similaridade ortográfica.	228
6.13 Exemplo de entradas incluídas no léxico de expressões formulaicas	236

Artigos da Tese

- Torres, L. S. e Aluísio, 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, 1. páginas 67-76.
- Torres, L. S.; Rodrigues, R.; Aluísio, S. M. e VALE, O. (2012). *Espanhol-Acadêmico-Br: Um Córpus de Aprendizes de Português Acadêmico Produzido por Nativos de Espanhol*. In: XI Encontro de Linguística de Corpus (ELC 2012), São Carlos/SP, 1. páginas 1-9.
- Torres, L. S. ; Paiva, J. G. S.; Minghim, R. e Aluísio, S. M. (2013). *Visual Classification of Documents Based on Writing Quality*. In: X Encontro Nacional de Inteligência Artificial e Computacional, Fortaleza/Ceará, 1. páginas 1-12.
- Torres, L. S.; Rodrigues, R. e Aluisio, S. M. (2014). *Espanhol-Acadêmico-Br: A Corpus of Academic Portuguese Learners Produced by Native Speakers of Spanish*. In: Aluisio, S. M.; Tagnin, S. E. O.. (Org.). *New Languages Technologies and Linguistic Research: a Two-Way Road*. 1ed. Cambridge, England: Cambridge Scholars Publishing, 2014, 1, páginas 98-111.
- Torres, L. S.; Duran, M. S. e Aluisio, S. M. (2014). *Generating a Lexicon of Errors in Portuguese to Support an Error Identification System for Spanish Native Learners*. In: The 9th International Conference on Language Resources and Evaluation, 2014, Reykjavik (Iceland), 1. páginas 3952-3957.
- Torres, L. S.; Duran, M. S. e Aluísio, S. M. (2014). *O Tratamento de Marcadores Discursivos em uma Ferramenta de Apoio à Escrita Acadêmica em Português Para Nativos de Espanhol*. In: *Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish*, São Carlos, 1, páginas 1-5.
- Torres, L.S.; Duran, M. S. e Aluísio, S.M. (2015). *Automatic Detection and Correction of Discourse Marker Errors Made by Spanish Native Speakers in Portuguese Academic Writing*. Submetido a *Language Resources and Evaluation*.

Sumário

1	Introdução	1
1.1	Motivação: O Problema Tratado	1
1.2	Contexto Teórico	5
1.3	Objetivos e Contribuições	11
1.4	Organização do Trabalho	16
2	Cópus de Aprendizes	17
2.1	Cópus de Aprendizes: Definição e Trabalhos da Literatura	17
2.1.1	Características dos Cópus de Aprendizes	26
2.2	Cópus de Hispano Falantes Aprendizes de Português	30
2.3	Tipologia de Erros para as Línguas Próximas Português e Espanhol	31
2.4	Considerações Finais	42
3	Cópus Paralelos e Geração Automática de Léxicos Bilíngues	45
3.1	Métodos Automáticos de Alinhamento e de Tradução	46
3.1.1	Métodos de Alinhamento Sentencial	49
3.1.2	Métodos de Alinhamento Lexical	51
3.1.3	Tradução Automática	57
3.1.4	Modelos de Língua	62
3.2	Métodos para Gerar Léxicos Bilíngues	67
3.2.1	Métodos para Criar Léxicos Bilíngues compostos por Multipalavras	69
3.2.2	Avaliação da Geração de Léxicos Bilíngues	75
3.3	Análises Linguístico-computacionais para Gerar os Léxicos de Interesse	77
3.3.1	Cognatos e Falsos Cognatos	77
3.3.2	Verbos	84
3.3.3	Marcadores Discursivos	86
3.3.4	Expressões Formulaicas	89
3.4	Considerações Finais	95
4	Ferramentas de Apoio à Escrita de Textos	97
4.1	Ferramentas e Métodos para Auxiliar a Escrita de Aprendizes . . .	97
4.2	Deteção e Correção de Erros Gramaticais	103

4.2.1	O Revisor Gramatical ReGra	114
4.2.2	Avaliação da Tarefa de Detecção de Erros Gramaticais . . .	118
4.3	Considerações Finais	121
5	Projeto HABLA: Compilação dos Córpus e Análises	123
5.1	Córpus de Referência	124
5.1.1	Córpus Paralelo espanhol-português	124
5.1.2	Córpus do Gênero Acadêmico	128
5.2	Córpus de Aprendizes Espanhol-Acadêmico-Br	130
5.3	Hispano Falantes Aprendizes de Português Acadêmico: Diversas Análises	132
5.3.1	Avaliação Holística da Qualidade da Escrita	134
5.3.2	Análise Inicial de Erros de Hispano Falantes	143
5.3.3	Anotação de Erros de Hispano Falantes e Nativos	148
5.4	Considerações Finais	167
6	Projeto HABLA: Resultados e Avaliação dos Métodos de Auxílio	173
6.1	Métodos de Auxílios à Escrita	173
6.1.1	Equivalências entre o Espanhol e o Português	175
6.1.2	Escolhas Lexicais em Português	185
6.1.3	Detecção e Correção de Erros	188
6.2	Geração de Léxicos Bílingues para Apoiar a Escrita dos Hispano Falantes	193
6.2.1	Cognatos e Falsos Cognatos	193
6.2.2	Verbos	202
6.2.3	Marcadores Discursivos	212
6.2.4	Expressões Formulaicas	229
6.3	Ferramenta de Auxílio à Escrita HABLA	237
6.4	Considerações Finais	241
7	Conclusões	245
7.1	Contribuições	246
7.2	Limitações	250
7.3	Trabalhos Futuros	251
	Referências Bibliográficas	255
A	Medidas de Associação	283
B	Questionário Hispano Falantes	285
C	Manual de Anotação	289
D	Sufixos	303
E	Verbos	307
F	Marcadores Discursivos	311

Introdução

1.1 Motivação: O Problema Tratado

O português é a sexta língua mais falada no mundo, com aproximadamente 272,9 milhões de falantes¹ (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 a maioria hispano falantes. Esse fato se faz notar também pelo número de inscritos no exame CELPE-Bras² (Exame de Proficiência da Língua Portuguesa). A primeira aplicação oficial do exame, em 1998, contou com 127 examinandos; em 2013 e 2014, somando as duas edições anuais, o total de examinandos ultrapassou 9 mil (em 2013 foram 9.213 e, em 2014, 9.136)³.

O aumento do número de estudantes estrangeiros que ingressam nas universidades brasileiras também incrementa o interesse pela aprendizagem do português. Realizamos uma análise da quantidade de estudantes e a porcentagem de estrangeiros, cuja língua nativa é o espanhol, que ingressaram nos programas de pós-graduação da USP de São Carlos⁴, de 2007 a 2010. Na Ta-

¹<http://pt.wikipedia.org/wiki/Língua-portuguesa>

²Este exame avalia a proficiência em língua portuguesa quanto à capacidade de agir no mundo em situações similares às reais. Mede o uso da língua considerando o conhecimento gramatical ou específico de vocabulários, principalmente, as formas adequadas de uso da língua em um ambiente comunicativo (Schoffen, 2009).

³<http://www.ufrgs.br/acervocelpebras/estatisticas/numero-de-examinandos/view>.

⁴<http://www.saocarlos.usp.br/>

bela 1.1 são apresentados esses números, os quais foram obtidos nas Seções de Pós-graduação de cada instituto do campus. Como podemos ver na tabela, somente o IQSC⁵ não apresenta um crescimento do número de estudantes de língua espanhola nesses anos, contrariando o cenário dos outros institutos em que o número de estudantes estrangeiros, cuja língua nativa é o espanhol, tem aumentado.

Tabela 1.1: Quantidade de estudantes e a porcentagem de estrangeiros, cuja língua nativa é o espanhol, nos programas de pós-graduação da USP de São Carlos.

	2007		2008		2009		2010	
	Total	%	Total	%	Total	%	Total	%
ICMC ⁶	104	3%	114	9%	228	6%	126	20%
IQSC	72	3%	72	0%	80	2%	87	0%
IFSC ⁷	326	2%	331	2%	355	3%	375	4%
EESC ⁸	1.246	3%	1.271	3%	1.244	5%	1.306	5%

Esse cenário propicia um aumento tanto na quantidade de pesquisas que buscam métodos mais eficientes de ensino/aprendizagem do português por hispano falantes (Evers et al., 2011; Carvalho et al., 2010; Henriques, 2004; Grannier e Carvalho, 2001; Durão, 1999), quanto em iniciativas para a criação de novos recursos, como o projeto *Portuguese for Spanish Speakers*, desenvolvido na Universidade do Arizona (Carvalho et al., 2010).

Pesquisadores da área de ensino/aprendizagem do português para estrangeiros (Evers et al., 2011; Carvalho et al., 2010; Henriques, 2004; Grannier e Carvalho, 2001; Durão, 1999) mostram uma crescente preocupação pela falta de recursos e métodos capazes de aprimorar o processo de ensino/aprendizagem, seja para aprendizes de língua próximas, como o espanhol quanto para aprendizes de línguas distantes, como o chinês ou japonês. Diferentemente da língua inglesa, a língua portuguesa disponibiliza um número reduzido de recursos ou métodos automáticos para auxiliar o aprendizado do português por estrangeiros. Isso é evidenciado no exame de proficiência da língua portuguesa, o CELPE-Bras, que ainda é realizado de forma manual, enquanto seu equivalente para o inglês, o TOEFL⁹, há tempo é automatizado. No entanto, alguns autores, como Da Silva (2010), Mohr (2007), Gomes (2002), Furtoso e Gimenez

⁵Instituto de Química de São Carlos (<http://www.iqsc.usp.br/>).

⁶Instituto de Ciências Matemáticas e de Computação (<http://www.icmc.usp.br/>);

⁷Instituto de Física de São Carlos (<http://www.ifsc.usp.br/>);

⁸Escola de Engenharia de São Carlos (<http://www.eesc.usp.br/portal/>)

⁹do inglês *Test of English as a Foreign Language*, <http://www.ets.org/toefl/>

(2000) e Santos (1999), têm estudado os problemas que os hispano falantes enfrentam na aprendizagem do português considerando as similaridades entre as línguas. Eles acreditam que, devido à similaridade entre o português e o espanhol, os hispano falantes possuem características diferenciadas em relação aos demais aprendizes.

O fato de 85% do léxico do português ter cognatos em espanhol e das estruturas linguísticas das duas línguas serem altamente coincidentes é considerado vantagem na aprendizagem do português para os hispano falantes (Furtoso e Gimenez, 2000; Henriques, 2004). Porém, essas similaridades geram alguns aspectos pouco positivos na aprendizagem do português, tais como: (i) o impedimento, por parte do aprendiz, da capacidade de identificar em qual das línguas ele está se comunicando; (ii) aproximadamente 20% dos cognatos são falsos (Furtoso e Gimenez, 2000; Henriques, 2004); (iii) fossilização da interlíngua¹⁰ no início do processo de aprendizagem devido ao entendimento recíproco entre os interlocutores; e (iv) plena aceitação da interlíngua do hispano falante pelos nativos de português, o que leva o aprendiz a uma acomodação no seu nível de proficiência (Furtoso e Gimenez, 2000). Além disso, há um número reduzido de fonemas vocálicos em espanhol e a inexistência de vogais abertas e nasais, bem como de todos os ditongos abertos em português, dificultando a aprendizagem.

Segundo Gomes (2002) e Furtoso e Gimenez (2000) a 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 interferências da língua espanhola na fala e na escrita em língua portuguesa. Além disso, segundo Scaramucci e Rodrigues (2004), as produções realizadas por hispano falantes também apresentam outros problemas, comuns a estrangeiros falantes de outras línguas e aos próprios brasileiros.

Diversos estudos (Furtoso e Gimenez, 2000; Grannier e Carvalho, 2001; Gomes, 2002; Rojas, 2006; Mohr, 2007) têm sido propostos para compreender os principais problemas que hispano falantes enfrentam na produção escrita do português. Uma avaliação dos textos produzidos no exame CELPE-Bras, apresentada por Mohr (2007) indicou que os hispano falantes têm vantagens em comparação a outros aprendizes estrangeiros na compreensão de aspectos explícitos do exame, mas têm dificuldades em compreender e produzir tex-

¹⁰Segundo o **Dicionário de términos clave de ELE** a interlíngua é o sistema linguístico do estudante de uma segunda língua nos diversos níveis do processo de aprendizagem. Este sistema linguístico é individual e dinâmico e media entre o sistema da língua materna e a segunda língua do aluno.

tos de gêneros diversos, em relacionar ideias e posicionar-se, adequando a linguagem a diferentes situações. Por outro lado, Durão (1999) e Grannier e Carvalho (2001) demonstraram que os nativos de espanhol enfrentam diversos problemas nos níveis lexical, sintático, semântico e discursivo. Por último, Henriques (2004) considera que, em etapas iniciais de aprendizagem, esses usuários enfrentam problemas com (i) falsos cognatos¹¹ ou interferências no nível semântico, que afetam a coerência do texto; (ii) marcadores discursivos¹², que comprometem a coesão e a coerência do texto; (iii) expressões idiomáticas; e, finalmente, (iv) em outros aspectos relacionados com o discurso: organização de ideias, estruturação do texto e progressão temática.

Os estudos apresentados por Mohr (2007) e Henriques (2004) e Durão (1999) concordam que, para aprimorar a aprendizagem dos hispano falantes, é necessário considerar aquilo que é igual nas duas línguas, aquilo que parece igual, mas não é, e aquilo que concretamente é diferente. Embora nesses estudos tenha havido uma preocupação por levantar os erros que os hispano falantes cometem na escrita em português, nenhum dos estudos formalizou a criação de *córpus* de aprendizes¹³, segundo os princípios da Linguística de *Córpus*. Além disso, os estudos também não disponibilizaram os erros detectados para serem utilizados como referências em processos automáticos. Em geral, as análises foram realizadas considerando um número pequeno de textos, o que inviabiliza o emprego dos textos em análises automáticas.

O processo de aprendizagem de escrita da língua portuguesa pode ser mais complicado ainda quando os aprendizes precisam escrever textos acadêmicos, pois os aprendizes têm que dominar as características do gênero e, além disso, a gramática e o léxico da língua. Uma das alternativas para tratar os problemas que aprendizes de uma língua apresentam tem sido o emprego de ferramentas computacionais que apóiam as diferentes etapas do processo de escrita (planejamento, composição e revisão) de um texto. Grande parte das ferramentas computacionais de auxílio à escrita enfoca o pós-processamento do texto, isto é, trabalha sobre um texto já construído, geralmente um rascunho, com o objetivo de corrigir erros e melhorar aspectos como clareza e

¹¹São pares de palavras em duas línguas que segundo a percepção são similares, mas que têm diferentes significados, por exemplo, “aula”(lição de uma disciplina, em português) - “aula”(sala onde se ministra uma disciplina, em espanhol).

¹²Palavras ou expressões que sinalizam relações entre as ideias do texto, por exemplo: “também”, “entretanto”, “isto é”, “por exemplo”, “além disso”, entre outras.

¹³São coleções de textos produzidos por estrangeiros, aprendizes de uma segunda língua, que exibem os problemas reais que eles enfrentam quando aprendem uma segunda língua ou língua estrangeira.

concisão. Nesse grupo, incluem-se os corretores ortográficos, gramaticais e estilísticos (Dayrell et al., 2012; Uzêda et al., 2008).

Especificamente para a escrita de textos científicos, existem também ferramentas que visam apoiar o usuário no processo de composição e não apenas na revisão. Para o português do Brasil, a ferramenta SciPo¹⁴ (*Scientific Portuguese*) (Feltrim, 2004) é a única ferramenta que auxilia na estruturação retórica de textos científicos escritos em português. O SciPo foi criado para auxiliar escritores nativos e iniciantes na escrita de textos científicos em português. Embora o SciPo se apresente como uma alternativa muito eficiente no auxílio à escrita de textos científicos em português, está direcionado a um usuário que talvez não conheça o gênero científico, mas que conhece o português, por ser sua língua nativa. Assim, é de se esperar que os hispano falantes enfrentem problemas na escrita dos textos para os quais o SciPo não foi projetado para resolver.

Atualmente, a ferramenta mais utilizada para apoiar a escrita em português brasileiro é o corretor gramatical inserido na ferramenta *Microsoft Office 2003, 2007 e 2010*, que em sua versão inicial foi chamado de ReGra (Nunes et al., 1996; Martins et al., 1998; Bonfante e Nunes, 2000; Pinheiro et al., 2001; Hasegawa et al., 2002). Segundo Duran (2008a), os corretores são amplamente utilizados por estrangeiros para consultar léxicos, substituindo as consultas a dicionários tradicionais para verificar a grafia correta das palavras. Apesar disso, Duran (2008a) considera que em geral essas ferramentas ainda não estão preparadas para aprendizes de línguas estrangeiras porque não apresentam recursos que possibilitem a substituição de palavras inexistentes, inferidas por esse tipo de usuários. O corretor gramatical ReGra não foge do cenário anterior, pois ele foi criado para resolver os problemas dos nativos do português.

Até o momento foi possível identificar diversas lacunas relacionadas com a inexistência de métodos e recursos que auxiliem os hispano falantes na escrita em português e que também podem ser úteis em processos de ensino/aprendizagem.

1.2 Contexto Teórico

Técnicas de Processamento de Línguas Natural (PLN) têm sido muito utilizadas no cenário de aprendizado de línguas, especificamente, na produção de

¹⁴<http://www.nilc.icmc.usp.br/scipo>

textos, tanto de nativos como de não nativos de uma língua. Com relação à análise de textos escritos por não nativos ou aprendizes de uma segunda língua (L2) ou língua estrangeira (LE), as técnicas de PLN podem ser utilizadas para analisar a escrita dos aprendizes com a finalidade de: (i) criar sistemas tutores inteligentes (ICALL)¹⁵; (ii) criar sistemas de avaliação da qualidade da escrita (AES)¹⁶ e (iii) analisar e anotar corpus de aprendizes (Meurers, 2013). O termo segunda língua (L2) vem do inglês, da área nomeada *English as a Second Language* (ESL). ESL refere-se ao estudo do inglês em países em que o inglês é a língua de comunicação diária. O termo língua estrangeira (LE), também, vem do inglês, da área *English as a foreign language* (EFL). EFL refere-se ao estudo do inglês em um ambiente em que outra língua é a língua usada diariamente.

Os tutores inteligentes foram propostos para oferecer sugestões individuais para atividades e exercícios desenvolvidos pelos aprendizes. Alguns sistemas desse tipo são Robo-Sensei (Nagata, 2013) e TAGARELA (Amaral e Meurers, 2011). O TAGARELA é um sistema criado, especialmente, para os aprendizes de português. Por outro lado, os sistemas que avaliam a qualidade da escrita (Attali e Burstein, 2006; Burstein et al., 2003) focam na extração de características lexicais, sintáticas, semânticas e de estruturação retórica dos textos e as correlacionam à qualidade do texto. A maioria desses sistemas surgiram após a análise de redações provenientes de testes em larga escala como o *Graduate Management Admissions Test*¹⁷ (GMAT), o TOEFL e o *General Certificate of Secondary Education* (GCSE). Tanto os sistemas ICALL quanto os sistemas AES são criados utilizando-se como referência corpus de aprendizes.

Tradicionalmente, os sistemas ICALL e AES focam na detecção de erros para oferecer os auxílios. Diversos autores consideram que a identificação de erros representa uma fonte de informação muito útil para aprimorar a escrita em uma segunda língua (Leacock et al., 2010; Carrió-Pastor e Mestre-Mestre, 2013). A detecção de erros sempre foi o ponto central de sistemas que auxiliam à escrita, também conhecidos como corretores ortográficos e gramaticais. Diferentemente dos sistemas ICALL e AES, os sistemas de auxílio à escrita têm o propósito de auxiliar os usuários na criação de um texto livre de erros de grafia ou gramaticais.

Os corretores ortográficos e gramaticais foram criados para apoiar a escrita dos nativos e, como consequência, não são adequados para tratar problemas

¹⁵Do inglês *Intelligent Computer-Assisted Language Learning*

¹⁶Do inglês *Automated Essay Scoring*

¹⁷<http://www.gmac.com/gmac/thegmat/>

de escrita de aprendizes de determinada língua (Leacock et al., 2010; Meurers, 2013). Em geral, erros de aprendizes são mais complexos, pois para solucioná-los é comum ter que realizar mais de uma modificação (Tschichold, 1999). Além disso, uma lista de palavras pode não ser suficiente para o aprendiz corrigir o erro, sendo preciso mostrar exemplos de uso da língua, que permitam tanto compreender o erro como corrigi-lo (Meurers, 2013).

O exemplo a seguir, extraído da Revista *Interciencia*, mostra a dificuldade da tarefa de corrigir erros de aprendizes. O erro aparece em *itálico e sublinhado*; nele é observado como o aprendiz usou a palavra "*aledanha*", que não existe nem em espanhol nem em português. Como a palavra gerada não existe em português, então o erro é detectado facilmente, contudo, para corrigi-lo uma alternativa seria analisar o contexto em que ele ocorreu. Para isso, seria necessário inferir que a expressão "*zona aledaña*" é uma colocação¹⁸ do espanhol, que no português significa "zona limítrofe" ou "ao redor". A alternativa proposta para corrigir o erro (mostrada entre parênteses e em **negrito**) considera uma análise que combina o espanhol e o português.

(1) Não há comunidades humanas dentro da reserva, mas *na sua zona aledanha* **(ao seu redor)** a maior parte das terras pertencem a ...

Larsen-Freeman e Long (1992) propuseram dividir as causas dos erros dos aprendizes nas seguintes duas categorias: (i) erros interlinguais, que são produzidos por interferências da língua materna e (ii) erros intralinguais que são produzidos com independência da língua materna e acontecem como produto de deficiências nos processos de ensino/aprendizagem da língua. Essa categorização é aceita atualmente na área de aprendizado de línguas. O exemplo apresentado anteriormente mostra um erro interlingual. Erros intralinguais são os mesmos que cometem os nativos da língua, principalmente em processos de aprendizado.

Diversas abordagens têm sido propostas para construir sistemas de auxílio à escrita que tratam erros produzidos por aprendizes. Os primeiros esforços para detectar erros gramaticais de nativos e não nativos foram baseados em estudos puramente linguísticos que usavam gramáticas e *parsers* sofisticados. Essas abordagens, combinadas com gramáticas de restrições, produziram análises sintáticas completas dos textos de entrada, sendo possível identificar diferentes tipos de erros. A maioria das abordagens anteriores envolveu um processo manual para a criação tanto das regras como das gramáticas.

¹⁸Segundo Nattinger e Decarrico (1992), colocações são sequências que contêm itens lexicais que coocorrem mutuamente.

Com o avanço de técnicas de processamento de línguas naturais, as abordagens baseadas em regras deram lugar a métodos baseados em dados (Leacock et al., 2010). A maioria dos trabalhos baseados em dados utiliza algoritmos de classificação para determinar a palavra correta em um determinado contexto. A abordagem baseada em dados tem sido muito utilizada para detectar e corrigir erros de artigos e preposições (Dahlmeier e Tou Ng, 2011; Tetreault et al., 2010; De Felice, 2008; Andersen, 2007; Genoves et al., 2007). Adicionalmente, há outra linha de pesquisa que emprega modelos de língua com a finalidade de detectar diversas categorias de erros (Atwell, 1997; Stehouwer e Van Zaanen, 2009; Gamon et al., 2008; Hermet e Désilets, 2009; Tsao e Wible, 2009). Nesses trabalhos, os erros são localizados nas regiões em que a probabilidade do modelo de língua é muito baixa.

A maioria dos estudos dedicados à detecção automática de erros é guiada por um *córpus* de aprendizes previamente anotado com as diferentes categorias de erros. A anotação de erros em um *córpus* de aprendizes, contudo, é uma tarefa demorada que requer anotadores treinados. Além disso, para ser um recurso útil na identificação automática de erros, é necessário que o *córpus* anotado contenha um grande número de erros, o que requer a anotação de um grande volume de dados. Em geral, os estudos que empregam modelos de língua para detectar e corrigir erros tentam driblar a necessidade de uso de *córpus* de aprendizes. Outra alternativa para driblar a ausência de erros anotados tem sido a geração de erros artificiais (Dicknson, 2010; Foster e Oistein, 2009). A geração de erros artificiais surge como um método para prever erros dos aprendizes, assim como para criar exemplos negativos para aprimorar algoritmos de aprendizado de máquina.

Erros de diversas categorias já foram tratados pelas abordagens anteriores. Os erros mais estudados na literatura são: uso inadequado de artigos definidos e indefinidos (Dahlmeier e Tou Ng, 2011; De Felice, 2008; Andersen, 2007; Genoves et al., 2007) e uso inadequado de preposições (Dahlmeier e Tou Ng, 2011; Tetreault et al., 2010; Andersen, 2007; De Felice, 2008). Erros lexicais (Leedham, 2012; Simpson-Vlach e Ellis, 2010; Alonso et al., 2010) têm sido menos tratados que os erros de ortografia e de gramática, embora alguns corretores ortográficos e gramaticais possam detectar alguns erros desse tipo. Erros lexicais têm sido mais estudados do ponto de vista linguístico (Carrió-Pastor e Mestre-Mestre, 2013; Castele e Collewaert, 2013; Aidinlou e Mehr, 2012; Dayrell e Aluísio, 2008; Fernández, 2005).

Segundo Alonso et al. (2010), os erros de escolha lexical são os erros mais

frequentes encontrados na produção de aprendizes de uma língua estrangeira, juntamente com erros de uso de preposições. Em geral, os erros lexicais são muito heterogêneos, pois aparecem com outros erros, como erros de preposições e artigos. Os erros lexicais costumam ser idiossincráticos, o que afeta o desempenho de sistemas que detectam essa categoria de erros. Atualmente, as categorias lexicais que têm recebido mais atenção são as colocações, sendo possível identificar diversos métodos para tratar problemas dessa categoria em diversas línguas (Alonso et al., 2010; Chang et al., 2008).

A abordagem lexical proposta por Lewis (1993) considera que os léxicos são a base de qualquer língua. Segundo Lewis (1993), tradicionalmente em processos de ensino e aprendizado é comum focar no estudo da gramática, sem compreender a importância do léxico. Aprendizes de L2 ou LE em etapas iniciais de aprendizado costumam escrever em L2 pensando na língua nativa, o que provoca a ocorrência de erros em diferentes níveis. Além disso, é comum encontrar alunos que empregam palavras e expressões básicas da língua ou que utilizam a tradução literal para produzir textos em L2 (Tang, 2012; Alonso et al., 2010; Duran, 2008a). Segundo Lewis (1993) o estudo de unidades lexicais¹⁹ como locuções adverbiais ("*at times*" / às vezes; "*in turn*" / "por sua vez"); colocações ("*strongly recommended*" / "altamente recomendado"); enunciados institucionalizados ("*If I were you...*" / "se eu fosse você...") e expressões formulaicas ("*The aim of the thesis is to explore*" / "O objetivo da tese é explorar") desempenham um importante papel na aprendizagem de L2 ou LE.

Carrió-Pastor e Mestre-Mestre (2013) consideram que, por meio do estudo das causas dos erros lexicais, podem ser identificados padrões recorrentes que ditariam uma série de medidas para corrigir os erros. Os erros lexicais em inglês identificados pelas autoras foram categorizados segundo a classificação tradicional dos erros lexicais, proposta por James (1998). Desta forma, foram separados em: i) escolha errada da palavra; ii) palavra que não existe em L2, mas existe em L1 e iii) palavra que não existe nem em L1 e nem em L2. Além disso, de uma perspectiva semântica os erros lexicais são classificados como: iv) uma palavra foi utilizada em um contexto quando outra devia ter sido usada e v) erros de colocações (a palavra que acompanha a outra é inapropriada).

Um estudo empírico desenvolvido por Tang (2012) mostrou que realmente a abordagem lexical contribuiu para aprimorar as habilidades escritas dos aprendizes, pois os aprendizes começam ter consciência da existência das unidades lexicais e começam a usá-las com maior frequência. Tang (2012)

¹⁹do inglês lexical chunks

considera a importância em disponibilizar recursos lexicais, dicionários e corpus para aprimorar a aprendizagem dessas expressões. As análises realizadas por Loucky (2013) e Aidinlou e Mehr (2012) e Duran (2008a) recomendaram a apresentação dos léxicos em L2 relacionados com os léxicos na língua nativa, pois assim podem aprimorar a aprendizagem dos itens paralelos.

Adicionalmente, Duran (2008a) acredita que para aprimorar os corretores ortográficos, de forma que auxiliem os aprendizes, é necessário focar nos erros lexicais, que devem ser identificados por meio de análise de corpus de aprendizes de uma mesma língua nativa. Para isso, Duran (2008a) propõe um roteiro com os erros que devem ser tratados: (i) uso de falsos cognatos como se fossem verdadeiros cognatos; (ii) uso de regras para formar palavras na língua estrangeira, a partir de radicais em língua estrangeira, que reflitam palavras da língua materna e (iii) uso de regras de formação de palavras em língua estrangeira a partir de radicais de língua materna. Duran (2008a) considera que léxicos da língua nativa e da segunda língua devem ser combinados para aprimorar os corretores. Além disso, supõe que a customização dos corretores representa uma oportunidade de promover o aprendizado, uma vez que alertará o aprendiz que escreve em ambiente digital sobre possíveis erros que estiver cometendo.

A customização proposta por Duran (2008a) considera importante combinar a análise das duas línguas: a língua materna e a língua estrangeira para ter sucesso na detecção e correção dos erros lexicais apontados. Seguindo essa linha e considerando que a abordagem de tradução automática baseada em probabilidades extraídas de corpus (*Statistical Machine Translation* (SMT)) se apresenta na literatura como um conjunto de técnicas muito promissoras na tradução automática de textos, neste projeto de doutorado propusemos a criação de recursos lexicais por meio dessas técnicas. O sucesso desses métodos estatísticos deve-se, em grande parte, pelo fato de eles se beneficiarem do alinhamento de textos paralelos²⁰ (Caseli, 2003a), tentando extrair o conhecimento de tradução contido em um corpus paralelo alinhado para utilizá-lo na criação de recursos úteis para sistemas de Tradução Automática (TA) (Caseli, 2007). Se esses textos possuírem marcas que identificam os pontos de correspondência entre o texto original e sua tradução eles são considerados alinhados (Caseli, 2003b).

A primeira hipótese levantada neste doutorado é a de que a utilização de

²⁰Segundo a terminologia estabelecida pela comunidade de Linguística Computacional, são textos acompanhados de sua tradução em uma ou várias línguas.

métodos da área de tradução automática permite a criação de recursos lexicais para auxiliar a escrita em português. Acredita-se que esses recursos possam sanar alguns dos problemas, principalmente as interferências do espanhol, que ainda não foram resolvidos nem com o emprego de ferramentas que ensinam, principalmente, a estruturação retórica dos textos, como é o caso do sistema SciPo, nem com corretores ortográficos para o português, que trabalham com nativos do português como usuários alvos. Uma segunda hipótese é a de que o uso de recursos linguístico-computacionais, desenvolvidos para apoiar à escrita no nível lexical, pode melhorar a qualidade linguística dos textos em português produzidos pelos nativos do espanhol.

Esta pesquisa de doutorado propõe uma aproximação entre as seguintes áreas de pesquisas: a tradução automática estatística e o apoio à escrita entre línguas muito próximas, considerando a análise de erros extraídos de corpus de aprendizes. Seguindo essa linha, levantaram-se as seguintes questões de pesquisa:

- (i) Os hispano falantes aprendizes do português conseguem produzir textos acadêmicos sem problemas causados pela interferência da língua espanhola? Se não, quais problemas poderiam ser tratados com recursos e ferramentas de auxílio à escrita?
- (ii) A proximidade entre as línguas é um ponto positivo ou mascara problemas que possam retardar o aprendizado da nova língua e seu uso efetivo no meio acadêmico?

A seguir são apresentados os objetivos e as contribuições desta pesquisa de doutorado.

1.3 Objetivos e Contribuições

Neste projeto de doutorado, denominado HABLA (Hispano falantes Adquirindo uma Base Linguística Acadêmica), aliam-se os seguintes objetivos:

Objetivo I: Analisar os problemas que nativos do espanhol possuem ao escrever teses e dissertações em português, considerando as principais interferências do espanhol que esses alunos pesquisadores mantêm na escrita do português;

Objetivo II: Avaliar os resultados de métodos e recursos léxico-sintático-semânticos baseados em corpus paralelos alinhados, que são utilizados em TA, na área de Sistemas de Auxílio à Escrita, para auxiliar na pós-edição e

composição de um texto científico escrito em português por um aluno pesquisador cuja língua materna é o espanhol.

Para cumprir os objetivos desta pesquisa foram realizados diversos experimentos e análises que resultaram nos recursos apresentados sinteticamente na Figura 1.1.

Como mostrado na Figura 1.1, um corpus de hispano falantes escrevendo textos acadêmicos em português foi compilado. Diversas análises realizadas no corpus de hispano falantes, denominado EspanholAcadêmico-Br, mostraram os principais problemas que hispano falantes enfrentam quando escrevem textos acadêmicos em português. As análises serviram como referência para coletar os subsídios sobre tipos de erros que motivaram a criação dos auxílios de apoio à escrita.

Figura 1.1: Auxílios oferecidos pelos léxicos bilíngues

Resultados de um questionário, aplicado aos hispano falantes aprendizes de português, mostraram que eles estão muito preocupados com a ocorrência de erros de falsos cognatos nos textos. Durante a análise do corpus EspanholAcadêmico-Br foram identificados alguns exemplos que mostraram erros de falsos cognatos. O Exemplo 1 mostra como o hispano falante utiliza a palavra "presentes" no sentido de "atuais". Desta forma, usa uma palavra que existe nas duas línguas. O problema se dá pois em português não é adequado usar a palavra "presente" para referir-se à "atualidade" como utilizado no exemplo. Detectar esses erros automaticamente é difícil, pois é preciso reconhecer que a palavra foi utilizada em um contexto errado.

Exemplo 1: o presente estudo realizou uma pesquisa entre antigos e presentes /(**atuais**) líderes de CTI, para realizar uma análise em profundidade sobre a efetividade e impacto das políticas públicas de CTI.

Além de erros de falsos cognatos, a anotação dos erros no corpus de aprendizes mostrou que os hispano falantes têm problemas em escolher os verbos mais comumente utilizados nas diferentes áreas de pesquisa. O Exemplo 2 mostra como o hispano falante utilizou o verbo "elucidar" em um texto da área de exatas, que não aparece frequentemente nos textos dessa área, sendo preferível utilizar o verbo "esclarecer".

Exemplo 2: O seguinte artigo de pesquisa está a serviço de alunos e docentes do nível superior de três diferentes contextos sociais do México, tendo como objetivo elucidar/ (**esclarecer**) se o uso da estratégia de porta-fólio eletrônico influi no processo de motivação e auto-regulação dos estudantes em matérias correspondentes à área de ciências sociais.

A anotação de erros também mostrou que os hispano falantes cometem erros interlinguais quando usam marcadores discursivos e expressões formulaicas. Observam-se exemplos em que expressões do espanhol foram adaptadas ao português. Nesses casos comprovou-se a geração de palavras e expressões que não existem em português. No Exemplo 3 apresenta-se a expressão "*pondo de manifesto*" que representa uma clara interferência do espanhol.

Exemplo 3: ... desta forma, detalhar propostas de melhoria para futuras pesquisas, pondo de manifesto/ (**evidencia que**) trabalhar sistemática e continuamente a estratégia para se alcançar o maior benefício.

Os erros de marcadores discursivos são similares aos erros de expressões formulaicas, no entanto nessa categoria foi observado que, às vezes, as expressões geradas existem em português, mas nunca fazendo função de marcador discursivo. Um erro desse tipo é apresentado no Exemplo 4 e representou um desafio para esta pesquisa, porque para detectar o erro é necessário identificar que o erro foi utilizado como marcador discursivo. No Exemplo 4, o hispano falante adaptou o marcador em espanhol "*por una parte* " quando devia ter utilizado "por um lado".

Exemplo 4: A aplicação deste modelo permitirá, por uma parte/(**por um lado**), efetuar o reconhecimento, descrição e medição das variáveis qualitativas ...

A partir dos erros e/ou inadequações lexicais detectadas durante a anotação de erros no corpus EspanholAcadêmico-Br foram propostas as formas de auxílio apresentadas na Figura 1.1:

1. Equivalências entre o espanhol e português;
2. Escolhas lexicais em português e
3. Detecção e correção de erros.

Como o nome indica, o primeiro dos auxílios parte do princípio que o hispano falante conhece o termo em espanhol e precisa de um auxílio para conhecer o melhor termo em português. Por outro lado, no segundo auxílio o hispano falante já conhece o termo em português, mas precisa de outra palavra ou expressão para substituí-lo. Finalmente, o terceiro auxílio identifica erros nos textos e tenta corrigi-los.

Para obter as equivalências entre as duas línguas foram testados três métodos, um deles baseado em medidas de similaridade e outros dois métodos baseados em tradução automática estatística. Para criar os auxílios relacionados com as escolhas lexicais foi proposto um método que usando a língua espanhola como pivô, procura sinónimos em português. Finalmente, para detectar e corrigir erros foi adotado um método que usando como referência os erros identificados no *corpus* EspanholAcadêmico-Br gera erros artificiais. O método de geração de erros tenta prever diferentes categorias de erros, sendo seu principal diferencial manter o controle da forma correta da palavra ou expressão. Assim, é possível oferecer a correção do erro ou pelo menos sugestões bem sucedidas.

Os auxílios à escrita foram testados na compilação dos seguintes recursos:

1. léxicos bilíngues de cognatos e falsos cognatos das línguas português e espanhol;
2. listas de pares de marcadores discursivos;
3. listas das expressões formulaicas que aparecem nos textos científicos;
4. listas de verbos relacionados com pesquisa científica em português, recorrentemente utilizados em textos.

Estes léxicos bilíngues servem de auxílio para aprimorar a escrita dos hispano falantes. Como mostrado na Figura 1.1, todos os recursos lexicais foram extraídos de *corpus* paralelos alinhados via métodos de tradução automática estatística. Dado que os léxicos bilíngues contém itens lexicais de diferentes tipos, a forma de compilar cada léxico e os tipos de erros associados a

eles foram diferentes. Esse fato representou um desafio nesta pesquisa, pois diferentes métodos e abordagem tiveram que ser propostas.

Dentre as principais contribuições deste projeto destacam-se:

Do ponto de vista teórico:

Contribuição I: Levantamento e quantificação dos principais problemas causados pelas marcas do espanhol, deixadas nos textos acadêmicos em português escritos por nativos do espanhol;

Outras contribuições foram:

1. Disponibilizar um corpus de hispano falantes produzindo textos acadêmicos em português;
2. Construir uma tipologia dos principais erros dos hispano falantes;
3. Fazer anotação manual de erros, gerando um corpus de treinamento;
4. Coletar subsídios sobre os tipos de erros que motivam os auxílios à escrita;

Do ponto de vista da criação de recursos linguísticos para a área de Processamento de Línguas Naturais:

Contribuição II: Compilação de recursos lexicais, a partir de resumos de artigos, teses e dissertações. Especificamente, a partir de um corpus paralelo de resumos em português e em espanhol, foram criados:

1. listas de cognatos e falsos cognatos das línguas português e espanhol;
2. listas de pares de marcadores discursivos;
3. listas das expressões formulaicas que aparecem nos textos científicos; e
4. listas de verbos relacionados com pesquisa científica em português, recorrentemente utilizados em textos acadêmicos.

Do ponto de vista da criação de subsídios para a área de auxílio à escrita científica:

Contribuição III: Projeto e avaliação de auxílios para apoiar à escrita científica em português por nativos do espanhol, que serviu para validar a hipótese desta pesquisa.

1.4 Organização do Trabalho

A tese está organizada em sete capítulos. No Capítulo 2 apresenta-se uma revisão dos principais *córpus* de aprendizes que aparecem na literatura, suas características e a finalidade para a qual foram compilados. Além disso, é realizado um levantamento dos *córpus* de hispano falantes escrevendo em português brasileiro, considerando as principais dificuldades que os hispano falantes enfrentam quando aprendem o português.

No Capítulo 3 apresenta-se a definição e principais utilidades dos *córpus* paralelos. Além disso, são apresentados os métodos de alinhamento dos *córpus* paralelos propostos para apoiar tarefas de tradução automática. Também são descritos os conceitos relacionados com a tradução automática estatística, para depois introduzir os conceitos dos modelos de língua baseado em *n*-gramas. Depois de introduzidos os aspectos teóricos relacionados com *córpus* paralelos e tradução automática, apresentam-se trabalhos que utilizam essas técnicas para gerar léxicos bilíngues, comentando os principais procedimentos adotados e as metodologias propostas para avaliar os léxicos gerados. Finalmente, apresentam-se métodos computacionais e análises linguísticas que se dedicam aos itens lexicais de interesse desta pesquisa.

No Capítulo 4 são comentadas algumas das ferramentas e métodos, propostos para auxiliar à escrita dos aprendizes de uma segunda língua ou língua estrangeira. Adicionalmente, são apresentadas as abordagens computacionais para tratar diversos tipos de erros gramaticais e as metodologias utilizadas para avaliar o desempenho dos métodos.

No Capítulo 5 é apresentada a base de conhecimento do projeto HABLA. Esta base de conhecimento é composta por *córpus* paralelos utilizados para extrair léxicos bilíngues e pelo *córpus* de aprendizes EspanholAcadêmico-Br, compilado para identificar os problemas que enfrentam os nativos de espanhol na escrita de textos acadêmicos. Nesse capítulo, também, são apresentadas todas as análises propostas para caracterizar a escrita dos hispano falantes.

No Capítulo 6 apresentam-se os métodos propostos para compilar os léxicos bilíngues e, como consequência, os auxílios para tratar alguns dos problemas que hispano falantes enfrentam quando produzem textos acadêmicos em português. Finalmente, no Capítulo 7 apresentam-se as contribuições, limitações e trabalhos futuros deste projeto de pesquisa.

Córpus de Aprendizes

Neste capítulo são apresentados diversos estudos que tiveram como finalidade analisar a produção escrita dos aprendizes de uma segunda língua (L2) ou língua estrangeira (LE). Por meio dos trabalhos identificados foi possível conhecer as abordagens utilizadas para solucionar alguns problemas teóricos e práticos propostos neste doutorado, ao mesmo tempo em que foi possível identificar algumas das lacunas existentes na área de pesquisa. Na Seção 2.1 apresentam-se trabalhos dedicados à criação de corpú de aprendizes. A análise destas pesquisas serviu para apontar as principais características destes recursos e suas utilidades. Além disso, na mesma Seção 2.1 foram inventariados os corpú de aprendizes que existem para o português do Brasil. Na Seção 2.2 identificam-se algumas deficiências dos corpú de aprendizes de português. Finalmente, na Seção 2.3 apresentam-se dois estudos que apontam os principais erros que hispano falantes, aprendizes de português, enfrentam quando escrevem nessa língua que serviram de base para a tipologia de erros propostas nesta pesquisa de doutorado.

2.1 Córpus de Aprendizes: Definição e Trabalhos da Literatura

Os corpú de aprendizes são coleções de textos produzidas por estrangeiros ou aprendizes de L2 ou LE, que têm como objetivo principal descrever a língua dos aprendizes (Dagneaux et al., 1998). Esses recursos costumam ser

muito úteis em pesquisas dedicadas a conhecer o processo de aprendizagem de uma língua estrangeira, assim como para procurar métodos mais eficientes no ensino/aprendizagem da língua. Seguindo essa linha, os pesquisadores empregam os corpúsculos de aprendizes para estudar questões teóricas ou educacionais, enquanto os educadores se concentram em identificar as necessidades dos aprendizes. De uma perspectiva pedagógica, diversos esforços têm sido desenvolvidos para criar ferramentas que auxiliem os aprendizes na escrita e leitura de aspectos lexicais, gramaticais, semânticos e/ou discursivos dos textos (Da Silva, 2010; Granger, 2004; Grannier e Carvalho, 2001; Durão, 1999; Berber Sardinha, 2001).

A linguística de corpúsculos tem utilizado três tipos de análises para, usando os corpúsculos de aprendizes como base de conhecimento, identificar os elementos que podem causar problemas no desempenho linguístico dos aprendizes. Estas análises são:

- a Análise Contrastiva (AC);
- a Análise de Erros (AE); e
- a Análise de Desempenho (AD) (Santos, 1999; Granger, 2004).

Segundo Santos (1999), a AC considera que o principal problema de um aprendiz de qualquer LE ou L2 é a interferência da sua língua materna (L1) no processo de aprendizagem. Desta forma, a AC concentra-se em comparar os elementos de L1 e L2 para reconhecer as principais dificuldades do aprendiz. Segundo Santos (1999), para realizar uma AC é necessário, em primeiro lugar, uma descrição das duas línguas, seguida da seleção de certos critérios de comparação das línguas, para depois realizar a comparação e, finalmente, prever os possíveis erros ou dificuldades. O procedimento proposto pela AC apresenta alguns problemas, uma vez que é impraticável comparar inteiramente duas línguas. Santos (1999) também comenta que a maior crítica à AC é o fato dela supor que os únicos erros cometidos por um aprendiz de uma LE estão relacionados com a interferência da língua nativa. Essa suposição é refutada por Grannier e Carvalho (2001), Durão (1999) e Rojas (2006), os quais identificam erros intralinguais¹ de simplificação e de super generalização na análise de erros de aprendizes de português brasileiro.

A segunda análise, a AE, também comentada por Santos (1999), é amplamente divulgada em pesquisas de ensino/aprendizagem de língua. Esse

¹São os mesmos erros que cometem as crianças no processo de aprendizagem da sua língua nativa.

enfoque analisa, na linguagem produzida pelo aprendiz de uma L2, o desvio sistemático das normas dessa mesma língua. A AE parte do pressuposto de que o processo de aquisição da língua nativa poderia ser o mesmo que o da L2 e a partir disso identifica similaridades entre os erros mais frequentes na aprendizagem de uma língua por um nativo e por um estrangeiro. Apesar de a AE não ser uma abordagem perfeita porque não consegue revelar alguns erros, segundo Santos (1999), há pesquisadores que consideram que não há necessidade de usar a AC porque AE é igualmente satisfatória, mais promissora e consome menos tempo. Porém, segundo Santos (1999), há alguns tipos de erros que a AE não consegue registrar.

Para superar as deficiências da AE, a outra abordagem utilizada é a AD (Santos, 1999), que faz uma análise geral do desempenho total do aluno, segundo vários critérios e sem restringir-se apenas à análise dos erros. A maioria dos exames de proficiência de línguas estrangeiras, por exemplo, o CELPE-Bras e o TOEFL, propõe avaliar o desempenho do estudante considerando aspectos gerais e não focam em erros específicos, seguindo, desta forma, a abordagem AD.

Métodos automáticos para analisar os *corpus* de aprendizes têm aprimorado os estudos de ensino/aprendizagem de uma L2 e, por meio deles, diversos recursos linguísticos e computacionais têm sido disponibilizados (Granger et al., 2007; Granger, 2004). Alguns métodos automáticos extraem desde informações básicas, tais como: total de palavras no texto, lista de palavras mais frequentes, número de parágrafos ou sentenças, número médio de palavras por sentenças e outras informações mais sofisticadas, como, ocorrência de colocações ou expressões fixas (Granger et al., 2007). Etiquetadores morfosintáticos e lematizadores são as ferramentas computacionais mais frequentemente utilizadas para anotar os *corpus* de aprendizes. Essas ferramentas são totalmente automáticas e trazem inúmeras vantagens para as análises linguísticas e computacionais dos *corpus*. Porém, essas últimas ferramentas devem ser, idealmente, treinadas nos *corpus* dos aprendizes porque há erros nesses textos que uma ferramenta produzida para textos de nativos não conseguiria classificar corretamente.

Por outro lado, além de empregar métodos automáticos nos *corpus* de aprendizes, comumente são criados processos manuais para anotar nos textos diversos tipos de informações. Nesses processos, as redações do *corpus* são anotadas seguindo um conjunto de etiquetas previamente definidas. Há outros tipos de anotações bem sucedidas e comumente utilizadas nos *corpus*

de aprendizes, algumas delas são linguísticas, i.e., marcações discursivas, semânticas, sintáticas, e anotação de outros tipos de aspectos, como, erros de diversas categorias.

A seguir são detalhados diversos trabalhos cuja finalidade foi analisar a escrita dos aprendizes de L2. Esses trabalhos empregaram a análise contrastiva e a análise de erros como metodologia para analisar os corpú de aprendizes. Como resultado do estudo confirmou-se que na AC a anotação dos corpú não é o foco principal, sendo mais importante fazer um levantamento dos vocabulários mais frequentes, marcadores discursivos, colocações ou expressões, itens gramaticais e estruturas sintáticas empregadas pelos aprendizes (Da Silva, 2010; Allen, 2009; Dayrell e Aluísio, 2008; Dayrell, 2008). Por outro lado, os trabalhos que seguem a AE, mostram sua preocupação pela anotação rigorosa do corpú seguindo etiquetas predefinidas de erros (Dahlmeier e Tou Ng, 2011; Genoves et al., 2007). Alguns dos erros mais estudados são: uso de artigos, preposições, erros lexicais e gramaticais, sendo minoria os trabalhos relacionados com o uso de tempos verbais.

Na literatura, são diversos os estudos que focam na construção de corpú de aprendizes para a língua inglesa (Dahlmeier e Tou Ng, 2011; Allen, 2009; Dayrell, 2008; Tagnin, 2006; Granger, 2004; Berber Sardinha, 2001). Alguns deles (Granger, 2004; Grannier e Carvalho, 2001; Berber Sardinha, 2001) surgiram na tentativa de aprimorar o ensino/aprendizagem da língua inglesa como L2. Em contraste, são discretos os esforços para criar corpú de aprendizes para o português de Brasil como L2. Nesse cenário, é possível identificar algumas pesquisas para estudar os principais problemas que hispano falantes enfrentam quando aprendem português (Da Silva, 2010; Grannier e Carvalho, 2001; Durão, 1999; Rojas, 2006). Porém, o número de textos analisados nestas pesquisas é pequeno e não se alinham com os princípios da linguística de corpú como será mostrado neste capítulo. A Profa. Sylviane Granger compilou uma longa lista² de mais de 100 corpú de aprendizes de várias línguas que merece destaque, pois mostra a quantidade de pesquisas concentradas na análise da produção escrita dos aprendizes de línguas estrangeiras.

O primeiro projeto de corpú de aprendizes de âmbito internacional foi o ICLE³ (*International Corpus of Learners English*), constituído de vários subcorpú de redações de aprendizes de inglês, dos níveis intermediário e avançado

²<http://www.uclouvain.be/en-cecl-lcworld.html>

³<http://www.uclouvain.be/en-cecl-icle.html>

de diferentes línguas maternas (Tagnin e Fromm, 2011; Duran, 2008a). Os subcórpus do ICLE permitem comparar as variações da língua inglesa de nativos e não-nativos, assim como as variações de não-nativos com diferentes línguas maternas. Desta forma, é possível identificar características comuns na aprendizagem do inglês, assim como características específicas dos nativos de diferentes línguas. A parte referente ao português brasileiro do projeto ICLE, o Br-ICLE⁴, está sendo construída desde 2001 por Berber Sardinha (2001). O Br-ICLE é composto por textos argumentativos produzidos em algumas universidades associadas ao projeto e em 2011, o córpus tinha um total de 80.000 palavras (Tagnin e Fromm, 2011). O projeto Br-ICLE disponibiliza um conjunto de recursos que podem ser utilizados livremente para diversos tipos de aplicações. Alguns desses recursos são: lista das 100 palavras mais frequentes, a concordância dessa lista de palavras e ferramentas para realizar buscas nos textos e para anotar erros com sua correção, como, o *toolkit Error Editor*.

Em uma versão do ICLE lançada em 2009, o projeto incluía 16 línguas diferentes, mas ainda não contava com a língua portuguesa. Atualmente, o córpus ICLE continua se expandindo e o subcórpus do português do Brasil ainda não pertence ao córpus do ICLE, tendo se formado como um projeto a parte, dada a disponibilização restrita dos textos do córpus. Córpus de outras línguas como o espanhol, francês, chinês e da língua russa já foram incluídos no ICLE. Alguns dos córpus referentes ao projeto ICLE são o córpus *Japanese English as a Foreign Language Learner (JEFL)*⁵, e o córpus *Cambridge Learner Corpus (CLC)*⁶.

O córpus CLC⁷ (*Cambridge Learner Corpus*) é um córpus de aprendizes aberto e comercial que auxilia o reconhecimento dos léxicos, padrões ou estruturas gramaticais que os aprendizes de inglês usam correta ou incorretamente. Esse córpus é frequentemente utilizado na criação de dicionários e outros recursos inseridos em processos de ensino do inglês. O córpus é composto, atualmente, por mais de 50 milhões palavras e os textos foram produzidos em exames de proficiência de inglês por aprendizes de diversas línguas maternas. Os textos do córpus CLC contêm as avaliações recebidas nos exames de proficiência. Desta forma, o córpus também é utilizado para

⁴<http://www2.lael.pucsp.br/corpora/bricle/>

⁵<http://jeffl.corpuscobo.net/>

⁶<http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/?sitelocale=enGB>

⁷<http://www.cambridge.org/us/cambridgeenglish/about-cambridge-english/cambridge-english-corpus>

identificar as dificuldades que os aprendizes apresentam nos diferentes níveis de proficiência.

Outro projeto significativo, com foco aos aprendizes brasileiros, com várias línguas além do inglês, é o *córpus de aprendizes COMAprend*⁸ (Tagnin, 2006). Esse *córpus* contém redações de alunos de cursos de diversas línguas estrangeiras como alemão, espanhol, francês, inglês e italiano, com a peculiaridade que todos os aprendizes têm nacionalidade brasileira. O projeto disponibiliza em um site todas as redações dos aprendizes, sendo possível a construção de sub*córpus* de aprendizes (de acordo com uma determinada língua, por exemplo) para estudos específicos. O site disponibiliza ferramentas de análise dos sub*córpus* tais como: gerador de listas de palavras, concordanciador e gerador de *n*-gramas (Tagnin e Fromm, 2011). Desde o ano 2011 há uma iniciativa conjunta da Universidade Federal de Rio Grande do Sul e da Universidade Federal de São Carlos, instituições que oferecem cursos de português para nativos de espanhol, de criar um sub*córpus* dentro do ambiente de COMAprend, com os textos produzidos nestes cursos. Isso representa um passo inicial para a formalização de um *córpus* de aprendizes de português por nativos de espanhol.

Allen (2009) apresenta o *córpus* de aprendizes de nativos do japonês escrevendo em inglês *ALESS - Active Learning of English for Science Students*⁹. Esse *córpus* é constituído por 847 textos acadêmicos em inglês, produzidos por alunos do curso de escrita científica da universidade de Tokyo. O *córpus* com um total de 731,612 palavras foi empregado para analisar a ocorrência de expressões formulaicas, específicas do gênero científico, nos textos escritos por falantes de japonês, aprendizes de inglês. Allen (2009) propõe uma análise contrastiva em que os resultados do *córpus* de aprendizes foram comparados com os resultados de um *córpus* de referência, composto por textos acadêmicos escritos por nativos do inglês, com a finalidade de avaliar os usos das expressões formulaicas em textos de aprendizes e em textos de nativos.

Dahlmeier e Tou Ng (2011), em colaboração com a *National University of Singapore (NUS)*, criaram o *córpus NUS Corpus of Learner English (NUCLE)*, constituído por 1.400 redações de diversos tópicos, produzidos por estudantes da universidade de Singapura. O *córpus* contém aproximadamente um milhão de palavras e está completamente anotado com etiquetas de erros e suas correções. Esse *córpus* é o primeiro disponibilizado livremente em conjunto com

⁸<http://www.fllch.usp.br/dlm/comet/comaprend.html>

⁹<http://aless.ecc.u-tokyo.ac.jp/>

a anotação para ser utilizado com fins de pesquisas. A análise do processo de anotação desse corpus de aprendizes foi apresentada por Dahlmeier et al. (2013). Essa anotação foi realizada após a definição de uma tipologia de erros, criada em um estudo piloto previamente desenvolvido. Os autores apresentam os resultados da tarefa de anotação considerando as concordâncias dos anotadores, as quais foram calculadas por meio da medida *Kappa*. O corpus foi anotado pela ferramenta *Writing, Annotation and Marking Platform (WAMP)*, desenvolvida pelo grupo de pesquisa dos autores. O Exemplo 1 mostra uma sentença anotada, sendo possível observar o erro, anotado entre colchetes, que pertence à categoria de tempo verbal, e sua correção.

Exemplo 1: A university [**had conducted** - **conducted**] the survey last year.

A estatística *Kappa* (*K*) (Carletta, 1996) é comumente usada em tarefas de anotação para medir a concordância entre as anotações de dois ou mais anotadores. A Equação 2.1 mostra o cálculo da *Kappa*.

$$K = \frac{P_A - P_E}{1 - P_E} \quad (2.1)$$

em que, P_A é a taxa de concordância observada e P_E é uma estimativa da concordância ao acaso. A medida *Kappa* é amplamente utilizada em estudos na área de Linguística Computacional para avaliar: a concordância entre humanos na execução de uma tarefa; entre humanos e um sistema computacional ou entre vários sistemas. A *Kappa* tem valores entre 0 e 1 que indicam total discordância e total concordância, respectivamente. A Tabela 2.1 foi obtida em um estudo guiado por Landis e Koch (1977) e serve como referência para interpretar os valores da *Kappa*. No entanto, segundo a literatura, não é adequado considerar que os intervalos apresentados na Tabela 2.1 são aplicáveis para todos os estudos, pois existem diversos problemas que são tratados de forma diferente, por isso cada autor deve interpretar os valores *Kappa* de acordo com o fenômeno que está sendo estudado. Por exemplo, em tarefas de análise de conteúdo uma *Kappa* de 0,8 faz com que a anotação seja confiável. Adicionalmente, diversos autores têm demonstrado que a tarefa de anotação de erros gramaticais é muito complexa, aceitando valores de *Kappa* mais baixos. Por exemplo, Leacock et al. (2010) reportaram valores de *Kappa* de 0,63 na tarefa de anotação de erros relacionados com o uso inadequado de preposições. Já, Dahlmeier et al. (2013) obtiveram valores de 0,38 na identificação de erros de diversas categorias; 0,54 na categorização dos erros identificados e 0,48 na tarefa de correção dos erros.

Tabela 2.1: Interpretação dos valores da medida *Kappa* (Landis e Koch, 1977)

Valores de Kappa	Nível de concordância
< 0	Nenhuma
(0 - 0,20)	Pouca
(0,21 - 0,40)	Mediana
(0,41 - 0,60)	Moderada
(0,61 - 0,80)	Substancial
(0,81 - 1)	Quase perfeita

Dayrell (2008) apresenta um *cópus* de resumos de textos acadêmicos em inglês, separados em dois grupos. O primeiro grupo contém um total de 159 resumos produzidos por estudantes da pós-graduação que têm o inglês como LE. O segundo grupo do *cópus* contém a mesma quantidade de resumos, mas desta vez os textos foram selecionados de artigos já publicados. O *cópus* é composto por 33.836 palavras e foi criado com o objetivo de examinar textos produzidos por estudantes brasileiros quando escrevem em inglês e textos em inglês já publicados, que podem ter sido escritos por nativos ou não nativos da língua inglesa, mas que foram revisados. O foco da análise foram as escolhas lexicais e as colocações utilizadas em cada um dos resumos acadêmicos.

Genoves et al. (2007) disponibilizou um *cópus* composto por 723 resumos de textos científicos provenientes dos jornais mais importantes de Física (369) e Ciências Farmacêuticas (354). Os resumos que integram o *cópus* foram escritos por nativos da língua inglesa. Esse *cópus* contém um total de 4.886 sentenças e 115.913 palavras e foi anotado morfossintaticamente. O principal objetivo do estudo de Genoves et al. (2007) era identificar de forma automática erros no uso dos artigos do inglês por nativos do português. Outro *cópus* com resumos produzidos por 78 estudantes de Física e Ciências Farmacêuticas, nativos da língua portuguesa foi criado para testar o desempenho de dois algoritmos de classificação de componentes da estrutura retórica de resumos. Este novo *cópus* de aprendizes é composto por 570 sentenças.

Os trabalhos apresentados até agora mostram *cópus* de aprendizes criados para o inglês como L2. A seguir, são apresentados alguns estudos, em que *cópus* de aprendizes para o português como L2 foram utilizados.

Uma das pesquisas dedicadas ao estudo do português produzido por nativos do espanhol foi dirigida por Da Silva (2010). O foco deste trabalho era analisar as peculiaridades da função dos pronomes como complementos verbais em produções orais em português por estudantes e professores de português na Argentina. A finalidade principal do estudo de Da Silva (2010)

era auxiliar em questões referentes ao uso da gramática no processo de ensino/aprendizagem. O estudo pretendia entender os aspectos sintáticos da interlíngua de um aprendiz de português cuja língua nativa é o espanhol; especificamente, observar a presença ou omissão dos complementos dos verbos e identificar suas diferentes formas de realização junto com os motivos para esses usos.

Da Silva (2010) partiu da premissa de que apesar do espanhol e o português terem muitas afinidades entre si, além de muitos aspectos linguísticos em comum, há pontos, por exemplo, os pronomes como complementos verbais, que distanciam as duas línguas. Para desenvolver a análise, Da Silva (2010) criou um *cópus* com as produções orais de seis falantes de espanhol com vários níveis de proficiência e experiência com a língua portuguesa. Todos os usuários que participaram da pesquisa estavam aprendendo o português e se formando para ensinar essa língua nas universidades argentinas. Os resultados dessa pesquisa indicaram que os aprendizes que participaram do estudo produziram formas de uso dos pronomes como complementos verbais semelhantes àsquelas produzidas por nativos do português, apesar de constatar que existem algumas diferenças.

Evers et al. (2011) descreveram uma proposta inicial de construção de um *cópus* de aprendizes de português de diferentes nacionalidades que representa um dos primeiros esforços no caminho da produção de recursos que contribuam com o desenvolvimento de materiais didáticos para a aprendizagem de português, além de materiais que apoiem a escrita de textos conseguindo oferecer *feedback* qualificados e imediatos. Dessa forma, Evers et al. (2011) apresentam a compilação de um *cópus* de aprendizes de 16 textos com um total de 8.873 palavras, obtido de textos produzidos para o blog de uma escola particular especializada em português como língua adicional¹⁰. Desse *cópus*, foram extraídas as listas de palavras e *clusters* de três palavras do *cópus* de estudo e do *cópus* de referência. Em conversas particulares com as autoras deste estudo foi-me apresentada a compilação de um *cópus* maior de textos extraídos do exame CELPE-Bras, compilado com a finalidade de ser anotado a partir dos erros que forem identificados. A ideia das autoras é estabelecer uma tipologia de erros que respeite os critérios de avaliação considerados no exame de proficiência da língua portuguesa, além de incluir outros critérios para aprimorar a avaliação proposta.

¹⁰A designação “língua adicional” é também utilizada quando se trata de ensino de língua estrangeira (LE) ou segunda língua (L2).

Rojas (2006) criou um *córpus* com trechos de entrevistas orais realizadas com seis hispano falantes. Na pesquisa, o autor fez um levantamento das características da interlíngua produzida pelos hispano falantes em diferentes estágios da aprendizagem do português. Após a análise das entrevistas, o autor propôs uma tipologia de erros, que mostra a natureza de alguns dos erros produzidos pelos hispano falantes. Além disso, Grannier e Carvalho (2001) criaram um *córpus* com 15 exames escritos no CELPE-Bras para fazer um levantamento dos principais erros que nativos do espanhol mantêm na escrita em português. Os textos no *córpus* de Grannier e Carvalho (2001) foram distribuídos em três grupos de acordo ao grau de proficiência do candidato (básico, intermediário e avançado). Os textos são de gêneros diversos e os erros identificados foram classificados utilizando critérios puramente linguísticos.

Finalmente, Durão (1999) analisou a produção escrita de 24 hispano falantes aprendizes de português e brasileiros aprendizes de espanhol. O *córpus* criado por Durão (1999) foi utilizado para criar uma tipologia dos principais erros que ambos aprendizes apresentam.

Na Tabela 2.2 apresenta-se um resumo com as principais características dos *córpus* de aprendizes que foram comentados nesta seção.

2.1.1 Características dos *Córpus* de Aprendizes

Analisando-se os trabalhos apresentados na Seção 2.1, é possível inventariar as características comuns dos *córpus* de aprendizes. Algumas dessas características são: tamanho do *córpus*, propósito do *córpus*, nível de proficiência dos aprendizes, anotações tanto manuais quanto automáticas, assim como diversas ferramentas ou insumos utilizados nos estudos dos *córpus*. Essas características já foram comentadas em Dagneaux et al. (1998), Granger (1998) e Pravec (2002) e a seguir serão relatadas em relação aos trabalhos apresentados na seção anterior.

A compilação de um *córpus* de aprendizes é uma tarefa meticulosa e demorada (Granger, 1998). Em geral, espera-se criar *córpus* extensos que descrevam com precisão a língua dos aprendizes. Porém, nos trabalhos relatados não há uniformidade no tamanho dos *córpus*. Em *córpus* criados para estudos longitudinais, em que, por exemplo, deseja-se avaliar o desempenho de um aluno em diversas etapas de aprendizado, não é necessário compilar muitas redações produzidas pelo estudante (Granger, 2004). A análise apresentada por Rojas (2006) e Grannier e Carvalho (2001) são exemplos de estudos longitudinais, nos quais deseja-se identificar o erro em função do tempo que o

Tabela 2.2: Resumo dos corpus de aprendizes e suas caractersticas

Autor ou C�rpus	L�ngua (L1) / L�ngua (L2)	Utilidade	Estat�sticas	G�nero
ICLE	v�rias l�nguas -> ingl�s	comparar as inter-l�nguas	80.000 pala- vras	Argumentativo
COMAprend	portugu�s -> v�rias l�nguas	Extrair informa��es para an�lises lin- gu�sticas	C�rpus Aberto, em crescente	Diversos
C�rpus CLC	v�rias l�nguas -> ingl�s	criar dicion�rios para ambientes de aprendizado e reconhecimento dos erros em di- versos n�veis de profici�ncia	50.000.000 palavras	Diversos
ALESS	japon�s -> in- gl�s	estudar express�es acad�micas	847 textos	Acad�mico
NUCLE	v�rias l�nguas -> ingl�s	identificar automa- ticamente erros gra- maticais	1400 textos	Diversos
Dayrell (2008)	portugu�s -> ingl�s	analisar as escolhas lexicais e coloca��es	159 resumos	Acad�mico
Genoves et al. (2007)	portugu�s -> ingl�s	estudar o uso de ar- tigos	570 senten�as	Acad�mico
Da Silva (2010)	espanhol -> portugu�s	uso da gram�tica no ensino/aprendizado	6 estudantes	Diversos (l�ngua falada)
Evers et al. (2011)	v�rias l�nguas -> portugu�s	criar materiais did�- ticos para aprendi- zagem	8.873 pala- vras	Diversos
Rojas, J. P. (2006)	espanhol -> portugu�s	identificar erros	6 hispano fa- lantes	Entrevistas
Grannier (2001)	espanhol -> portugu�s	identificar erros	15 textos	Diversos
Dur�o (1999)	espanhol por- tugu�s -> por- tugu�s espa- nhol	identificar erros	24 alunos	Diversos

aprendiz mora no Brasil ou segundo o desempenho no exame CELPE-Bras, respectivamente. Por outro lado, *córpus* extensos são requeridos quando eles foram criados para apoiar processos automáticos, como os de identificação de erros, por exemplo. Seguindo essa linha, o tamanho do *córpus* está relacionado com a aplicação para a qual o *córpus* será utilizado. Em geral, *córpus* pequenos são construídos em pesquisas que fazem análises pontuais (Evers et al., 2011; Da Silva, 2010; Dayrell, 2008; Genoves et al., 2007; Rojas, 2006), enquanto em outras, como é o caso do ICLE, COMAprend, ALESS e NUCLE, *córpus* maiores foram construídos por meio de um processo iterativo e incremental, de forma que novos textos e novas análises são incluídas ao *córpus*, até se criar um recurso sólido que descreva adequadamente o fenômeno que está sendo analisado (Pravec, 2002).

Outra característica importante dos *córpus* de aprendizes é o domínio ao qual pertencem os textos do *córpus* (Granger, 1998). A maioria dos *córpus* apresentados na seção anterior contém textos escritos, porém os *córpus* compilados por Rojas (2006) e por Da Silva (2010) são compostos por entrevistas orais posteriormente transcritas. Na categoria de *córpus* escritos, o gênero dos textos também é uma aspecto importante, pois influencia nas escolhas lexicais (Granger, 1998). Por exemplo, os textos incluídos nos *córpus* ALESS e NUCLE são do gênero acadêmico, enquanto os textos incluídos no projeto COMAprend foram produzidos em salas de aula e pertencem a gêneros diversos. Adicionalmente, outra característica importante é se o texto foi ou não produzido por meio de uma tarefa guiada com contagem de tempo (Granger, 1998). Este último caso é representado pelos *córpus* CLC, que contém alguns textos produzidos em exames de proficiência da língua inglesa e pelo *córpus* compilado por Grannier e Carvalho (2001) que contém textos do CELPE-Bras. Outro exemplo de *córpus* com essa característica aparece no trabalho apresentado por Da Silva (2010), no qual foi planejado um questionário de perguntas relacionadas com filmes brasileiros assistidos como tarefas de sala de aula.

Como consequência da difícil e demorada tarefa de se criar um *córpus* de aprendizes, a acessibilidade deste recurso converte-se em um elemento de grande valor para pesquisadores, educadores e até para aprendizes da língua. Segundo Pravec (2002), o *córpus* CLC está disponível unicamente para os desenvolvedores do projeto, enquanto o *córpus* ICLE está disponível para pesquisas linguísticas também dos envolvidos no estudo dos sub*córpus*. Em contraste, o *córpus* NUCLE é o único *córpus* com anotação de erros gramaticais de aprendizes de inglês que se encontra disponível para pesquisas nessa

língua. O *córpus* do projeto COMApend também está disponível para pesquisas linguísticas. Por outro lado, os *córpus* apresentados por Rojas (2006), Da Silva (2010) e Evers et al. (2011) são *córpus* pequenos que podem ser acessados mediante pedidos aos autores das pesquisas.

A presença de informações que descrevam as características dos aprendizes que produziram os textos no *córpus* são valiosas em *córpus* de aprendizes, porque permitem criar sub*córpus* a partir de atributos previamente definidos e realizar comparações entre eles (Granger, 1998). Por exemplo, comparar a produção escrita de mulheres e homens, levantar características de aprendizes de diferentes níveis de proficiência, etc.. No projeto COMApend é mantido um banco de dados com as informações dos aprendizes que produzem os textos, desta forma é mantida uma relação entre os textos dos *córpus* e as características dos autores. Da Silva (2010) e Rojas (2006) também levantaram informações relacionadas com o gênero, idade, tempo de aprendizagem do português entre outras características dos autores.

Finalmente, um dos insumos mais valiosos dos *córpus* de aprendizes são os diferentes tipos de anotação, tais como: anotação morfossintática, semântica, discursiva e anotação de diversas categorias de erros. Essas anotações são realizadas nos textos de forma manual ou automática. No primeiro caso é necessário o treinamento de humanos para desenvolver o processo de anotação, já os processos de anotação automática requerem ferramentas robustas que consigam anotar as diferentes informações com precisão. Segundo Pravec (2002), o *córpus* CLC é o único em que mais de um quarto do *córpus* encontra-se anotado considerando os erros dos aprendizes de inglês. Por meio dessas etiquetas é possível realizar buscas pelos diferentes erros no *córpus*. O *córpus* NUCLE também disponibilizou um grande conjunto de textos anotados com diferentes categorias de erros e diferentemente do *córpus* CLC, o *córpus* NUCLE está disponível para diversas pesquisas. Os *córpus* apresentados por Rojas (2006), Durão (1999) e Grannier e Carvalho (2001), embora tenham resultado em uma tipologia de erros, não apresentam a metodologia seguida para a anotação manual dos textos. Além disso, não fazem referência a nenhum outro tipo de anotação.

2.2 Crpus de Hispano Falantes Aprendizes de Portugus

No melhor do nosso conhecimento, ainda no existe um crpus de hispano falantes aprendendo o portugus como lngua estrangeira que seja equivalente a alguma das iniciativas apresentadas na Seo 2.1 deste captulo, que trata do aprendizado do ingls. Embora diversas iniciativas tenham sido desenvolvidas para levantar os problemas que hispano falantes apresentam quando aprendem o portugus, os crpus criados nessas pesquisas so pequenos e mostram exemplos de diferentes tipos de erros, mas no possvel afirmar que os estudos conseguem mostrar uma generalizao desses erros. Como consequncia, por serem pequenos, no podem ser utilizados como insumo para pesquisas computacionais, em que o crpus seja a base de conhecimento.

Adicionalmente ao tamanho do crpus, os estudos que existem para o portugus, embora descrevam com rigor a metodologia utilizada para compilar os crpus, no descrevem o processo de anotao realizado. As pesquisas realizadas por Rojas (2006), Duro (1999) e Grannier e Carvalho (2001) apresentam tipologias de erros, mas nenhuma delas disponibiliza os erros anotados e suas correes. O processo de etiquetamento dos erros foi realizado para analisar os principais problemas e no para uma anlise computacional. Esses autores combinaram uma anlise contrastiva com uma anlise de erros e criaram a tipologia de erros; eles analisaram os textos e, em funo disso, fizeram um levantamento dos erros mais frequentes. Alm disso, nas pesquisas de Rojas (2006), Duro (1999) e Grannier e Carvalho (2001) no se especificam a quantidade de pessoas envolvidas na anlise dos erros, diferentemente dos crpus anotados para o ingls; e a concordncia entre os anotadores envolvidos na tarefa no foi estimada. Finalmente, no levantamento feito para o portugus no foi identificado nenhum estudo em que um crpus do gnero cientfico fosse compilado; em geral, essas pesquisas estudam erros produzidos em textos de gneros diversos.

Para analisar os problemas dos aprendizes do portugus do Brasil, uma fonte importante de conhecimento o exame de proficincia CELPE-Bras¹¹. Esse exame prestado por alunos de diferentes lnguas nativas, sendo a maioria hispano falantes. Antes de fazer o exame, diferentes informaes dos alunos so compiladas, obtendo uma relao do texto e as caractersticas do aluno que o produziu. Com a compilao dos textos produzidos nesse exame

¹¹<http://www.ufrgs.br/acervocelpebras/acervo/>

seria possível criar um corpus de aprendizes adequado para diversas aplicações tanto linguísticas quanto computacionais. A dificuldade se dá pelo fato desse exame ainda ser realizado de forma manual, sendo a transcrição dos textos uma tarefa demorada, que requer tempo e cuidado para sua execução.

2.3 Tipologia de Erros para as Línguas Próximas Português e Espanhol

Devido ao aumento no número de hispano falantes interessados na aprendizagem do português como L2, as pesquisas dedicadas a procurar métodos mais eficientes para a aprendizagem do português por nativos do espanhol têm aumentado nos últimos anos (Carvalho et al., 2010; Henriques, 2004; Granier e Carvalho, 2001; Durão, 1999), juntamente com as iniciativas para a criação de recursos que auxiliem a aprendizagem desse grupo de usuários. Um exemplo dessas iniciativas é o projeto *Portuguese for Spanish Speakers*¹², desenvolvido na Universidade de Arizona e comentado por Carvalho et al. (2010). Outra iniciativa acontece no Celin - Centro de Línguas e Interculturalidade¹³, da Universidade Federal do Paraná. O Celin é um espaço representativo para o desenvolvimento de práticas no campo de aquisição e aprendizado de línguas estrangeiras por nativos de diversas nacionalidades. Essas instituições, que oferecem cursos de português para estrangeiros, estabelecem dois tipos de cursos: um dirigido para estrangeiros de qualquer nacionalidade e outro dirigido, especificamente, para nativos do espanhol. Essa divisão das turmas ocorre porque vários autores (Da Silva, 2010; Mohr, 2007; Gomes, 2002; Furtoso e Gimenez, 2000; Santos, 1999) concordam que, devido à similaridade entre o português e o espanhol, os falantes do espanhol possuem características diferenciadas em relação aos demais aprendizes da língua portuguesa.

Na literatura, a similaridade entre o português e o espanhol apresenta-se como um elemento positivo, que muitas vezes tem se convertido em um obstáculo, devido ao fato da semelhança e da proximidade ocultarem muitas vezes as diferenças entre as línguas e impedirem frequentemente o domínio das mesmas, mantendo, tanto na fala quanto na escrita da língua portuguesa, as interferências da língua nativa, o espanhol (Gomes, 2002; Furtoso e Gimenez, 2000). De fato, as interferências do espanhol têm sido os aspectos mais enfatizados pela literatura que trata da produção escrita dos hispano falantes

¹²<http://www.portspan.cercll.arizona.edu/atividades/atividades.htm>

¹³<http://www.celin.ufpr.br/>

(Mohr, 2007; Scaramucci e Rodrigues, 2004). Contudo, segundo Scaramucci e Rodrigues (2004) as produções escritas pelos hispano falantes também apresentam outros problemas, comuns a estrangeiros falantes de outras línguas e a brasileiros escrevendo na sua língua materna: o português.

Henriques (2004) apresenta um experimento para estimar a proximidade entre o léxico do espanhol e do português. No experimento, a autora fez uma seleção de vários textos científicos curtos para quantificar a frequência de vários tipos de palavras tais como: cognatos, não cognatos e falsos cognatos. De um total de 1.198 palavras que foram analisadas, 1.087 (90,8%) palavras foram identificadas como cognatos (com ortografia idêntica e não idêntica), 56 (4,7%) palavras análogas etimologicamente (porém com algumas divergências semânticas), 27 empréstimos de outras línguas (2,3%), 19 (1,6%) falsos cognatos e 9 (0,76%) termos científicos. Desta forma, Henriques (2004) conclui que todo o conhecimento que o aprendiz/falante transfere, com acerto, da sua língua nativa, o espanhol, para a segunda língua (ou língua estrangeira), o português, facilita o processo de aquisição/aprendizagem e a compreensão (oral e/ou escrita).

Os resultados apresentados por Henriques (2004) mostraram que existe um alto grau de intercompreensão entre os falantes de português e de espanhol, em relação à:

1. compreensão de textos (desde que os falantes usem estratégias de inferência lexical para ajudá-los na compreensão global do texto);
2. tradução (já que existe um alto índice de palavras cognatas com ortografia idêntica ou similar) e
3. inferência de itens lexicais (com a ajuda de estratégias de vocabulário e de gramática).

Contudo, existe uma percentagem de palavras (falsos cognatos, marcadores discursivos e expressões idiomáticas) de difícil inferência, mesmo com a ajuda do contexto em que ocorrem, em qualquer tipo de texto e sem previsão de ocorrência. Com esses resultados, Henriques (2004) considera que em etapas iniciais de aprendizagem é recomendável focar em aspectos relacionados com:

- falsos cognatos ou interferências no nível semântico que afetam a coerência do texto;

- marcadores discursivos, que comprometem a coesão e, consequentemente, a coerência do texto;
- expressões idiomáticas e, finalmente,
- outros aspectos relacionados com o discurso: organização de ideias, estruturação do texto, ponto de vista e progressão temática.

Nas Seções 2.3.1 e 2.3.2 são apresentadas duas tipologias, cujo foco principal foi levantar os principais problemas que hispano falantes enfrentam quando estudam português. Essas pesquisas combinaram as análises AC e a AE para especificar os principais erros dos hispano falantes. Esses estudos tentaram compreender as origens dos diferentes erros, propondo tipologias de erros baseadas na natureza dos problemas. As tipologias de Grannier e Carvalho (2001) e Durão (1999) foram construídas com a finalidade de aprimorar o processo de ensino/aprendizagem do português. Outra tipologia interessante foi apresentada por Rojas (2006), porém ela não será detalhada, pois se propôs a analisar erros que ocorrem na fala dos hispano falantes e foge do interesse desta pesquisa.

2.3.1 Tipologia de Grannier e Carvalho (2001) - Problemas que Hispano Falantes Enfrentam na Escrita em Português

Grannier e Carvalho (2001) realizaram um levantamento dos principais erros que nativos do espanhol cometem quando escrevem em português. O conjunto de erros foi identificado após a análise de 15 provas escritas do exame CELPE-Bras do ano 2000. As provas foram distribuídas em três grupos de acordo com o grau de proficiência do candidato (básico, intermediário e avançado). Os textos são de gêneros diversos e os erros identificados foram classificados utilizando critérios puramente linguísticos.

Os autores classificaram os erros seguindo as categorias:

1. erros lexicais, que reúnem aqueles que incidem em algum aspecto da seleção lexical (na sua forma ou na sua adequação semântica);
2. erros morfossintáticos, são aqueles que afetam a estrutura interna da palavra e seus vínculos com outras palavras, tais como erros de regência e de gênero;
3. erros sintáticos, tais como erros de ordem dos constituintes ou dos conectivos;

4. erros léxico-sintático-semânticos, são 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 e
5. inadequações de registro e de estruturas¹⁴.

Na Tabela 2.3, são mostradas as quatro primeiras categorias de erros, além de alguns exemplos que evidenciam a ocorrência de determinados problemas. A última categoria, **Inadequação de registro**, é apresentada na Tabela 2.4. A seguir, são detalhados os problemas que ocorreram com maior frequência, inclusive em aprendizes com nível de proficiência avançado. Esses erros mais frequentes foram considerados pontos críticos ou erros reiterativos. Na tipologia apresentada por Grannier e Carvalho (2001) apontam-se os significados dos erros mais frequentes, assim como em que momento do processo de ensino/aprendizagem podem ser aproveitados ou indicados. O terceiro tipo de erro, dentro da categoria **Lexicais** (1.3) foi um erro recorrente. Nessa categoria de erros, os autores classificaram produções em que os aprendizes combinaram regras do espanhol com regras do português, como no exemplo da Tabela 2.3. Grannier e Carvalho (2001) consideram que esse erro é um bom indicador para revelar o desenvolvimento do processo de aquisição ou aprendizagem da nova língua.

Dentro da categoria de erros morfossintáticos, foram observados vários problemas relacionados com regência verbal ou uso de preposição em complementos diretos ou em complementos preposicionais. Tanto a falta da preposição quanto o emprego equivocado de uma preposição em lugar de outra foram aspectos frequentemente identificados. Segundo Grannier e Carvalho (2001), esses últimos erros são úteis para identificar os pontos em que o nativo do espanhol consegue se conscientizar das igualdades e diferenças das duas línguas. Erros da categoria léxico-sintático-semântica também foram identificados, sendo que Grannier e Carvalho (2001) acreditam que esses estão relacionados com as distribuições em estruturas diferentes e/ou recortes semânticos com algumas diferenças de uma língua para outra.

A seguir, é apresentada outra tipologia de erros de hispano falantes quando escrevem em português. Essa tipologia também foi obtida após a análise de produções escritas em um ambiente educacional.

¹⁴Os autores não explicam este tipo de erro.

Tabela 2.3: Categorização dos erros que hispano falantes cometem quando escrevem em português (Grannier e Carvalho, 2001).

Tipos de Erros	
1 Lexicais 1.1 Uso da forma em espanhol Compartir (E: <i>compartir</i> ; P:compartilhar) precios (E: <i>precios</i> ; P:preços) 1.2 Uso da forma (quase) em português <i>tambem</i> (P: também) <i>éxtase</i> (P: êxtase) 1.3 Uso de formas mistas 1.3.1 Interferência ortográfica <i>aprossimar</i> (E: <i>aproximar</i> , P: aproximar) <i>assessorio</i> (E: <i>accesorio</i> , P: acessório) 1.3.2 Interferência fonológica <i>posivilidade</i> (E: <i>posibilidadade</i> , P: possibilidade) <i>povre</i> (E: pobre, P:pobre) 1.3.3 Regularização morfológica <i>durme</i> (E: <i>duerme</i> , P: dorme) <i>vevir</i> (E: <i>vivir</i> , P: viver) 1.3.4 Inadequação semântica <i>alargar</i> (E: <i>alargar</i> , P: alongar) <i>trocou</i> (E: <i>cambió</i> , P: mudou) 3 Sintáticos 3.1 Ordem dos constituintes <i>Ajudar para que o indivíduo</i> (P: ajudar o indivíduo para que) <i>Eu só lhe posso dizer</i> (P: eu só posso lhe dizer) <i>Meu chato dia</i> (P: meu dia chato) <i>Não todos somos iguais</i> (P: não somos todos iguais)	2 Morfossintáticos Usos de preposições 2.1 Regência verbal <i>aprender deles</i> (P:aprender com eles) <i>descordar com</i> (P: discordar de) 2.2 Outros usos de preposição <i>de mal em pior</i> (P: de mal a pior) <i>nesse respeito</i> (P: a esse respeito) <i>no longo prazo</i> (P: a longo prazo) 2.2.1 Gênero <i>as outras costumes</i> (P:os outros costumes) <i>o viaje</i> (P: a viagem) 2.2.2 Determinantes e indefinidos <i>a longo do trajeto</i> (P: ao longo) <i>a outra única coisa</i> (P: a única coisa) <i>algum</i> (P: alguém) <i>tudos</i> (P: todos) 4 Questões léxico-sintático-semânticas <i>Estou a favor</i> (E: <i>estoy a favor</i> , P: sou a favor) <i>Prestar ouvidos</i> (P: dar ouvidos/ prestar atenção) <i>É tão mau</i> (E: <i>está tan mal</i> , P: está tão mau) <i>Feria</i> (E: <i>feria</i> , P: feira)

Tabela 2.4: Continuação da categorização dos erros que hispano falantes cometem quando escrevem em português (Grannier e Carvalho, 2001).

Tipos de Erros
5 Inadequações
De registro
5.1 Uso de pronome átono em texto informal
<i>Tem que faze-lo agora</i> (P: deve fazê-lo agora/ tem que fazer isso agora)
5.2 De estrutura
(a) Uso de relativa por estrutura com possessivo
<i>Em relação à natureza, não devemos esquecer que a qualidade mais importante que tem é [...]</i>
(P: [...] à natureza, não devemos esquecer que a (sua) qualidade mais importante)
(b) Uso de relativa por estrutura com infinitivo
<i>Em relação à natureza, [...] a qualidade mais importante que tem é que é única</i>
(P: em relação à natureza, [...] a qualidade mais importante é ser única)

2.3.2 Tipologia de Durão (1999) - Problemas que Hispano Falantes Enfrentam na Escrita em Português

Durão (1999) desenvolveu um estudo relacionado ao aprendizado de espanhol por nativos do português e ao aprendizado do português por hispano falantes. O estudo está baseado na constatação, em um viés pedagógico, de que os pontos de intersecção do português com o espanhol favorecem o processo de ensino/aprendizagem, ao mesmo tempo em que ocultam aspectos diferenciados que se manifestam posteriormente na produção linguística dos aprendizes nos níveis fonético-fonológico, ortográfico, morfossintático e semântico. Devido ao fato da análise contrastiva considerar que as interferências da língua materna são as principais dificuldades para um aprendiz de língua estrangeira e após concluir que as interferências da língua materna não são as únicas, a autora combina a análise contrastiva com a análise de erros na tarefa de identificação dos erros mais frequentes. Desta forma, Durão (1999) concentra-se em determinar estratégias de simplificação, descrever erros léxicos, sintáticos, semânticos e tentar identificar os fatores que provocam esses erros.

Dado o fato da presente pesquisa ter foco na produção escrita em português de nativos de espanhol, nesta seção somente serão comentados os erros dos hispano falantes quando escrevem em português, identificados por Durão (1999). Para analisar esses erros, Durão (1999) selecionou 24 alunos do pri-

meiro ano do Curso de Português oferecido na Universidade de Valhadril, Espanha. Todos os sujeitos eram nativos de países com língua espanhola e com um nível de proficiência entre básico e intermediário.

Na Tabela 2.5, são apresentadas as três categorias de erros apontadas por Durão (1999), assim como as quantidades de cada tipo e sua percentagem com relação ao total. Como a tipologia descrita em Durão (1999) é uma tipologia bem detalhada, às vezes até específica demais, serão comentados e exemplificados os erros que os corretores ortográficos do *Microsoft Word 2007* e do editor de texto incluído no *Latex* não conseguem identificar ou que foram considerados por Durão (1999) como erros graves.

Tabela 2.5: Categorização dos erros que hispano falantes cometem quando escrevem em português (Durão, 1999).

Tipos de Erros		
	Erros	Percentagem
Erros fonológicos e gráficos	595	51,00%
Erros léxicos	386	33,10%
Erros gramaticais	186	15,90%

A primeira categoria - **Erros gráficos e fonológicos** - está relacionada com deficiências tipicamente ortográficas ou motivadas por dificuldades fonológicas. Essa categoria encontra-se subcategorizada em oito erros, apresentados na Tabela 2.6, que além de mostrar a categoria dos erros mostra exemplos mais relevantes, acompanhados de sua correção.

Os erros de **acentuação ortográfica** estão relacionados com o uso errado do acento ou a falta de acentuação, sendo a maioria destes erros identificados por um corretor ortográfico. Os erros relacionados com o **uso errado da crase**, às vezes, é devido ao fato dos aprendizes não diferenciarem entre crase e acento agudo, desconhecendo que a crase aparece após a junção de duas vogais iguais e antecedendo um substantivo feminino. As **marcas erradas de nasalização** são comuns em nativos do espanhol quando escrevem em português, pelo fato de não existir esse som em espanhol. A categoria **separação e união de palavras** mostra que os nativos do espanhol apresentam problemas para identificar quais palavras devem estar separadas ou unidas. Nesse último caso, sempre aparecem as seguintes perguntas entre os hispano falantes: Será que estas palavras são separadas ou unidas? E caso unidas como deve ser a união, usando ou não o hífen? A categoria **separação e união de palavras** inclui erros de palavras compostas separadas por um hífen, além da

Tabela 2.6: Erros fonológicos e gráficos que hispano falantes cometem quando escrevem em português (Durão, 1999).

Erros fonológicos e gráficos		
	Erros	Percentagem
Acentuação ortográfica errada	324	54,50%
Uso errado da crase	9	1,50%
Ex: Voltada para à clientela...-> Voltada para a clientela...		
Marcas erradas de nasalização	136	22,90%
Ex:... não é extranho ...-> Ex:... não é extranho ... Ex:... serem maiores ...-> Ex:... serão maiores ...		
Separação e união de palavras	16	2,70%
Ex: <i>Sub região</i> -> Sub-região // <i>Sub córpus</i> -> Subcórpus Ex: é difícil <i>darse</i> conta -> é difícil dar-se conta // <i>estudarlo</i> -> estudá-lo		
Confusão de Fonemas	47	7,90%
Ex: ...Tenho certeza do resultado <i>sim</i> nenhuma dúvida... -> ...Tenho certeza do resultado sem nenhuma dúvida...		
Confusão de grafemas com fonemas	51	8,50%
Ex: ... <i>devido</i> a...-> ...devido a... // <i>façer</i> -> fazer		
Consoante Surda	12	2,00%
<i>acção</i> -> ação // acto -> ato		

união de pronomes complementos do verbo.

Outro tipo de erro classificado por Durão (1999) como um erro fonológico ou gráfico é a **confusão de fonemas**. Essa classe inclui erros relacionados com palavras similares do espanhol e do português, por exemplo, com alternância vocálica do tipo e/i ou o/u. Outra categoria foi a **confusão de grafemas**, na qual encontram-se os erros provocados pela dificuldade dos aprendizes do espanhol em diferenciar entre v/b, ou ç/z. Finalmente, o erro de **consoantes surdas**, não foram interessantes na pesquisa de Durão (1999), porque esses erros estão relacionados com a versão do português europeu, mas comumente os nativos do espanhol erram quando escrevem palavras como *ação* e *ato*, colocando um “c ”após o primeiro “a ”.

A Tabela 2.7 mostra a categorização feita por Durão (1999) para os erros lexicais, assim como a distribuição de cada conjunto. O primeiro erro da classificação é o **uso de falsos cognatos**. De acordo com Durão (1999), falsos cognatos são as palavras que apesar de sua semelhança não têm o mesmo significado nas duas línguas. A definição anterior é unanimidade entre os linguistas. Segundo Durão (1999), nos pares português e espanhol o tipo anterior não é o mais comum e sim é frequente achar palavras que compartilham um ou mais significados, enquanto outros são distintos. Desta forma, a autora resume o problema de falsos cognatos à frequência de uso de uma palavra, em que em uma língua opta-se pelo uso da palavra, enquanto em outra se usa um sinônimo da palavra, o que não significa que não se tenha a mesma palavra nas duas línguas com o mesmo significado semântico em determinado contexto.

Na categoria **outros problemas semânticos** estão incluídos erros que aparecem devido ao número reduzido do vocabulário que o aprendiz domina e termina sobreusando esses sem identificar um erro na escolha lexical. Exemplo desse tipo de erro é o uso da palavra “gente ”que em espanhol (“*la gente* ”), significa conjunto de pessoas; em português tem o mesmo significado, mas também é usado no sentido do pronome pessoal nós. Outro caso é a palavra “viver ”, que em espanhol tem o sentido de “morar ”e de “residir ”, já em português essa palavra não é comumente usada no sentido de “morar ”. Os **casos de fronteira**, segundo Durão (1999), são empréstimos da língua materna que sofreram alguma adaptação fonética ou gráfica. Finalmente, os erros relacionados com **transferência linguística** referem-se a léxicos da língua materna deixados nos textos na língua estrangeira. Em geral, os erros desse tipo identificados por Durão (1999), são empréstimos que foram produzidos por trans-

Tabela 2.7: Erros lexicais que hispano falantes cometem quando escrevem em português (Durão, 1999).

Erros lexicais		
	Erros	Percentagem
Uso de falsos cognatos	57	14,80%
Ex: Duas prefeituras <i>acordam</i> estabelecer o plano ... -> Duas prefeituras combinam/decidem estabelecer o plano ...		
Outros problemas semânticos	96	24,90%
Ex: É <i>certo</i> que espanhol...-> É verdade que // espanhol // ... Que <i>formam parte</i> de...-> ... Que participam de ... //		
Hoje eu <i>conozco</i> que tenho...-> Hoje eu entendo que tenho...		
Casos de fronteira	168	43,50%
Ex: <i>Ensenhar</i> a língua...-> Ensinar a língua		
Transferência linguística	65	16,80%
Ex: <i>Algunos</i> com uma tradição.. -> Alguns com uma tradição...		

ferência do significado espanhol para um significado que existe em português, mas com outra forma.

Por último, Durão (1999) apresenta uma classificação para erros gramaticais que é apresentada na Tabela 2.8. Dentro das categorias mostradas na Tabela 2.8, há outras categorias mais específicas que serão explicadas a seguir.

Tabela 2.8: Erros gramaticais que hispano falantes cometem quando escrevem em português (Durão, 1999).

Erros Gramaticais		
	Erros	Percentagem
Paradigmas nominais	61	32,80%
Paradigmas verbais	39	20,00%
Outros paradigmas	86	46,20%

Os erros classificados como **paradigmas nominais** estão separados em dois grupos. O primeiro grupo contém erros de mudanças de gênero e o segundo grupo refere-se ao uso inadequado dos artigos. Segundo Durão (1999), os erros de mudança de gênero são produzidos, em geral, por interferências

do espanhol dado que há substantivos com ortografia e significado idênticos, mas com gêneros diferentes. Alguns exemplos são “*o linguagem*” e “*as costumes*”. A maioria dos erros de mudança de gênero são identificados pelo corretor ortográfico do *Word*, enquanto que o corretor incluído no editor do *Latex* não consegue indicar erros desses tipos. Por outro lado, os problemas referentes ao uso dos artigos estão relacionados tanto com a ausência ou presença de artigos quanto a erros sintáticos. Em português, há casos em que o uso do artigo é desnecessário, por exemplo, com verbos relacionados com aprendizagem, tais como “aprender, ensinar,” etc.. Em outros casos, torna-se necessário o uso do artigo, por exemplo, quando são usados nomes próprios ou geográficos. Talvez por desconhecer essas regras o hispano falante comete os seguintes erros identificados por Durão (1999): “*aprender o espanhol*”, “*...só Brasil fala...*”.

Na categoria de **paradigmas verbais** estão incluídos desvios provocados pela regência verbal (Duração, 1999). Esse conjunto de erros inclui problemas relacionados com as conjugações dos verbos irregulares. Durão (1999) separa os erros desse conjunto em dois tipos: o primeiro quando o erro não leva a uma mudança do significado e o segundo quando a mudança acontece. A seguir, exemplos desses tipos de erros junto com suas correções são apresentados:

- *considerar-os* -> *considerá-los*,
- *se não foram duas línguas* -> *se não fossem duas línguas*

Finalmente, na categoria **outros paradigmas**, proposta por Durão (1999), estão incluídos erros relacionados com: confusão no emprego do adjetivo *grande*; confusão no uso de outros elementos gramaticais (considero que estas duas primeiras categorias são muito específicas, por isso não serão explicadas neste documento) e uso incorreto de preposições. O uso incorreto de preposições está associado a erros no momento de contrair as preposições “*em/de*” ou em construções diferentes do espanhol e do português, de forma que em uma língua a preposição é exigida e na outra não. É importante salientar que esses erros não são identificados por nenhum dos corretores gramaticais testados:

- *...longe de fronteira...* -> *...longe da fronteira...*
- *...há em Espanha...* -> *...há na Espanha...*
- *A prova de os estudantes...* -> *A prova dos estudantes...*

- *Por as necessidades...-> Pelas necessidades...*

A tipologia de erros mostrada por Durão (1999) aponta a maioria das deficiências que nativos do espanhol apresentam quando escrevem em português. Considerando a percentagem de erros em cada categoria, é possível identificar alguns erros que foram mais recorrentes que outros. Na categoria de **erros fonológicos e gráficos**, os erros ortográficos foram os de maior incidência. Já na categoria **erros lexicais**, os casos de fronteira e os erros relacionados com outros problemas semânticos foram os que ocorreram um maior número de vezes. Finalmente, os erros de uso incorreto de preposições, erros morfológicos, erros sintáticos, todos dentro da categoria de **outros paradigmas** e erros de mudanças de gênero e de uso incorreto de artigos, ambos da categoria **erros nominais**, foram os de maior ocorrência.

Após o levantamento realizado, Durão (1999) considera que na aprendizagem tanto do espanhol por nativos de português, quanto do português por nativos do espanhol devem-se combinar elementos da análise contrastiva com elementos da análise de erros. Em outras palavras, é interessante que o aprendiz conheça as semelhanças entre as duas línguas e os pontos críticos que podem ser traduzidos em erros.

2.4 Considerações Finais

Neste capítulo foram comentados alguns trabalhos que estudaram a produção escrita dos aprendizes de L2. Constatou-se que a maioria destes trabalhos emprega textos escritos por aprendizes de uma língua para analisar suas características e dessa forma foram apresentados alguns dos *cópus* de aprendizes desenvolvidos. Para o português, o conjunto de redações mais utilizadas para analisar a escrita dos aprendizes foram produzidas no exame CELPE-Bras (Mohr, 2007; Grannier e Carvalho, 2001); no caso da língua inglesa, a prova de proficiência TOEFL tem sido amplamente utilizada.

Especificamente, para o português do Brasil observou-se que alguns esforços têm sido realizados, sendo ainda discretos os estudos e recursos disponíveis para serem empregados na tarefa de detecção automática de erros de aprendizes. Contudo, essa área de pesquisa apresenta-se como uma área em ascensão devido ao aumento do interesse pelo aprendizado da língua portuguesa por nativos de diversas nacionalidades. A revisão da literatura realizada nesta seção permite identificar lacunas especificamente em áreas de pesquisas como ensino/aprendizagem da língua portuguesa, mostrando a necessidade

de criar novos métodos e recursos que auxiliem este processo. Salienta, também, que são muitos os profissionais preocupados com o aprimoramento do ensino da língua portuguesa, seja para qualquer aprendiz ou especificamente para aprendizes de línguas próximas, como o espanhol.

Os estudos apresentados para levantar os problemas que enfrentam os hispano falantes aprendizes de português foram guiados por pesquisadores que utilizando como base a análise contrastiva conseguiram realizar comparações entre o português e o espanhol. Essas análises mostraram o conhecimento comum entre as duas línguas, assim como as divergências. Além disso, os pesquisadores concordaram na importância do aprendiz ou professor conhecer estas similaridades e diferenças para aprimorar o processo de ensino/aprendizagem.

Os estudos baseados na análise de erros apresentaram duas tipologias de erros por meio das quais foram classificados os tipos de erros. As tipologias apresentadas mostram erros em diferentes níveis linguísticos, assim como a frequência com que eles ocorrem nos textos analisados. Porém, nas tipologias apresentadas, ainda faltam outros tipos de problemas, mais específicos e difíceis de tratar, tais como: coesão, coerência, adequação a determinados gêneros discursivos, estruturação do texto, entre outros. As tipologias apresentadas fazem uma definição granulada dos erros, que embora, relacionem adequadamente o erro e sua natureza, é difícil distinguir a fronteira que separa um tipo de erro de outro. Provavelmente, a riqueza no detalhe da classificação dos erros é proveitosa em estudos linguísticos, perdendo clareza na hora de apresentar o erro para um aprendiz que talvez não conhece aspectos da gramática da língua estrangeira. Adicionalmente, o detalhamento na categorização dos erros pode ser um problema na anotação dos erros em um corpus de aprendizes, pois às vezes é complexo diferenciar uma categoria da outra. O detalhe na categorização dos erros também reflete em uma ferramenta computacional, pois pode provocar desbalanceamentos nas categorias de erros. Concluindo, é importante apontar que a tipologia apresentada por Durão (1999) é mais completa e os erros são menos específicos, mas os erros foram produzidos por alunos aprendizes de português como segunda língua e não como aprendizes estrangeiros.

Córpus Paralelos e Geração Automática de Léxicos Bilíngues

Segundo Alonso et al. (2010), erros lexicais são os mais frequentes na escrita de aprendizes de língua estrangeira. Análises realizadas por Loucky (2013), Aidinlou e Mehr (2012) e Duran (2008a) recomendam a apresentação de léxicos bilíngues com palavras e/ou expressões na língua nativa e na segunda língua, para aprimorar a aprendizagem dos itens incluídos nos recursos paralelos e, como consequência, diminuir o número de erros lexicais. Essa recomendação foi uma das principais motivações para a proposta de geração automática de léxicos bilíngues para auxiliar hispano falantes na escrita em português.

Seguindo essa linha e considerando que a abordagem de tradução automática baseada em probabilidades extraídas de córpus paralelos (*Statistical Machine Translation* (SMT)) se apresenta na literatura como um conjunto de métodos promissores, neste projeto de doutorado foi proposta a criação de recursos lexicais via esses métodos. O sucesso desses métodos estatísticos deve-se, em grande parte, pelo fato deles se beneficiarem do alinhamento de textos paralelos (Caseli, 2003b), tentando extrair o conhecimento de tradução contido em um córpus paralelo alinhado para utilizá-lo na criação de recursos úteis para sistemas de tradução automática (Caseli, 2007).

A Seção 3.1 define córpus paralelos e suas principais utilidades. Além disso, nessa seção são apresentados os métodos de alinhamento dos córpus

paralelos propostos para apoiar tarefas de tradução automática. Também são apresentados os conceitos relacionados com a tradução automática estatística, para depois introduzir os conceitos dos modelos de língua baseado em n-gramas, que são de grande utilidade em diversas aplicações. Depois de introduzidos os aspectos teóricos relacionados com córpus paralelos e tradução automática, na Seção 3.2 são comparados e contrastados trabalhos que utilizam esses métodos para gerar léxicos bilíngues. Dessa forma, serão apresentados os principais procedimentos adotados com essa finalidade e as metodologias propostas para avaliar os léxicos gerados. Finalmente, na Seção 3.3, apresentam-se métodos computacionais e análises linguísticas que se dedicam aos itens lexicais de interesse desta pesquisa: (i) listas de cognatos e falsos cognatos das línguas português e espanhol; (ii) listas de pares de marcadores discursivos; (iii) listas das expressões formulaicas que aparecem nos textos científicos e (iv) listas de verbos relacionados com pesquisa científica em português.

3.1 Métodos Automáticos de Alinhamento e de Tradução

Córpus paralelos são pares de textos acompanhados de sua tradução em uma ou várias línguas, que têm sido utilizados em diversas áreas, especialmente em áreas que estudam as relações existentes entre duas ou mais línguas. Estes recursos são mais úteis, ainda, quando estão alinhados, isto é, quando possuem marcas que identificam pontos de correspondência entre o texto original e sua tradução. Inicialmente, os métodos de alinhamento de textos paralelos surgiram com o intuito de diminuir esforços no desenvolvimento de recursos para tradução automática (Caseli, 2003a; Caseli et al., 2004; Fattah et al., 2007; Tiedemann, 2011), contribuindo para a aquisição automática de recursos linguísticos como: regras de tradução, dicionários bilíngues e listas de palavras.

Embora a tradução automática seja uma das áreas que mais se beneficia dos métodos de alinhamento de textos paralelos (Tiedemann, 2011; Indurkha e Damerau, 2010), esses métodos desempenham um importante papel em diferentes aplicações práticas, tais como: tradução automática (Caseli, 2003a; Caseli et al., 2004; Fattah et al., 2007; Caseli, 2007) e recuperação de informação, lexicografia (monolíngue ou bilíngue) (Langlais et al., 1998; Fung, 1998; Tiu e Roxas, 2008; Li et al., 2011), desambiguação de sentido no nível lexico-

gráfico (Van Der Plas e Tiedemann, 2006; De Melo e Weikum, 2009), aquisição automática de conhecimento (construção de dicionários, extração de terminologias de textos técnicos) (Begoña e Tiedemann, 2006) e aprendizagem e apoio à escrita de língua assistido por computadores (Nerbonne, 2000; Huang et al., 2000; Hayashibe et al., 2012; Dai et al., 2014).

Devido à importância dos *corpus* paralelos, diversas abordagens têm sido propostas para automaticamente extrair da Web esse tipo de recursos. Aziz e Specia (2011) propuseram um método simples que rastreou o site da Revista Pesquisa Fapesp para compilar pares de *corpus* inglês português e português espanhol. Sridhar et al. (2011) e Barbosa et al. (2012) propuseram abordagens para rastrear a Web com a finalidade de identificar sites com informações bilíngues (Sridhar et al., 2011) e multilíngues (Barbosa et al., 2012). Para isso, Sridhar et al. (2011) e Barbosa et al. (2012) combinaram técnicas de recuperação de textos e programação dinâmica e utilizando a estrutura das páginas HTML e os enlaces compilaram *corpus* paralelos. Sridhar et al. (2011), Barbosa et al. (2012) e Aziz e Specia (2011) testaram os *corpus* construídos automaticamente em tarefas de tradução automática, mostrando ótimos resultados nos três estudos. Esse tipo de abordagem é muito útil, porém quando o *corpus* paralelo precisa ser de um gênero específico, por exemplo, científico, é necessário avaliar se os textos incluídos no recurso pertencem ou não a esse gênero.

Produto do sucesso dos métodos de alinhamento de textos paralelos em aplicações de tradução automática, uma grande variedade de técnicas de alinhamento de textos tem sido proposta. As abordagens existentes incluem desde técnicas simples baseadas na contagem de palavras ou caracteres até métodos mais sofisticados que envolvem dados linguísticos, difíceis de automatizar (Indurkha e Damerau, 2010). Existem técnicas que fazem uso de características morfológicas, enquanto outras empregam palavras cognatas ou empréstimos linguísticos. O tipo de alinhamento é crucial para o sucesso do algoritmo de alinhamento (Tiedemann, 2011). De forma geral, o processo de alinhamento de textos paralelos consiste na identificação da correspondência entre o texto original e sua tradução (Caseli, 2003a). Essas correspondências podem se dar em diferentes níveis: do nível do documento completo ao nível de suas partes componentes (capítulos, seções, parágrafos, sentenças, constituintes sintáticos, colocações, palavras e finalmente caracteres) (Caseli, 2003a; Indurkha e Damerau, 2010; Tiedemann, 2011). Os alinhamentos mais estudados na atualidade são os sentencias e lexicais (Caseli, 2003b).

Basicamente, os algoritmos de alinhamento automático recebem como entrada dois textos paralelos, os quais são segmentados a partir da granularidade selecionada. A saída deste processo é um mapa dos textos paralelos que mostram as ligações entre os elementos dos textos individuais, por exemplo, sentenças ou palavras. Na Figura 3.1 mostram-se dois textos paralelos – em espanhol e em português – que foram alinhados sentencialmente. Esse alinhamento foi obtido com a ferramenta visual *VisualTCA*¹. A representação visual da Figura 3.1 representa os diferentes alinhamentos entre as sentenças das duas línguas. Por exemplo, a cor verde representa que uma sentença em espanhol foi alinhada com outra sentença em português. A cor amarela mostra que uma sentença em espanhol foi alinhada com duas sentenças em português e a rosa mostra que duas sentenças em espanhol foram alinhadas com uma sentença em português.

Figura 3.1: Alinhamento sentencial dos pares de resumos em português e espanhol, selecionados da Revista Latino Americana de Enfermagem.

Uma representação visual do alinhamento lexical do mesmo par de resumos é apresentada na Figura 3.2. Nessa Figura, observam-se alinhamentos de vários tipos, assim como alguns erros no alinhamento, que mostra a dificuldade da tarefa. Alguns dos erros nos alinhamentos lexicais aparecem sublinhados na Figura 3.2. Por exemplo, o algoritmo não conseguiu alinhar o verbo "disfarçar" do português com o verbo "esconder" do espanhol. Também falhou o alinhamento entre os verbos e suas preposições "refletir acerca", do português e "reflexionar sobre", do espanhol. Por outro lado, o alinhador conseguiu alinhar a palavra "mídia", com as palavras "medios de comunicación",

¹<http://www.lalic.dc.ufscar.br/portal/>

sendo possível mostrar como um substantivo composto em espanhol pode ser transformado para um simples em português. O algoritmo também alinhou as palavras "papel", em português e "rol" em espanhol. Esse último alinhamento é bastante interessante, dado que a palavra "rol" é um falso cognato das duas línguas.

Figura 3.2: Alinhamento lexical de pares de resumos em português e espanhol, selecionados da Revista Latino Americana de Enfermagem.

A tarefa principal do algoritmo de alinhamento é aperfeiçoar o mapeamento considerando uma função de correspondência (Caseli, 2003a). Para garantir o sucesso dos alinhamentos é comum a combinação de diversos métodos. Na Seção 3.1.1 são apresentados alguns métodos utilizados para alinhar sentencialmente textos paralelos, enquanto que na Seção 3.1.2 são comentados métodos para realizar alinhamentos lexicais.

3.1.1 Métodos de Alinhamento Sentencial

Diversos métodos têm sido desenvolvidos para alinhamento sentencial de textos paralelos. Os trabalhos apresentados por Gale e Church (1991), Gale e Church (1993), Brown et al. (1991) e Kay e Röscheisen (1993), todos citados por Indurkha e Damerau (2010), Tiedemann (2011) e Fattah et al. (2007), constituem a base para a maioria dos métodos dedicados ao alinhamento sentencial de textos paralelos.

Uma das principais características dos métodos de alinhamento sentencial está relacionada às categorias de alinhamentos que eles consideram. A categoria mais comum é a 1-1 na qual uma sentença fonte corresponde exa-

tamente a uma sentença alvo; porém outras categorias são possíveis, como: 0-1, 1-0, 1-2, 2-1, 2-2, etc.. A maioria dos métodos de alinhamento sentencial considera seis casos: eliminação (1-0), inserção (0-1), substituição (1-1), expansão (1-2), contração (2-1) e união ($n - m$ com $n, m > 1$) (Caseli, 2003a).

Segundo Caseli (2003a) e Fattah et al. (2007), os métodos de alinhamento sentencial propostos na literatura podem ser classificados como: métodos empíricos, métodos linguísticos ou métodos híbridos. Os métodos empíricos são aqueles que em seu processo de alinhamento não empregam nenhum tipo de informação linguística; eles utilizam informações estatísticas, tais como a frequência de ocorrência das palavras ou a distribuição delas no texto e técnicas de reconhecimento de padrões que consideram como possíveis traduções mútuas duas palavras de grafias similares. Os métodos linguísticos utilizam informações específicas sobre as línguas envolvidas no processo de alinhamento, como léxicos, listas de palavras âncoras, glossários e etiquetamento morfossintático. Por último, os métodos híbridos, são os mais populares na literatura dado que eles conseguem agrupar a simplicidade dos métodos empíricos com a precisão oferecida pelos métodos linguísticos.

Dentre os métodos empíricos estão os estudos propostos por Gale e Church (1991), Gale e Church (1993) e Brown et al. (1991) que realizaram alinhamentos sentenciais considerando o tamanho das sentenças. A hipótese desses trabalhos é que sentenças grandes em uma língua teriam traduções grandes em outra, enquanto sentenças pequenas teriam traduções pequenas. A principal diferença desses métodos está na forma em que é calculado o tamanho das sentenças. Gale e Church (1991, 1993) calculam o tamanho da sentença em função do número de caracteres e Brown et al. (1991) faz o cálculo considerando o número de palavras presentes nas sentenças.

Um exemplo da abordagem linguística foi apresentado por Simard et al. (1992) que alinhou as sentenças paralelas considerando as palavras cognatas presentes nelas. Esse método obtém um alinhamento lexical grosseiro e é útil para línguas próximas, por exemplo, português e espanhol. Outros exemplos de métodos linguísticos, foram apresentados por Papageorgiou et al. (1994) e Piperidis et al. (2000) que consideraram a quantidade de substantivos, verbos, adjetivos e advérbios presentes nas sentenças para obter os alinhamentos sentenciais.

Os métodos híbridos combinam as duas abordagens anteriores em um único modelo para obter o alinhamento sentencial. Os métodos mais difundidos desta abordagem são: *Geometric Mapping and Alignment* (GMA) (Melamed,

2000), também conhecido como GSA+ e o método *Translation Corpus Alinger* (TCA) (Hoflan, 1996).

O método GSA+ é uma extensão do método GMA que emprega uma lista de cognatos e de palavras âncoras como critérios de alinhamento sentencial. O método TCA utiliza nomes próprios, caracteres especiais, listas de palavras âncoras e o tamanho da sentença para realizar o alinhamento sentencial. Uma implementação do método TCA, chamada de *TCAalign* foi proposta por Caseli e Nunes (2003) e baseia-se em diferentes critérios de alinhamentos para estabelecer as correspondências sentenciais entre dois textos paralelos, tais como listas bilíngues de palavras âncoras, palavras com letra inicial maiúscula (possíveis nomes próprios), caracteres especiais, como é o caso de pontos de exclamação, listas de palavras cognatas e comprimento das sentenças (Caseli et al., 2009). A ferramenta visual *VisualTCA* é uma implementação do método *TCAalign*.

Aziz et al. (2008b) utilizou o algoritmo *TCAalign* para alinhar sentenças entre duas línguas próximas: o português do Brasil e o espanhol Europeu. Os experimentos realizados em Aziz et al. (2008b) mostraram que a abordagem estatística oferece alinhamentos satisfatórios sem necessidade de considerar grandes esforços para seu desenvolvimento. Além disso, Aziz et al. (2008b) consideram que o processo de alinhamento entre duas línguas próximas é mais simples do que entre duas línguas mais distantes. Os experimentos realizados em Aziz et al. (2008b) mostraram que o método de alinhamento sentencial alinhou incorretamente 9,43% das sentenças.

3.1.2 Métodos de Alinhamento Lexical

O alinhamento lexical de palavras e de multipalavras é um processo crítico que desempenha um papel importante em diferentes aplicações de processamento de línguas naturais, tais como tradução automática baseada em exemplos (Caseli, 2007), tradução automática estatística (Deng e Byrne, 2005; Aziz et al., 2008b,a), lexicografia bilíngue (Fung, 1998; Tiu e Roxas, 2008; Li et al., 2011), desambiguação de sentido (De Melo e Weikum, 2009), entre outras.

A principal diferença entre os métodos de alinhamento sentencial e lexical é que na maioria das línguas não é razoável assumir um alinhamento lexical monotônico², além de ser menos provável achar correspondências do tipo 1-1 entre as palavras dos textos. Desta forma, ao contrário do alinhamento

²Os elementos alinhados acontecem na mesma ordem nos textos fonte e alvo (Indurkha e Damerau, 2010).

sentencial, no alinhamento lexical a segmentação das palavras é mais difusa, sendo possível aceitar mapeamentos alternativos entre várias palavras e até mesmo aceitar alinhamentos nulos, ou seja, dizer que palavras ficaram sem correspondência. Esses fatores contribuem para aumentar a complexidade das técnicas de alinhamento lexical (Aziz et al., 2008a; Tiedemann, 2011). Por isso, vários autores (Brown et al., 1988, 1993; Langlais et al., 1998; Begoña e Tiedemann, 2006; Van Der Plas e Tiedemann, 2006; Tiedemann, 2011) recomendam aplicar primeiramente uma técnica de alinhamento sentencial para depois realizar o alinhamento lexical entre os textos.

Em geral, os métodos de alinhamento lexical dependem das aplicações em que serão utilizados. Em alguns casos, é útil encontrar um mapeamento completo entre todas as palavras de um texto, aceitando determinados ruídos³ e em outros é preferível um alinhamento parcial, em que é necessário alinhar com sucesso segmentos de diferentes comprimentos (Tiedemann, 2011).

Os métodos de alinhamento lexical são divididos de acordo com a técnica que utilizam para encontrar as ligações entre as palavras. A primeira abordagem de alinhamento lexical é baseada no modelo gerativo e foi originada nas pesquisas dedicadas à tradução automática estatística. A segunda abordagem é conhecida como modelo discriminativo. Nesta pesquisa de doutorado, métodos de tradução automática estatística foram utilizados para obter léxicos bilíngues, por isso, é importante compreender os aspectos relacionados com o alinhamento lexical com a abordagem gerativa. Quanto ao modelo discriminativo, ele consiste em maximizar uma função que mapeia objetos de um conjunto de entrada em objetos de um conjunto de saída.

No alinhamento lexical, para um par de sentenças paralelas (s, t) , busca-se o alinhamento $\hat{a}$ de forma que:

$$\hat{a} = \arg \max_a \sum_{i=1}^n \lambda_i f_i(a, s, t) \quad (3.1)$$

em que: s^4 é a sentença na língua fonte, t a sentença na língua alvo, f_i são os atributos e λ_i os pesos. Para isso, usualmente empregam-se atributos e métodos de aprendizado de máquina treinados com exemplos corretos.

As técnicas de tradução estatística procuram construir um modelo estocástico capaz de traduzir de uma língua natural (fonte) s para outra língua

³Existem elementos em alguma parte dos textos para os quais não houve correspondência.

⁴Na literatura em inglês de tradução automática é comum denotar a língua fonte com a letra s (*source*) e utilizar a letra t (*target*) para identificar a língua alvo. Neste documento vamos adotar a mesma notação.

(alvo) t . Esse processo de tradução estatística é modelado como um problema de busca, no qual a partir de uma sentença de entrada é necessário determinar suas traduções mais prováveis. Dessa forma, o modelo de alinhamento gerativo considera que os dados foram gerados a partir de uma distribuição de probabilidade $P(t|s)$ e o problema seria estimar a probabilidade de uma sentença alvo t dada uma sentença fonte s . O modelo de tradução automática será detalhado na Seção 3.1.3.

Um modelo gerativo $P(t|s)$, responsável por traduzir uma sentença fonte s para uma sentença alvo t , deve representar as relações básicas entre as palavras que compõem cada sentença. Uma das alternativas propostas para caracterizar essas relações é incluir no modelo o alinhamento entre as palavras de cada sentença. Para isso, é necessário incorporar uma nova variável a , que represente os alinhamentos entre os *tokens* das sentenças, obtendo o seguinte modelo $P(t, a|s)$.

Como o alinhamento entre os *tokens* não é uma variável conhecida, é necessário considerar todas as possíveis alternativas de alinhamento e suas probabilidades. Após isso, é razoável somar todos os alinhamentos possíveis e suas probabilidades como se mostra na Equação 3.2 e inferir o modelo de tradução automático.

$$P(t|s) = \sum_a P(t, a|s) \quad (3.2)$$

A seguir, são explicados os métodos mais conhecidos de alinhamento lexical. Todos eles recebem como entrada principal pares de sentenças em duas línguas diferentes. Os primeiros modelos de alinhamento lexical baseados no modelo gerativo foram propostos por Brown et al. (1988, 1993) para apoiar a tarefa de tradução automática estatística. Em geral, esses modelos de alinhamento lexical são caracterizados a partir de diferentes critérios. Primeiramente, é preciso estabelecer um critério para segmentar as palavras, também conhecido como *tokenizador*, considerando que a sentença fonte é composta por N *tokens* ($s_1, \dots, s_N$) e a sentença alvo por M *tokens* ($t_1, \dots, t_M$).

Depois de tokenizar as sentenças paralelas, o modelo de alinhamento lexical tem que especificar os tipos de alinhamentos entre as palavras das sentenças fonte e alvo. Quanto maior for a diversidade dos tipos de alinhamentos do modelo, os alinhamentos obtidos estarão mais próximos do alinhamento real. Por exemplo, um modelo que permite alinhamentos entre *tokens* individuais é bom, já um modelo que, adicionalmente, permita alinhamentos entre um *token* da língua fonte e vários da língua alvo é melhor ainda e, finalmente,

se também permite alinhar palavras que aparecem em posições diferentes o modelo é mais próximo de um alinhamento real.

Os modelos propostos em Brown et al. (1988, 1993) são baseados na ideia de que palavras em uma língua foram geradas a partir de palavras em outra, então definem uma função de alinhamento $a : m \rightarrow n$, para formalizar o mapeamento entre as palavras de uma língua fonte e alvo. Pode ser observado que a função a permite a correspondência entre um *token* s_n na sentença fonte e outro t_m na sentença alvo, sendo possível representar o alinhamento entre duas sentenças através de um vetor $s : a = a_{1M} = a_1, \dots, a_M$, em que $0 \leq a_m \leq N$ para todo $a_m \in a$. Desta forma, o vetor mostra a posição de cada *token* fonte na sentença e permite representar o valor 0 referente a um *token* que não foi alinhado.

Finalmente, a última consideração relevante em modelos de alinhamento lexical está relacionada com o processo de treinamento. Nesse caso, é crucial determinar o conjunto de parâmetros que melhor vai descrever o alinhamento entre as palavras. Embora o processo de alinhamento consista unicamente em procurar as correspondências entre as palavras individuais dentro das sentenças, esse processo é bem mais complexo dado que os alinhamentos entre as palavras não dependem somente das palavras, também dependem do contexto em que elas estão inseridas. Consequentemente, o modelo de alinhamento tem que construir uma estrutura complexa capaz de mapear uma sentença na outra, respeitando o contexto, da mesma forma que é necessário que a estrutura possa ser descomposta de forma que facilite o treinamento do modelo, permitindo realizar inferências estatísticas. Dessa forma, o modelo estatístico é decomposto em vários tipos de probabilidades, como se observa na Equação 3.3.

$$P(t, a|s) = P(M|s) \prod_{m=1}^M P(a_m|t_1^{m-1}, a_1^{m-1}, s_1^N) P(t_m|t_1^{m-1}, a_1^m, s_1^N) \quad (3.3)$$

Como resultado da decomposição, as probabilidades a seguir são obtidas:

Probabilidade de comprimento: $P(M|s)$, é a probabilidade de uma tradução de comprimento M dada a sentença fonte s.

Probabilidade de alinhamento: $P(a_m|t_1^{m-1}, a_1^{m-1}, s_1^N)$, é a probabilidade de um alinhamento na posição a_m de um *token* na sentença alvo, t^m , dado todos os *tokens* anteriores a t^m , com seus alinhamentos e dada a sentença $s = S_1^N$, na língua fonte.

Probabilidade lexical: $P(t_m|t_1^{m-1}, a_1^m, s_1^N)$, é a probabilidade de um *token* t_m ,

na sentença alvo dado seu alinhamento, dados todos os *tokens* prévios a ele na sentença alvo, com seus respectivos alinhamentos e dada a sentença fonte s_1^N .

Os modelos propostos por Brown et al. (1988, 1993) utilizam o modelo gerativo como ponto de partida e são conhecidos como modelos IBM 1-5. Cada um desses modelos se diferencia basicamente nos parâmetros empregados para estabelecer o alinhamento lexical e seus níveis de dificuldades vão aumentando da mesma forma que aumenta a sucesso dos algoritmos.

O modelo IBM 1 é o mais simples e considera total independência entre as palavras e seus alinhamentos. É baseado somente na probabilidade de tradução dos léxicos $p(t|s)$, a probabilidade de comprimento não é considerada no treinamento do modelo e a probabilidade de alinhamento assume total independência dos *tokens* individuais e seus alinhamentos. Considera-se que os possíveis alinhamentos dos *tokens* estão distribuídos uniformemente.

O modelo IBM 2 é uma extensão do modelo IBM 1, que adiciona no modelo a probabilidade de alinhamento dependendo das posições absolutas dos *tokens*, $P(a_m|t_1^{m-1}, a_1^{m-1}, s_1^N) = P(a_m|m, N, M)$. Esse modelo é dividido em dois procedimentos: (i) tradução lexical e (ii) probabilidade de alinhamento. Ele continua assumindo independência entre os parâmetros. Uma das principais limitações desse modelo é que não consegue modelar alinhamentos do tipo $1 - N$ entre as palavras da sentença fonte e alvo.

Como solução ao problema anterior, o modelo IBM 3 adiciona um novo parâmetro ao modelo de alinhamento $p_n(\phi|s)$, chamado de fertilidade. Esse parâmetro se refere ao número $\phi = 0, 1, 2, \dots$, de *tokens* na sentença alvo que podem ser gerados dado um *token* s . Dessa forma, é possível modelar alinhamentos de um *token* para múltiplos *tokens*, além de permitir a eliminação de *tokens* que não tiveram nenhuma correspondência. O IBM 3 também usa uma nova classe de parâmetros $P(m|n, N, M)$, chamada de distorção, a qual prediz a posição absoluta m do *token* alvo, dada a posição n do *token* fonte que o gerou. Embora o modelo IBM 3 consiga modelar muitas das propriedades de correspondência entre *tokens*, ele ainda assume independência entre os parâmetros do modelo.

O modelo IBM 4 introduz algumas extensões do parâmetro de fertilidade apresentado no modelo IBM 3. Primeiramente, considera o parâmetro de distorção baseando-se na localização relativa do *token* alvo e não na posição absoluta, como acontecia no modelo IBM 3. Assim, a localização da palavra alvo é baseada na localização das palavras geradas anteriormente. Esse mo-

delo assume pela primeira vez uma dependência de primeira ordem entre os alinhamentos, tentando capturar a tendência de que trechos de uma sentença fonte são movimentados conjuntamente na sentença alvo. Nesse modelo, também, é introduzido o conceito de funções das palavras, avaliando as possíveis localizações das palavras a partir da função que desempenha, por exemplo, adjetivos, substantivos, etc.. Assim, além de considerar dependência entre os alinhamentos o modelo assume dependências lexicais.

O modelo IBM 5 tem, basicamente, as mesmas propriedades que o modelo IBM 4, sendo que a principal diferença é ser o primeiro a acrescentar parâmetros no modelo de alinhamento para determinar alinhamentos que não são permitidos. Segundo Tiedemann (2011), esse modelo torna o processo de treinamento mais complexo e usualmente não apresenta melhoras muito significativas no alinhamento se comparado com o modelo IBM 4. Por isso, o modelo IBM 4 é considerado um modelo adequado para tarefas de tradução automática.

Embora o modelo IBM 4 seja um modelo que consegue generalizar o processo de alinhamento, algumas deficiências são identificadas. A formulação do modelo é tão complexa que dificulta a estimativa dos parâmetros empregando o método de Máxima Verossimilhança. Nesse sentido, o modelo de alinhamento é aproximado, aplicando o método de busca *hill-climbing*. Consequentemente, a estimativa dos parâmetros é um processo dificilmente paralelizável, lento e que requer alto consumo de memória. Além disso, o cálculo de estatísticas a partir do modelo IBM 4 é complexo, o qual é uma limitação para o uso desse modelo, em outras tarefas de modelagem diferentes do alinhamento lexical. Como alternativa para solucionar alguns dos inconvenientes, é apresentado o modelo oculto de *Markov* como um modelo de alinhamento lexical (Deng e Byrne, 2005; Tiedemann, 2011).

O modelo oculto de *Markov* (HMM) foi proposto por Vogel et al. (1996) e comentado em Deng e Byrne (2005) e Tiedemann (2011). O modelo HMM usa os mesmos parâmetros do modelo IBM 2, mas inclui dependências de primeira ordem p_a entre os alinhamentos.

$$p_{hmm}(a_m | t_1^{m-1}, a_1^{m-1}, s_1^N) = p_a(a_m | a_{m-1}, N) \quad (3.4)$$

Nesse modelo, a posição de alinhamento a_m da palavra alvo t_m depende da posição de alinhamento a_{m-1} , da palavra alvo anterior, em vez de depender somente da posição m da palavra na sentença alvo, como acontecia no modelo 2. Essa é uma ideia similar à discutida no modelo IBM 4, que parte da premissa

que os *tokens* tentam se movimentar em conjunto com outros. O treinamento do modelo HMM é mais simples, empregando uma variação do algoritmo Maximização da Expectância, isso devido a algumas propriedades que o modelo apresenta. O alinhamento obtido com o algoritmo HMM é efetivo e comumente é utilizado como um alinhamento *pipeline* entre os modelos 2 e 3. Além disso, é possível adicionar no modelo HMM as dependências lexicais, como proposto no modelo 4.

Embora diversas ferramentas tenham sido disponibilizadas na literatura para realizar o alinhamento lexical de sentenças paralelas, o *framework* Giza++ (Och e Ney, 2000b) é o mais utilizado nas pesquisas de alinhamento lexical e faz parte do estado da arte deste tema. O Giza++ utiliza os modelos IBM e o modelo HMM para identificar o melhor alinhamento entre *tokens* de duas línguas diferentes. Na ferramenta Giza++ versão 2.0 estão incluídas iterações dos modelos IBM-1, IBM-3, IBM-4 e HMM. Gao e Vogel (2008) apresentam o algoritmo MGiza++ que paraleliza o processo de alinhamento da ferramenta Giza++. Esse algoritmo paralelo foi criado com o intuito de conseguir melhorar o desempenho dos alinhamentos de grandes volumes de textos. Os alinhamentos obtidos com MGiza++ se mantiveram com a mesma qualidade que os obtidos com a ferramenta Giza++, mostrando um significativo aumento na velocidade do processamento. Caseli et al. (2009) utilizou o *framework* Giza++ para obter alinhamentos entre o espanhol e o português, conseguindo uma precisão de 93,70% para esses par de língua.

3.1.3 Tradução Automática

Na tradução automática estatística usa-se uma abordagem que cria um modelo probabilístico para obter a melhor tradução de uma sentença fonte S para uma sentença alvo T , mantendo o sentido da sentença alvo (fidelidade) e gerando uma sentença aceita na língua fonte (fluência) (Jurafsky e Martin, 2009).

Para introduzir os conceitos da tarefa de tradução automática se assume que a tradução acontece de uma sentença da língua estrangeira $F = f_1, f_2, \dots, f_m$, que pode ser o espanhol, para uma sentença $E = e_1, e_2, \dots, e_n$ da língua alvo, que pode ser o português. Nesta seção, emprega-se a mesma notação adotada na literatura de tradução automática que utiliza a letra F para denotar a língua estrangeira e E para denotar a outra língua. No modelo probabilístico, a melhor tradução $E = e_1, e_2, \dots, e_n$ é aquela com maior probabilidade $P(E|F)$. Dessa forma, o modelo estatístico de tradução é semelhante ao modelo do ca-

nal ruidoso⁵, que pode ser descrito usando o teorema de *Bayes* na equação 3.5.

$$\hat{E} = \operatorname{argmax}_E P(E|F) = \operatorname{argmax}_E \frac{P(F|E)P(E)}{P(F)} = \operatorname{argmax}_E P(F|E)P(E) \quad (3.5)$$

Como a sentença na língua estrangeira é fixa, na Equação 3.5 o denominador $P(F)$ é ignorado, pois seu valor é constante. Desta forma, para identificar a melhor tradução para a língua alvo é necessário o modelo de tradução $P(F|E)$ e um modelo de língua $P(E)$.

$$\hat{E} = \operatorname{argmax}_{E \in \text{português}} \overbrace{P(F|E)}^{\text{modelo de tradução}} \underbrace{P(E)}_{\text{modelo de língua}} \quad (3.6)$$

Para compreender melhor a aplicação do canal ruidoso é importante pensar que a sentença na língua estrangeira é uma sentença da língua alvo distorcida. Assim, a análise que se faz é tentar descobrir qual foi a sentença na língua alvo que gerou a sentença que se deseja traduzir. Na Figura 3.3, mostra-se uma representação do canal ruidoso para visualizar melhor esses conceitos. A Figura 3.3 é uma adaptação da figura apresentada por Jurafsky e Martin (2009). Na Figura 3.3 observam-se os três principais componentes que intervêm no processo de tradução estatística:

- modelo de língua;
- modelo de tradução;
- decodificador, que dada a entrada F , produz a saída E mais provável.

O modelo de língua $P(E)$ é um modelo baseado em n -gramas que será detalhado na Seção 3.1.4, dada a sua importância em várias aplicações do PLN, como, reconhecimento automático de fala, recuperação da informação, sumarização, entre outros cenários de aprendizado de máquina, porém os outros dois componentes são detalhados a seguir.

O objetivo do modelo de tradução é obter a probabilidade de que E gere a F , dada a sentença alvo E e a sentença fonte F . Veja que há uma troca: procura-se na língua alvo a sentença que gerou a sentença na língua estrangeira. Uma

⁵do inglês *noisy channel model*

Figura 3.3: Modelo do canal ruidoso na tradução automática.

alternativa para calcular essa probabilidade é considerar a probabilidade de traduzir cada uma das palavras da expressão, no entanto os modelos de tradução estatística atuais fazem o cálculo considerando as traduções de frases. Esses modelos atuais são conhecidos como modelos de tradução baseados em frases. Devido ao sucesso dessa abordagem, vários modelos têm sido propostos, sendo o modelo criado por Koehn et al. (2003) um dos mais conhecidos.

O modelo de tradução baseado em frases (Koehn et al., 2003) segue o modelo gerativo, apresentado na Seção 3.1.2, e é separado em três etapas. Primeiramente, agrupa as palavras da língua alvo E em frases $\bar{e}_1, \bar{e}_2 \dots \bar{e}_l$. Depois, traduz as frases $\bar{e}_i$ na língua alvo para a língua fonte $\bar{f}_i$. Finalmente, ordena as frases na língua fonte; esse último passo é opcional.

O modelo de tradução baseado em frases é obtido após aplicar a equação 3.7. Esse modelo depende dos dois fatores principais que aparecem na Equação 3.7: a probabilidade de tradução $\Phi(\bar{f}_i, \bar{e}_i)$ e a probabilidade de distorção $d(a_i - b_{i-1})$. A probabilidade de tradução é a probabilidade de que a frase $\bar{f}_i$ seja gerada pela frase $\bar{e}_i$. A probabilidade de distorção é uma medida de distância entre as frases da língua fonte e alvo que estima a probabilidade de que uma frase na língua alvo apareça em uma posição diferente na língua fonte; isso pode acontecer porque as palavras aparecem em ordens diferentes ou porque as frases foram separadas por outras palavras.

$$P(F|E) = \prod_{i=1}^l \Phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1}) \quad (3.7)$$

Para facilitar a compreensão de como se estima o modelo representado na Equação 3.7, utiliza-se o exemplo da Figura 3.4. A Figura 3.4 mostra as posições das palavras de uma sentença em português e como elas foram alinhadas com palavras da sentença em espanhol. Os alinhamentos apresentados na Figura 3.4 foram obtidos com os algoritmos de alinhamento lexical, apresentados na Seção 3.1.2. Para estimar o modelo de tradução é necessário treinar um modelo, que utilize os alinhamentos do exemplo, para realizar as traduções. Para isso, é preciso usar um grande cópulo paralelo com as sentenças alinhadas, por meio do qual o modelo de tradução vai criar uma tabela de tradução que contém as palavras e frases das duas línguas com o cálculo da probabilidade $\Phi(\bar{f}_i, \bar{e}_i)$. Esse cálculo é realizado considerando a quantidade de vezes que os pares de frases ou palavras aconteceram no cópulo paralelo, como mostrado na Equação 3.8.

Posição	1	2	3	4	5		6	7		8
Português	os	meios	de	comunicação	foram		essenciais	na		educação
Espanhol	<i>los</i>	<i>medios</i>	<i>de</i>	<i>comunicación</i>	<i>han</i>	<i>sido</i>	<i>esenciales</i>	<i>en</i>	<i>la</i>	<i>educación</i>

Figura 3.4: Exemplo de alinhamentos entre palavras e frases usados no treinamento do modelo de tradução.

$$\Phi(\bar{f}_i, \bar{e}_i) = \frac{total(\bar{f}, \bar{e})}{\sum_{\bar{f}} total(\bar{f}, \bar{e})} \quad (3.8)$$

Como mencionado, o modelo de distorção também é considerado na escolha da melhor tradução. No exemplo da Figura 3.4, todas as frases e palavras das duas línguas se mantiveram na mesma posição, dessa forma a distância que representa a distorção é 1. Na Equação 3.9, aparece o cálculo do modelo de tradução para o exemplo apresentado na Figura 3.4.

$$\begin{aligned}
P(F|E) = & P(\text{los}|\text{os}) * d(1) * P(\text{medios}|\text{meios}) * d(1) * P(\text{de}|\text{de}) * d(1) * \\
& P(\text{comunicación}|\text{comunicação})d(1)* \\
& P(\text{han sido}|\text{foram})d(1) * P(\text{esenciales}|\text{essenciais})d(1)* \\
& P(\text{en la}|\text{na})d(1) * P(\text{educación}|\text{educação}) * d(1)
\end{aligned} \quad (3.9)$$

O último componente do modelo de tradução estatística é o decodificador. O decodificador tem a função de dada a sentença na língua estrangeira achar

a melhor tradução na língua alvo, considerando o produto do modelo de tradução e do modelo de língua, como indicado na Equação 3.6. O problema anterior é resolvido com um algoritmo de busca que precisa achar a melhor solução. O decodificador do modelo de tradução estatística é um algoritmo de busca informada, baseado na busca primeiro o melhor⁶. Essa técnica de busca utiliza heurísticas relacionadas com o domínio do problema para identificar as melhores alternativas de solucionar os problemas. A implementação do decodificador na tradução estatística é o algoritmo A_* que usa como heurísticas o cálculo da probabilidade do modelo de tradução e a probabilidade do modelo de língua para andar pelos nós. Nesse caso, os nós são as palavras e frases dentro das sentenças; a análise de cada nó é um passo para determinar a melhor sentença E dada a sentença em língua estrangeira F .

Diversas ferramentas (Koehn et al., 2007; Cer et al., 2010) têm sido propostas para simular o processo de tradução estatística baseada em frases. A ferramenta mais conhecida e implementada, principalmente, para apoiar pesquisas acadêmicas é o *toolkit* Moses⁷(Koehn et al., 2007). O *toolkit* Moses contém todos os componentes para processar os dados, treinar modelos de tradução e modelos de língua. Além disso, possui ferramentas para avaliar a saída do modelo de tradução e incorpora ferramentas externas como Giza++ para obter alinhamentos sentenciais e SRILM para criar modelos de língua.

Avaliação da Tradução Automática

A avaliação da saída do método de tradução automática é essencial, porém é uma tarefa subjetiva. Na avaliação deseja-se verificar a qualidade da tradução considerando a fidelidade - o sentido da sentença fonte foi mantido na sentença alvo - e a fluência - aceitação da sentença na língua alvo - (Jurafsky e Martin, 2009).

A avaliação mais precisa é realizada por um humano que avalia se a sentença gerada é aceita ou não na língua alvo. Para isso, é comum considerar níveis de inteligibilidade, que são medidos de diversas formas. A avaliação anterior refere-se à fluência da saída do método de tradução. Já, para avaliar a fidelidade da sentença gerada é comum ter uma ou várias sentenças na língua alvo que foram geradas por humanos. Esse conjunto de sentenças serve como referência para verificar a saída do sistema. A verificação costuma ser realizada por meio de medidas de similaridade, como a distância de *Damerau-*

⁶do inglês *best-first*

⁷<http://www.statmt.org/moses/>

Levenshtein (Levenshtein, 1966; Damerau, 1964a) (ou distância de edição).

A avaliação da saída utilizando humanos é uma tarefa demorada, que requer pessoal treinado. Para amenizar essa dificuldade, os conjuntos de testes gerados por humanos com sentenças na língua fonte e suas traduções na língua alvo são utilizados para, por meio de métricas automáticas, avaliar a saída dos métodos de tradução. Algumas das métricas propostas foram: Precisão, Cobertura, BLUE, NIST, TER e METEOR, que aparecem detalhadas em Jurafsky e Martin (2009).

A precisão e a cobertura são medidas utilizadas em diversos sistemas de PLN e serão empregadas nesta pesquisa para avaliar o desempenho dos métodos propostos para compilar os léxicos bilíngues. Essas medidas são calculadas por meio de mapeamentos entre as entradas candidatas e as entradas criadas como referência, como se mostram nas Equações 3.10 e 3.11. Dessa forma, a precisão representa o número de candidatos corretos considerando o total de itens candidatos, enquanto a cobertura indica o total de itens candidatos corretos em relação ao total de itens incluídos nas referências.

$$\text{precisão} = \frac{\text{candidatos} \in \text{referência}}{\text{candidatos}} \quad (3.10)$$

$$\text{cobertura} = \frac{\text{candidatos} \in \text{referência}}{\text{referência}} \quad (3.11)$$

O cálculo da medida BLUE, que consiste, também, no mapeamento das palavras e n-gramas da saída da tradução com os itens da saída gerada por um humano é baseado no cálculo da precisão.

3.1.4 Modelos de Língua

Um modelo de língua baseado em n-gramas é um modelo probabilístico que prediz a próxima palavra a partir da anterior $N - 1$. Um n-grama é uma sequência de *tokens* de diversos tamanhos, por exemplo 2-grama, também chamado de bigrama, é uma sequência com duas palavras, 3-grama ou tri-grama é uma sequência de três palavras. Estimar a probabilidade da próxima palavra está relacionado com estimar a probabilidade de uma sequência de palavras.

Os modelos de língua têm sido utilizados em diversas aplicações, pois por meio deles é possível reconhecer/gerar as sequências que fazem parte de uma determinada língua. Modelos de língua têm sido utilizados em tarefas de reconhecimento de fala (Oparin, 2008), de identificação de erros (Hermet e Dé-

silets, 2009; Gamon et al., 2009), de etiquetamento morfossintático, de geração de língua natural (Seno, 2010), de similaridade de palavras (Stehouwer e Van Zaanen, 2009), de reconhecimento de autoria (Johnson et al., 2006), de extração de sentimentos (Hu et al., 2007) e de predição textual (Trnka, 2008). Como comentado na Seção 3.1.3, modelos de língua baseados em n-gramas são uma parte essencial da tradução automática estatística.

Os modelos de língua utilizam *córpus* para computar a probabilidade de uma palavra w considerando um histórico h , $P(w|h)$. Supondo que o histórico é a sequência "os meios de comunicação", deseja-se calcular a probabilidade de que a palavra que segue a sequência anterior é "foram". Para calcular a $P(\text{foram}|\text{os meios de comunicação})$ uma alternativa é quantificar a quantidade de vezes que a expressão "os meios de comunicação foram" acontece no *córpus* dividido pelo total de vezes que somente acontece a expressão "os meios de comunicação". No entanto, o cálculo anterior não é uma alternativa adequada, pois a língua é muito variada e a probabilidade de que aconteça a sequência completa ou possíveis variações é baixa.

Uma alternativa mais inteligente de calcular a probabilidade das sequências de palavras é considerando a regra da cadeia. Formalizando algumas notações: a probabilidade de uma variável aleatória X_i que toma o valor "foram" é representada como $P(X_i = \text{foram})$ ou $P(\text{foram})$. Uma sequência N de palavras é representada como $w_1...w_n$ ou w_1^n e o valor da probabilidade conjunta da sequência de palavras é $P(w_1, w_2, ..., w_n)$. Para calcular a probabilidade da sequência $P(w_1, w_2, ..., w_n)$ a probabilidade é decomposta segundo a regra da cadeia:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^{n-1}) = \prod_{k=1}^n P(w_k|w_1^{k-1}) \quad (3.12)$$

A Equação 3.12 mostra como é possível calcular a probabilidade conjunta de toda a sequência, a partir das probabilidades condicionais. Porém, também não há forma de calcular a probabilidade de uma sequência muito grande de palavras, $\prod_{k=1}^n P(w_k|w_1^{k-1})$. Considerando a hipótese de *Markov* é possível realizar o cálculo anterior, pois a hipótese diz que a probabilidade de uma palavra depende, somente, da palavra anterior. Os modelos de *Markov* são uma classe de modelos probabilísticos que pressupõem que para prever a probabilidade de uma unidade é suficiente considerar o histórico próximo, sem necessidade de analisar históricos muito antigos. Para compreender melhor o significado

da hipótese de *Markov* no cálculo da probabilidade dos n-gramas no modelo de língua, retoma-se o cálculo $P(\text{foram}|\text{os meios de comunicação})$. A hipótese diz que é possível fazer uma aproximação como a apresentada na Equação 3.13:

$$P(\text{foram}|\text{os meios de comunicação}) \approx P(\text{foram}|\text{comunicação}) \quad (3.13)$$

A Equação 3.14 é uma generalização da aproximação apresentada na equação 3.13.

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1}) \quad (3.14)$$

Dada a aproximação, é possível calcular a probabilidade da sequência completa, substituindo a Equação 3.14 na Equação 3.12.

$$P(w_1^n) \approx \prod_{k=1}^n P(w_k|w_{k-1}) \quad (3.15)$$

A estimativa da probabilidade obtida com a Equação 3.15 é calculada por meio da estimativa da máxima verossimilhança (MLE). A MLE consiste em contar a quantidade de vezes que os n-gramas acontecem no *córpus* utilizado para criar o modelo de língua. Por exemplo, para computar a probabilidade de um bigrama com a palavra y dada a palavra anterior x , realiza-se a contagem do bigrama $C(xy)$ no *córpus* e a soma é normalizada dividindo-se pelo total de bigramas que começam com a palavra x , como indicado na Equação 3.16. Veja na Equação 3.16 que o denominador é a soma da quantidade de vezes que acontece a palavra x no *córpus*. O cálculo da Equação 3.16 é chamado de cálculo da frequência relativa e é usado para calcular a probabilidade de n-gramas de qualquer tamanho.

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})} \quad (3.16)$$

Como comentado, os cálculos de probabilidade dos n-gramas que integram um modelo de língua são realizados a partir de um *córpus*, utilizado como *córpus* de treinamento do modelo. Além do treinamento, é comum utilizar outro *córpus* chamado de *córpus* de teste para testar o desempenho do modelo de língua. Evidentemente, o *córpus* de treinamento é diferente do *córpus* de teste. O modelo de língua, como qualquer outro modelo estatístico, é dependente do conjunto de treinamento, em outras palavras, as probabilidades obtidas no

conjunto de treinamento podem ser diferentes em outros conjuntos. Além disso, quanto maior for o tamanho do n-grama melhor será o desempenho do modelo de língua (Jurafsky e Martin, 2009).

A melhor forma de avaliar um modelo de língua é de forma extrínseca, isto é, inserindo o modelo em uma tarefa específica, em função do desempenho da tarefa é avaliado o modelo. Por outro lado, os modelos de língua também são avaliados intrinsecamente, por meio do cálculo da perplexidade (PP). A perplexidade é uma medida baseada na estimativa das probabilidades das sentenças do conjunto de teste. Intuitivamente, quanto melhor for a probabilidade da sentença do conjunto de teste, melhor será o desempenho do modelo de língua. Formalmente, a perplexidade é uma função da probabilidade que o modelo de língua deu para o conjunto de teste. Para um conjunto de teste $W = w_1w_2...w_n$ a perplexidade é a probabilidade do conjunto de teste, normalizada pelo número de palavras:

$$PP(W) = P(w_1w_2...w_n)^{-\frac{1}{N}} = \sqrt[n]{\frac{1}{P(w_1w_2...w_n)}} \quad (3.17)$$

Analisando a Equação 3.17 observa-se que quanto maior for a probabilidade menor a será perplexidade. Para determinar o melhor modelo de língua, é comum comparar vários modelos criados com diferentes corpú ou com diferentes configurações e indicar o modelo com menor perplexidade.

Como o modelo de língua depende de um corpú de treinamento, um fenômeno comum é encontrar palavras em outros conjuntos que não foram incluídas no conjunto de treinamento, essas palavras são chamadas de palavras fora do vocabulário⁸. Para algumas aplicações costuma-se treinar o modelo de língua considerando que não há palavras fora do vocabulário, assumindo que todas as palavras estão incluídas no conjunto de treinamento, porém esse não é um cenário real. Por outro lado, é comum criar modelos de língua que estimam uma probabilidade para as palavras que não apareceram no conjunto de treinamento, considerando a quantidade dessas palavras que aconteceram em um conjunto de testes. Essas palavras são reconhecidas pelo modelo por meio da etiqueta <UNK>.

Como os modelos de língua são utilizados em diversas aplicações, inúmeros estudos têm sido utilizados para aprimorar as estimativas realizadas pelos modelos (Jurafsky e Martin, 2009). Muitos métodos foram criados para aprimorar o cálculo da probabilidade de n-gramas. O cálculo da probabilidade com o mé-

⁸do inglês *out of vocabulary words* (OOV)

todo MLE não é efetivo quando acontecem n-gramas com baixa frequência no corpus de treinamento (Jurafsky e Martin, 2009). Esse fenômeno é bastante usual, pois os n-gramas são esparsos, isto é, alguns acontecem com muita frequência, enquanto outros acontecem uma única vez e são totalmente aceitos na língua. Para tratar o problema anterior, métodos de suavização⁹ têm sido propostos. Os métodos de suavização tentam modificar a probabilidade dos n-gramas que receberam probabilidade zero de forma incorreta. Como o nome indica os métodos de suavização tentam diminuir as variações na área representada pelas probabilidades do modelo de língua, equiparando as probabilidades dos n-gramas mais frequentes com os menos frequentes. Os métodos de suavização mais conhecidos são: *Laplace Smoothing*; *Good-Turing Smoothing*; *Witten-Bell* and *Kneser-Ney smoothing* (Jurafsky e Martin, 2009). *Kneser-Ney smoothing* é o método mais utilizado para calcular as probabilidades considerando a suavização do modelo. Neste capítulo não serão apresentados esses algoritmos, pois somente deseja-se dar noções básicas dos modelos de língua, como eles são estimados e suas aplicações.

Adicionalmente aos métodos para suavizar os valores das probabilidades dos modelos de língua, outra alternativa proposta para aprimorar a estimativa de probabilidade é o emprego da hierarquia dos n-gramas para realizar o cálculo. Essa alternativa consiste na combinação das sub-cadeias de n-gramas para calcular a probabilidade do n-grama maior. Por exemplo, supondo que se deseja calcular a probabilidade $P(w_n|w_{n-2}w_{n-1})$, mas o trigramma $w_{n-2}w_{n-1}w_n$ não acontece no corpus de treinamento, então é possível fazer o cálculo considerando a probabilidade do bigrama $P(w_n|w_{n-1})$. Caso o bigrama não aconteça no corpus, então se usa a probabilidade do unigrama $P(w_n)$. Com esta finalidade, dois modelos foram propostos: (i) interpolação e (ii) *backoff*. Na interpolação, o cálculo da probabilidade do n-grama acontece mediante a combinação linear das probabilidades de todas as sub-cadeias, misturam-se os valores de probabilidade do trigramma, bigrama e unigrama. No modelo *backoff* as probabilidades das sub-cadeias não são combinadas, mas caso não seja possível estimar a probabilidade do trigramma, usa-se a probabilidade do bigrama, caso não exista a probabilidade do bigrama, então usa-se o unigrama.

Um modelo de língua é, geralmente, armazenado em um arquivo no formato ARPA. Esse arquivo é composto pela lista de n-gramas com probabilidade diferente de zero e ordenados pelos tamanhos dos n-gramas, aparecendo primeiro os unigramas, seguido dos bigramas e por último os trigramas. Cada

⁹do inglês *smoothing*

n-grama aparece no arquivo acompanhado do valor da probabilidade calculada segundo algum método de suavização e segundo o método de interpolação ou *backoff* α . Ambos os valores aparecem em função do $\log_{10}$ e os valores de *backoff* somente são calculados para os n-gramas que são sub-cadeias de outros n-gramas. Desta forma, os n-gramas maiores não têm esse valor. A representação de um arquivo ARPA de um modelo de trigramas é apresentada a seguir:

unigrama:	$\log P^*(w_i)$	w_i	$\log \alpha(w_i)$
bigrama:	$\log P^*(w_i w_{i-1})$	$w_{i-1}w_i$	$\log \alpha(w_{i-1}w_i)$
trigrama:	$\log P^*(w_i w_{i-2}, w_{i-1})$	$w_{i-2}w_{i-1}w_i$	

Vários *toolkits* têm sido desenvolvidos para construir modelos de língua. Um dos mais utilizados é o SRLIM (Stolcke, 2002). O *toolkit* está disponível livremente e disponibiliza diversos algoritmos para criar e avaliar o modelo de língua. SRLIM recebe como entrada um texto cru, que contém uma sentença tokenizada por linha. Além disso, recebe outros parâmetros adicionais, tais como: o tamanho do modelo e o tipo de suavização. Ele disponibiliza *Good-Turing* e *Kneser-Key* e outros limiares para realizar a estimativa do modelo. O *toolkit* também permite calcular a perplexidade dado um conjunto de teste considerando um modelo de língua, previamente criado.

3.2 Métodos para Gerar Léxicos Bilíngues

Os léxicos bilíngues são recursos de grande valor para diversas áreas de PLN (Caseli, 2003a; Lee e Swales, 2006; Caseli et al., 2009; De Melo e Weikum, 2009), pois especificam as correspondências entre palavras e/ou expressões entre duas línguas. Esses recursos costumam ser utilizados para aprimorar algoritmos de tradução automática.

Nesta seção, apresenta-se uma definição formal de léxicos bilíngues. Em seguida, são comentados alguns dos métodos utilizados para construir léxicos bilíngues. Devido ao fato dos recursos lexicais propostos nesta pesquisa conterem palavras e expressões, nesta seção também são apresentados os métodos da área de geração de léxicos bilíngues de multipalavras.

Um léxico bilíngue para as línguas S e T pode ser formalmente definido como um subconjunto do produto cartesiano das palavras de S e das palavras de T . Cada entrada X do léxico bilíngue é um par (w_s, w_t) , em que $w_s \in S$ e $w_t \in T$ (Melamed, 1996).

Em alguns métodos de geração de léxicos bilíngues, as entradas são acompanhadas de um valor que indica o nível de associação entre os léxicos de ambas as línguas, por exemplo, a probabilidade ou medidas de associação, caso o léxico tenha entradas multipalavras. Quanto maior for o valor, melhor a tradução entre os elementos de entrada. A construção desses recursos linguísticos adaptados para vários domínios é custoso, por isso os pesquisadores tentam automatizar a tarefa.

Córpus paralelos alinhados têm sido muito utilizados para a extração de léxicos bilíngues porque representam melhor as correlações específicas entre os contextos dos textos de línguas diferentes, o que contribui para a eliminação de possíveis ambiguidades (Fung, 1998; Tiu e Roxas, 2008; De Melo e Weikum, 2009; Caseli et al., 2009). Outros autores têm utilizado os corpúscos comparáveis¹⁰ nessa tarefa, pois são recursos mais abundantes e de fácil acesso na internet (Gaussier et al., 2004; Li et al., 2011).

Diversos métodos apresentam-se na literatura para compilar léxicos bilíngues. Alguns métodos procuram correspondências bilíngues utilizando estatísticas, cognatos ou métricas de similaridade nos textos paralelos. Outros métodos realizaram modificações do algoritmo proposto por Brown et al. (1993) para obter alinhamentos.

Wu e Xia (1994) combinaram modificações do algoritmo de Brown et al. (1993) com um procedimento iterativo de maximização da expectância para extrair um léxico bilíngue inglês-chinês de um grande corpúscos paralelo. Fung (1995) criou um léxico bilíngue de substantivos, sintagmas nominais e nomes próprios tratando o problema de compilação como um problema de casamento de padrão, em que as palavras da língua fonte e alvo compartilhavam algumas características comuns, tais como: as posições no corpúscos e a tendência de formarem segmentos alinhados. Koehn e Knight (2002) combinaram várias heurísticas e usando corpúscos comparáveis criaram um léxico bilíngue alemão-ínglês. As entradas do léxico foram obtidas a partir das seguintes heurísticas: (i) palavras idênticas ou que diferem em uma única letra; (ii) similaridade ortográfica; (iii) contextos similares e (iv) frequência das palavras.

Em geral, as entradas dos léxicos influenciam diretamente no método proposto para construir o léxico bilíngue. Por exemplo, usando medidas de similaridade ortográfica é possível obter palavras cognatas entre duas línguas, método que é mais eficiente se as línguas são próximas. Já, o método anterior

¹⁰Um corpúscos comparável é composto por uma coleção de documentos que pertencem a um mesmo domínio ou assunto, mas que foram escritos por diferentes autores e em diferentes datas (Fung, 1998)

não é válido para obter um léxico com palavras que ortograficamente não são similares. Por esse motivo, abordagens de tradução automática se mostram mais eficientes para compilar os léxicos bilíngues. A escolha da abordagem de tradução também depende das palavras e expressões incluídas na análise. Por exemplo, diversas abordagens podem ser utilizadas para gerar léxicos com palavras ou com multipalavras. Quando o léxico contém multipalavras é comum combinar métodos de tradução estatística com medidas de associação para extrair os léxicos e depois obter as equivalências entre as duas línguas.

3.2.1 Métodos para Criar Léxicos Bilíngues compostos por Multipalavras

As multipalavras¹¹ são muito requisitadas em diversas aplicações de PLN e nesse grupo estão incluídas palavras compostas, frases pré-fabricadas, colocações e expressões idiomáticas. Em aplicações dedicadas à aquisição de uma segunda língua ou língua estrangeira o conhecimento desse grupo é especialmente importante, pois segundo Lewis (1993) o estudo dessas expressões aprimora a aprendizagem da segunda língua. Esse fato incentivou o estudo de métodos para gerar léxicos bilíngues com multipalavras.

Sag et al. (2002) definem as multipalavras como combinações de palavras, para as quais as propriedades sintáticas ou semânticas de toda a expressão não podem ser obtidas a partir de seus componentes individuais. Segundo Rasmisch (2012) a ambiguidade na definição de multipalavras torna o tratamento automático dessas palavras um problema em aberto, de forma que ainda não há uma abordagem que seja unânime para tratar automaticamente as expressões multipalavras. No entanto, a tarefa anterior foi estruturada pelos organizadores do *MWE workshop*, 2009 pelas seguintes etapas:

- **Identificação:** A identificação de expressões multipalavras é um desafio, devido à variabilidade dessas expressões, sendo assim não é suficiente coletar e utilizar uma lista fixa de multipalavras, é preciso criar regras complexas e algoritmos de aprendizado para desenvolver com sucesso a tarefa.
- **Interpretação:** As expressões multipalavras têm significados próprios. Nesta etapa é importante identificar o sentido da expressão independente do contexto em que ela foi utilizada.

¹¹do inglês *multiwords expressions* (MWE)

- **Desambiguação:** Está relacionada com a etapa anterior; ambas se concentram na classificação das multipalavras. Além das multipalavras terem sentidos próprios, elas podem ter mais de um sentido. Inclusive uma mesma expressão em um contexto pode ser uma multipalavra, enquanto em outro contexto não é. Seguindo essa linha, a desambiguação diferencia-se da interpretação em que no último caso a classificação é realizada independente do contexto da multipalavra.
- **Aplicações:** O tratamento das multipalavras é importante para diferentes áreas de PLN, por isso, os organizadores do *MWE workshop* acreditam que a maioria dos léxicos foram criados para um propósito específico e em diferentes áreas de PLN podem surgir métodos interessantes para construir esses recursos.

Atualmente, a identificação de multipalavras, em um contexto monolíngue, é baseada na definição de critérios específicos, mediante os quais é criada uma lista de expressões candidatas. Os critérios de seleção da lista candidata são baseados em técnicas de reconhecimento de padrões, que consideram desde padrões simples como sequências de n-gramas ou listas de expressões previamente definidas, até padrões mais complexos como sequências estruturadas de etiquetas morfossintáticas ou relações sintáticas. Os padrões selecionados para extrair as multipalavras dependem do tipo de informação que se deseja extrair e da língua que se está analisando. É importante salientar que, para obter padrões mais complexos, é preciso dispor de alguns recursos, tais como anotadores morfossintáticos e *parsers*.

Depois de criar a lista com expressões candidatas, comumente são implementados filtros para depurar essa lista. Alguns dos filtros propostos incluem: listas de *stopwords*; limiares de contagem das expressões nos corpúscos ou medidas de associação lexical. As medidas de associação avaliam a coocorrência das palavras que compõem a expressão candidata. Desta forma, elas têm sido muito utilizadas na classificação das expressões candidatas, mantendo na lista final as expressões maiores que um limiar previamente definido. Espera-se que quanto maior for o valor da medida de associação maior será a probabilidade da expressão candidata ser uma multipalavra real. O cálculo de algumas das medidas de associação mais utilizadas é apresentado no Apêndice A deste documento. Preferiu-se colocar essas equações em um apêndice para ganhar clareza na apresentação da revisão literária de geração de léxicos bilíngues. Seguindo essa linha, a literatura reconhece que o processo de

identificação de multipalavras em um contexto monolíngue pode ser separado em duas etapas: (i) extrair as expressões candidatas e (ii) ordenar ou filtrar essas expressões (Ramisch, 2012).

Um *framework* bastante utilizado para a extração de multipalavras é o *mwetoolkit* (Ramisch, 2012), disponível livremente e dedicado à aquisição automática de expressões a partir de um *corpus* monolíngue. O *mwetoolkit* disponibiliza um procedimento em *pipeline* que, empregando como base de conhecimento um *corpus* monolíngue, extrai uma lista de expressões candidatas. O *corpus* utilizado pela ferramenta pode ou não estar previamente anotado com etiquetas morfossintáticas ou *parsers* de dependência. O *framework* permite a definição de padrões em diversos níveis seguindo um formalismo similar ao das expressões regulares. Esses padrões são definidos seguindo informações linguísticas, exemplos do *corpus*, anotações manuais, etc. e são utilizados na criação da lista de expressões candidatas.

Para filtrar a lista de expressões candidatas, o *mwetoolkit* implementa diversos métodos de ordenação, tais como, limiares baseados na contagem de frequências, sequências de *n*-gramas e medidas de associação. A lista com as expressões obtidas pode ser inserida diretamente em qualquer aplicação ou pode ser avaliada manualmente para verificar, das expressões selecionadas, quais realmente podem ser classificadas como multipalavras. Esse *framework* é independente de língua e já foi utilizado para extrair automaticamente verbos pronominais do português do Brasil (Duran et al., 2013).

Em um contexto bilíngue, como é o caso do projeto HABLA, em que se deseja obter uma lista de multipalavras em uma língua fonte e suas equivalências mais prováveis na língua alvo, diversas abordagens têm sido propostas. Em geral, as abordagens seguem o mesmo procedimento usado no contexto monolíngue para construir uma lista de multipalavras na língua fonte, sendo necessário um passo adicional em que as possíveis traduções para cada entrada da lista são procuradas. Outras abordagens obtêm as multipalavras e suas traduções ao mesmo tempo, porém em qualquer dos casos o léxico bilíngue precisa ser validado.

A maioria das técnicas de extração de expressões no contexto bilíngue é baseada em *corpus* paralelos (Caseli et al., 2009), mas há outras abordagens que empregam *corpus* comparáveis (Daille et al., 2004; Morin e Prochasson, 2011) para criar os recursos lexicais. De qualquer forma, estudos (Wu e Zhou, 2003a; Morin e Prochasson, 2011) que combinaram *corpus* monolíngues, com *corpus* bilíngues, na tarefa de aquisição de multipalavras obtiveram melhores

valores de precisão e cobertura.

Caseli et al. (2009) obtiveram uma lista bilíngue de expressões em português e inglês por meio do alinhamento lexical de um corpus paralelo. O corpus foi alinhado em ambas as direções utilizando o *toolkit* Giza++ com a heurística *grow-diag-final*. A heurística *grow-diag-final* é implementada no Giza++ para simetrizar os alinhamentos lexicais obtidos nas duas direções. Essa e outras heurísticas foram propostas por (Och e Ney, 2003) para aprimorar os alinhamentos entre frases. A lista extraída a partir dos alinhamentos foi filtrada considerando diversos padrões, baseados em regras, até se obter uma lista com expressões na língua fonte e seus correspondentes na língua alvo. O método baseado em alinhamento se mostrou eficiente, quando comparada a lista obtida com uma lista criada por meio de um algoritmo monolíngue de extração de n-gramas. No entanto, a cobertura do método foi bastante limitada.

ZarrieB e Kuhn (2009) combinaram a análise sintática de um corpus paralelo com o alinhamento do mesmo, para extrair listas bilíngues de frases verbais em alemão e inglês. O corpus também foi alinhado lexicalmente utilizando o Giza++. Os autores definiram várias heurísticas para garantir que as expressões obtidas eram realmente frases verbais:

- um verbo na língua fonte foi alinhado com um verbo na língua alvo;
- o objeto do verbo na língua fonte é um substantivo tanto na língua fonte, como na língua alvo e
- existe uma relação sintática na língua fonte entre o verbo e o substantivo.

O método recuperou 82,1% das traduções corretamente; delas 60% foram consideradas realmente multipalavras. Isso mostra que o método, em 40% dos casos, levanta expressões que não são consideradas expressões multipalavras.

Outros autores têm aproveitado os corpus comparáveis na tarefa de gerar léxicos bilíngues. Daille et al. (2004) e Morin e Prochasson (2011) criaram abordagens similares para obter léxicos bilíngues em francês e inglês. Nos dois estudos as listas na língua alvo e fonte foram criadas de forma individual e depois se estabeleceram as equivalências de tradução por meio de medidas de distância entre os contextos em que aconteciam as expressões. As saídas dos estudos foram avaliadas com léxicos bilíngues extraídos de corpus paralelos e na tarefa de criar léxicos bilíngues para dicionários terminológicos da medicina. Os métodos propostos utilizaram métodos de alinhamento com heurísticas para gerar os léxicos bilíngues. As heurísticas ou filtros propostos

são em sua maioria para determinar se a entrada do léxico é multipalavra ou não. Isso evidencia a dificuldade em identificar expressões fixas. A tradução automática das multipalavras também é uma tarefa complexa, pois, em geral, elas são traduzidas como uma unidade e não por meio da tradução das palavras individuais que as compõem.

Métodos de tradução baseados em frases, como o implementado no *tool-kit* Moses, por exemplo, conseguem traduzir expressões multipalavras com sucesso. No entanto, às vezes falham, pois esses métodos não consideram informações do contexto das expressões durante a tradução. Em caso de falhas é comum observar expressões incompletas ou com palavras que não pertencem à multipalavra (Durrani et al., 2013; Carpuat e Diab, 2010). Como foi comentado, o contexto em que a multipalavra aconteceu pode determinar se a expressão é ou não multipalavra.

Bai et al. (2009) apresenta um algoritmo que procura, em um corpus paralelo previamente alinhado, as possíveis traduções de cada uma das entradas de uma lista de expressões dada. O método ordena as traduções candidatas utilizando uma medida de associação que incorpora informação no nível lexical, chamada *Normalizing Correlation Count*. Uma das propostas dos autores foi criar essa medida para ordenar as traduções mais prováveis com a finalidade de tratar alguns dos problemas que as medidas de associação tradicionais apresentam: (i) traduções candidatas com palavras ruidosas e (ii) traduções candidatas que contêm palavras utilizadas com muita frequência nos textos, como é o caso de *stopwords*. O algoritmo mostrou bom desempenho para diferentes tipos de multipalavras: frases, palavras simples e palavras compostas. Além disso, o algoritmo foi integrado no *framework* de tradução automática Moses para os pares de línguas inglês e chinês, obtendo melhores traduções quando comparadas com as de referência. O Algoritmo 1 mostra

um resumo do procedimento apresentado por Bai et al. (2009).

Algoritmo 1: Algoritmo apresentado por Bai et al. (2009)

Entrada: Dada uma lista de expressões em uma língua fonte f , em que cada expressão $w = t_1, t_2, \dots, t_r$ é composta por $t, t > 0$ e um cópulo paralelo entre as línguas fonte f e alvo e .

Procedimento

Para cada expressão da lista de entrada

Passo 1: Diminuir o espaço de busca (Selecionar as sentenças paralelas $f = f_1 + f_2 + \dots + f_n$ e $e = e_1 + e_2 + \dots + e_n$). De forma que w é uma subsequência de f .

Passo 2: Selecionar as palavras e na sentença e que estão correlacionadas com w para isso:

Supõe-se que:

Cada palavra e da sentença e pode ter sido gerada por alguma palavra r na sentença fonte. Empregando o modelo de tradução proposto por Brown et al. (1993).

Se existe r , então a relação de r e w pode ser alguma das seguintes:

i) Todas as palavras em r pertencem a w . Ex: $r \subseteq w$. Então r é gerada unicamente por w ;

$NormalizingCorrelationCount = 1$

ii) Nenhuma palavra em r pertence a w , Ex: $r \subseteq f - w$. Então e foi gerado por palavras que pertencem ao contexto de w $NormalizingCorrelationCount = 0$

ii) Algumas palavras em r pertencem a w , enquanto outras são palavras de contexto;

$NormalizingCorrelationCount = weightedcorrelationcount(wcc)$

Passo 3: Estimar *weighted correlation count* (wcc). Medida de associação proposta, que avalia a relação entre a palavra candidata e e a expressão w para cada par de sentenças (f, e) .

$$wcc(e; e, f, w) = \frac{\sum_{\forall f_i \in w} p(e|f_i) + \Delta w}{\sum_{\forall f_j \in f} p(e|f_j) + \Delta f} \quad (3.18)$$

Passo 4: Normalizar a correlação. Normalizando a medida de associação de cada candidata no subcópus.

$$ncf = \sum_{i=1}^n wcc(e; e^{(i)}, f^{(i)}, w) \quad (3.19)$$

Passo 5: Após normalizar a wcc, o valor da medida ncf é utilizado para ranquear as n palavras candidatas à tradução.

Passo 6: Construir as sequências de traduções candidatas: As palavras candidatas são marcadas nas sentenças do cópus. Se aparecem várias palavras candidatas na mesma sentença, a sequência dela é marcada como uma tradução candidata. Além disso, as *stopwords* adjacentes à palavra candidata também são incluídas na sequência de traduções candidatas.

Passo 7: Gerar possíveis traduções (Aumentar a cobertura do método) Como a tradução de uma expressão w pode ser uma palavra ou uma subsequência de palavras das sequências candidatas criadas no Passo 6. Estas sequências candidatas podem ser falsos positivos, desta forma todas as subsequências da sequência candidata são separadas e consideradas como traduções candidatas.

Passo 8: Calcular *weighted correlation count* (wcc) para filtrar as subsequências comuns e ranquear as sequências candidatas. As sequências m , obtidas no Passo 6 foram identificadas nas sentenças, assim parte-se do pressuposto que a $wcc(m) = 1$. Desta forma, a *weighted correlation count* (wcc) de t , uma sequência de m é definida como a seguir:

$$wc = (t; e, f, m, w) = \prod_{\forall e \in m-t} (1 - wcc(e; e, f, w)) \quad (3.20)$$

Passo 9: Ranquear lista de traduções candidatas empregando a medida de associação DICE.

Passo 10: 10. As traduções candidatas foram avaliadas segundo a cobertura do método, que foi estimado após uma anotação manual, como se mostra a seguir:

$$cobertura = \frac{1}{n} \sum_w \frac{A_w \subset C_w}{|A_w|} \quad (3.21)$$

Diferentemente dos métodos de Caseli et al. (2009) e ZarrieB e Kuhn (2009), que juntam a obtenção da tradução com a extração da multipalavra, Bai et al. (2009) recebem como entrada a expressão multipalavra e procuram a melhor tradução, analisando o contexto em que a expressão aconteceu. O método de Bai et al. (2009) utiliza sentenças alinhadas e as medidas de associação que propõe são baseadas no cálculo da probabilidade de traduzir palavras. O método de Bai et al. (2009) é interessante quando a etapa de extrair multipalavras já foi resolvida. Na geração de léxicos bilíngues para criar auxílios à escrita a separação da etapa de extração do item na língua fonte e a obtenção de seu equivalente na língua alvo permite gerar léxicos de forma mais controlada e, como consequência, de maior utilidade.

3.2.2 Avaliação da Geração de Léxicos Bilíngues

A geração de léxicos bilíngues, comumente, é avaliada de forma intrínseca e extrínseca (Bai et al., 2009; Ramisch, 2012; Caseli et al., 2009; Morin e Prochasson, 2011). Na avaliação intrínseca, as entradas do léxico são avaliadas considerando-se o conteúdo que representam, de acordo com alguma métrica de interesse. Na avaliação extrínseca, o léxico é inserido em uma aplicação de PLN e em função do desempenho da tarefa é avaliado o desempenho do léxico gerado.

A avaliação intrínseca pode acontecer de forma manual ou automática. Na avaliação intrínseca manual as entradas do léxico bilíngue são verificadas por juízes humanos que julgam, segundo algum critério, a corretude dos léxicos na língua fonte e alvo. Já, a avaliação intrínseca automática consiste em comparar o léxico bilíngue com outro léxico, previamente criado, seja por outro método ou de forma manual. Em ambos os casos costuma-se considerar as medidas de precisão e cobertura.

Nesse cenário, a avaliação intrínseca é fraca porque não permite confirmar que os mesmos resultados vão ser alcançados para dados com características diferentes das testadas (Caseli, 2007). Isso acontece porque o desempenho do algoritmo de aquisição automática de léxicos bilíngues depende de diversos parâmetros, tais como, tamanho do corpus, pré-processamento do corpus, ferramentas automáticas inseridas na tarefa, como anotadores morfossintáticos ou *parsers*. Além disso, os léxicos bilíngues dependem da língua e da tarefa para as quais foram criados. Apesar disso, a avaliação intrínseca parece ser a metodologia mais comumente empregada, pois foi utilizada na maioria dos estudos consultados (Daille et al., 2004; Morin e Prochasson, 2011;

ZarrieB e Kuhn, 2009; Caseli et al., 2009; Wu e Xia, 1994; Fung, 1995).

Como comentado, na avaliação extrínseca o léxico bilíngue é inserido em uma tarefa de PLN e o desempenho da tarefa se correlaciona com o desempenho do léxico. Essa avaliação pode acontecer de forma manual, em que humanos avaliam a saída da tarefa verificando se o léxico consegue alguma melhoria no desempenho. Nenhum dos métodos estudados emprega essa metodologia. Além disso, a avaliação extrínseca do léxico bilíngue pode acontecer de forma automática. Esse tipo de avaliação foi observada nos trabalhos de Bai et al. (2009) e Koehn e Knight (2002), que em ambos os casos o léxico foi testado para aprimorar a tradução automática.

Quando há multipalavras no léxico bilíngue, além de avaliar as equivalências apresentadas no léxico, é preciso avaliar a tarefa de identificação das expressões apontadas como multipalavras. Nesse caso, as metodologias de avaliação intrínseca e extrínseca também são aplicadas. As medidas de precisão e cobertura são as utilizadas para quantificar o desempenho dos métodos na seleção das entradas do léxico bilíngue. Quando a extração das expressões acontece de forma automática, não é possível calcular a cobertura, porque não há conhecimento do total de expressões que deviam ser identificadas.

A avaliação intrínseca pode acontecer de forma manual ou automática. No primeiro caso, juízes são treinados para identificar as multipalavras reais a partir da lista de candidatas. É frequente realizar uma classificação binária em que o juiz indica um, caso a entrada seja multipalavra e zero, caso contrário. A avaliação também pode acontecer de forma automática comparando-se a lista de candidatas com uma lista previamente compilada. Porém, essa última avaliação não é realista, pois não há evidências de que se a expressão candidata não estiver no dicionário, esta não seja um multipalavra real, pois dicionários são recursos lexicais finitos.

A avaliação da extração de multipalavras é um momento adequado para realizar a interpretação e desambiguação das multipalavras candidatas. Nessa etapa tanto métodos manuais quanto automáticos podem ser inseridos. Para aplicar métodos automáticos é necessário disponibilizar ferramentas de desambiguação de sentidos. Por esse motivo, a maioria das metodologias propostas nessa etapa são análises manuais.

3.3 Análises Linguístico-computacionais para Gerar os Léxicos de Interesse

Nesta seção, são apresentados os estudos da literatura que deram base para a compilação dos léxicos propostos nesta pesquisa:

- listas de cognatos e falsos cognatos das línguas português e espanhol;
- listas de pares de marcadores discursivos;
- listas das expressões formulaicas que aparecem nos textos científicos;
- listas de verbos relacionados com pesquisa científica em português.

Para alguns itens será mostrada sua importância em processos de aprendizado ou para apoiar a escrita de textos. Em outros casos serão apresentados estudos que utilizaram os métodos apresentados na Seção 3.2 para compilar léxicos bilíngues. Além disso, é mostrado um grupo de estudos que fazem análises para comparar a produção de itens lexicais dos nativos e dos aprendizes. Nesse cenário predominam as análises linguísticas, identificando-se vários estudos que analisam o uso de colocações, marcadores discursivos (Aidinlou e Mehr, 2012), expressões formulaicas, etc.. Para apoiar a produção escrita também há vários trabalhos que tentaram criar recursos e propor abordagens para aprimorar a escrita de colocações, por exemplo.

3.3.1 Cognatos e Falsos Cognatos

Nesta pesquisa, considera-se a definição de cognatos e falsos cognatos proposta por Frunza e Inkpen (2009), em que: **cognatos** entre duas línguas são pares de palavras que além de serem traduções mútuas aparentemente são similares. Para vários autores (Simard et al., 1992; Fattah et al., 2007; Frunza e Inkpen, 2009; Nakov, 2009) cognatos são pares de palavras idênticas ou muito similares entre si, considerando sua ortografia e sua fonética. Os pares de palavras em espanhol e português, respectivamente, "amor – amor" e "jefe – chefe" são exemplos de cognatos entre as línguas português e espanhol.

Os **falsos cognatos** são pares de palavras em duas línguas que, segundo a percepção, são similares, mas que têm diferentes significados, por exemplo, "aula" (lição de uma disciplina, em português) – "aula" (sala onde se ministra uma disciplina, em espanhol). Os falsos cognatos podem ter diferentes significados dependendo do contexto em que foram utilizados, por exemplo, a

palavra "gente" em espanhol significa pluralidade de pessoas, pessoas do povo e coloquial de família. Em português, a mesma, significa conjunto de pessoas com o mesmo nome, raça ou família. Nesse caso, tem o mesmo significado que em espanhol, mas a frase "a gente" é usada no sentido do pronome pessoal nós. Outros exemplos de pares de falsos cognatos são *firma* – assinatura / *asignatura* - disciplina.

Diversas aplicações de processamento de língua natural se beneficiam da identificação de cognatos e faltos cognatos. Algumas dessas aplicações incluem: alinhamento sentencial (Simard et al., 1992; Melamed, 1999; Koehn e Knight, 2002), alinhamento de textos paralelos (Gomes, 2009), indução de léxicos bilíngues (Mann e Yarowsky, 2001), identificação de nomes de medicamentos confundíveis (Kondrak e Dorr, 2004) e tradução automática estatística (Simard et al., 1992; Melamed, 1999; Fattah et al., 2007).

Em um cenário de aprendizagem de língua estrangeira, os cognatos e falsos cognatos desempenham um importante papel. O reconhecimento dos cognatos ajuda o aluno na compreensão da leitura, além de contribuir com a expansão do novo vocabulário. Por outro lado, a existência de falsos cognatos representa um problema, pois seu uso incorreto acarreta em erros semânticos produzidos pela interferência da língua nativa. Furtado (2012) analisou a importância do estudo dos cognatos e falsos cognatos em sala de aula e Gonçalves (2013) fez uma análise de erros destes itens em produções escritas em espanhol por nativos do português. Ambas as autoras concordam com a importância de apresentar esses itens em diversos estágios do aprendizado da língua. Embora Gonçalves (2013) tenha observado erros deste tipo em maior frequência em etapas iniciais do aprendizado, entre alunos com níveis de aprendizado mais avançados também foram observados esses erros. Como alternativa para aprimorar o uso desses itens lexicais, Furtado (2012) e Gonçalves (2013) consideram importante a criação de listas que incluíam esses itens para apresentar para os alunos. Adicionalmente, Gonçalves (2013) conclui que chamar a atenção do aluno quando um falso cognato está presente em um texto ou tarefa de sala de aula, seria uma alerta importante para levar o aprendiz ao nível mais próximo possível da língua-alvo.

A criação manual de listas de cognatos e falsos cognatos é uma tarefa demorada que requer muito esforço dos lexicógrafos. Uma alternativa para solucionar este problema é a criação automática de dicionários de cognatos e falsos cognatos a partir de corpus bilíngues.

Diversas abordagens têm sido propostas para identificar de forma automá-

tica cognatos entre pares de línguas (Simard et al., 1992; Kondrak e Dorr, 2004; Kondrak, 2005). A identificação automática de falsos cognatos tem recebido menos atenção (Mitkov et al., 2007; Frunza e Inkpen, 2009; Nakov, 2009). Segundo Mitkov et al. (2007), isso poderia estar relacionado com o fato de que em muitas aplicações é interessante identificar palavras equivalentes entre duas línguas e não identificar falsos cognatos ou expressões que não são equivalentes.

Entre as abordagens para identificar os falsos cognatos citam-se:

- a procura por expressões que coocorrem nos corpúscos paralelos;
- a utilização de dicionários para encontrar similaridades entre expressões;
- o uso de características ortográficas das palavras e, finalmente,
- comparações entre os pares de palavras analisando a semântica das palavras.

Segundo Mitkov et al. (2007) e Nakov (2009), considerando as abordagens utilizadas, os estudos dedicados à identificação de cognatos e falsos cognatos podem ser divididos em 3 grupos:

1. métodos que utilizam medidas de similaridades ortográficas e fonéticas;
2. métodos estatísticos que identificam os cognatos e os falsos cognatos baseando-se em corpúscos paralelos;
3. métodos que utilizam uma abordagem semântica.

A seguir são apresentados alguns dos trabalhos que há na literatura para identificar tanto cognatos quanto falsos cognatos.

Os trabalhos que propõem a abordagem ortográfica e fonética computam as similaridades entre os pares de palavras utilizando medidas de distância. Há diversos critérios para estimar a distância entre duas palavras. Do ponto de vista ortográfico, uma das medidas mais empregadas é a distância de edição¹², comentada em Mitkov et al. (2007) e Frunza e Inkpen (2009). A distância ED pode ser utilizada entre palavras de duas línguas que compartilham o mesmo dicionário e se concentra no cálculo do número mínimo de operações (eliminações, inserções ou substituições) necessárias para transformar uma palavra em outra. A distância ED normalizada (NED) também tem sido

¹²do inglês *Edit Distance (ED)*

utilizada e é calculada dividindo a soma de todas as operações para transformar uma sequência em outra, pelo comprimento da maior sequência. Outra medida popular e efetiva é o coeficiente LCSR (*Longest Common Subsequence Ratio*) proposto por Melamed (1999). O cálculo desse coeficiente é realizado procurando-se a sequência comum de maior comprimento entre duas palavras, dividido pelo comprimento da maior palavra. Por exemplo, o valor da medida LCSR para os cognatos "*mejorar*" e "*melhorar*" é $6/8 = 0,75$. Com relação às similaridades fonéticas, as medidas mais utilizadas são SOUNDEX e EDITEX. Essas medidas aproveitam as características individuais das palavras para estabelecer similaridade (Mackay e Kondrak, 2005). O algoritmo fonético SOUNDEX foi proposto para o inglês com o objetivo de indexar nomes com similaridade na pronúncia. O algoritmo transforma todas as letras da palavra seguindo um código numérico e mantendo a primeira letra da sequência. Após a transformação, o código é truncado, sendo a palavra representada por quatro caracteres. Por outro lado, a medida EDITEX combina as propriedades da distância de edição com a estratégia de transformação usada pelo algoritmo SOUNDEX.

Simard et al. (1992) identifica cognatos para auxiliar a tarefa de alinhamento sentencial de textos paralelos. No estudo, o autor trabalha com os pares de línguas francês-inglês e considera cognatas aquelas palavras que compartilhem os quatro primeiros caracteres. A partir disso, foi proposta a medida PREFIX, que calcula o comprimento do prefixo comum para duas palavras, dividido pelo comprimento da maior palavra. Por exemplo, o prefixo comum dos cognatos "*mejorar*" e "*melhorar*" é 2 (as duas primeiras letras), dividido por 8 (maior comprimento entre as duas palavras) é 0,25. A medida PREFIX é utilizada em Kondrak (2005) para beneficiar técnicas de alinhamento lexical a partir da identificação previa de palavras cognatas. Para desenvolver a tarefa Kondrak (2005) também emprega outras medidas ortográficas tais como: LCSR, PREFIX, DICE e IDENT. O coeficiente DICE é uma medida de similaridade ortográfica, baseada nos números de n-gramas coincidentes entre as palavras. O coeficiente divide a quantidade de bigramas comuns para duas palavras pela soma dos bigramas de cada palavra, como mostra a Equação 3.22.

$$DICE(X, Y) = \frac{2 * |bigramas(X) \cap bigramas(Y)|}{|bigramas(X) + bigramas(Y)|} \quad (3.22)$$

em que, $bigramas(X)$ é o conjunto de caracteres bigramas na palavra X . Por exemplo, o coeficiente DICE para as palavras "*mejorar*" e "*melhorar*" é $8/13 =$

0.61 (os bigramas compartilhados são me-or-ra-ar). Outras medidas foram definidas a partir da medida DICE, como é o caso da medida TRIGRAM, que se diferencia do coeficiente DICE, dado que TRIGRAM trabalha com trigramas e não com bigramas. A medida IDENT é uma das mais simples, sendo utilizada, comumente como *baseline*. Ela retorna 1 se as palavras são idênticas e 0 caso contrário.

Kondrak e Dorr (2004) combinam as medidas de similaridade DICE, TRIGRAM, LCSR, ED, NED e SOUNDEX, para tratar o problema de diferenciar nomes de medicamentos com pronúncias muito similares. Os autores realizam uma análise para identificar os problemas que apresentam as medidas de similaridade baseadas no número de n-gramas coincidentes entre palavras. A partir disso, eles propõem as medidas BI-SIM e TRI-SIM que caracterizam várias das medidas anteriormente mencionadas e mostram um desempenho adequado após serem comparadas com outras medidas. Os autores consideram que o melhor desempenho foi alcançado após a combinação de várias medidas.

Vários trabalhos (Inkpen et al., 2005; Mitkov et al., 2007; Frunza e Inkpen, 2009; Nakov, 2009) utilizam técnicas estatísticas e de aprendizado de máquina para caracterizar cognatos e falsos cognatos a partir da combinação de medidas ortográficas, fonéticas e semânticas.

Inkpen et al. (2005) identificam cognatos e falsos cognatos combinando técnicas de aprendizado de máquina com medidas de similaridades ortográficas e fonéticas. Os autores fazem uma análise detalhada do impacto das medidas tanto de forma individual quanto a combinação na tarefa de classificação. Algumas das medidas de melhor desempenho foram BI-SIM e TRI-SIM.

Frunza e Inkpen (2009) propõem o uso de técnicas de aprendizado de máquina para separar cognatos de falsos cognatos, utilizando várias medidas de similaridade entre pares de palavras como características de classificação. O estudo disponibiliza um corpus paralelo dos pares de línguas francês-inglês e as medidas de similaridade tanto ortográfica quanto fonética utilizadas foram: IDENT, PREFIX, DICE, TRIGRAM, XDICE, XXDICE, LCSR, NED e SOUNDEX, além das medidas BI-SIM e TRI-SIM propostas por Kondrak e Dorr (2004). Os autores avaliam o impacto das medidas de similaridade tanto de forma individual como a combinação delas na tarefa de diferenciação de cognatos e falsos cognatos, além de estabelecer limiares específicos para cada medida com a finalidade de separar cognatos de falsos cognatos. Especificamente, nesse estudo são introduzidas evidências semânticas para diferenciar cognatos de

falsos cognatos. Eles empregam um dicionário bilíngue seguindo a hipótese de que cognatos aparecerão no dicionário com uma única tradução, enquanto falsos cognatos podem aparecer com mais de um significado. Para isso, os autores treinam um classificador para distinguir entre cognatos e falsos cognatos, por meio de um corpus paralelo.

Em geral, as abordagens semânticas para diferenciar cognatos de falsos cognatos são baseadas em medidas de similaridade semântica. Estas medidas são computadas a partir de estruturas taxonômicas de representação dos léxicos, como é caso da *WordNet*¹³, ou por meio de algoritmos que avaliam similaridade entre as distribuições de um corpus. Algumas das medidas de similaridade semântica utilizadas para computar distância entre os elementos de estruturas taxonômicas são a distância Euclidiana, a distância cosseno, a distância *Kullback-Leibler*, a distância entre conceitos, apresentada em Wu e Palmer (1994) e a distância *Leacock and Chodorow's*, utilizada em Mitkov et al. (2007).

Em Kondrak (2001) é incorporada uma análise semântica para separar cognatos de falsos cognatos. Nesse trabalho, o autor parte da premissa que palavras cognatas mostram similaridade fonética e semântica. Seguindo essa ideia, introduz um procedimento para identificar palavras relacionadas a partir da estimativa de similaridades semânticas entre glosas. O autor identifica a palavra chave da glosa e empregando a *WordNet* procura sinônimos nas glosas e desta forma similaridades entre as palavras. No estudo são combinadas as evidências semânticas com medidas de similaridade ortográfica e fonética.

Mitkov et al. (2007) propõem uma abordagem para identificar de forma automática cognatos e falsos cognatos, que está dividida em dois passos. Primeiramente, são extraídos de um corpus bilíngue as palavras candidatas para formar uma lista de pares de palavras. Após a extração, cada par da lista é classificado em cognato ou em falso cognato. A extração dos pares candidatos é baseada em medidas de similaridade ortográfica, enquanto a classificação das palavras extraídas é realizada a partir de medidas de similaridade semântica. As medidas de similaridade ortográfica utilizadas foram a LCSR e a NED. A base para estabelecer similaridades semânticas no estudo foi a estrutura taxonômica *EuroWordNet*¹⁴ que inclui as quatro línguas testadas (inglês-francês, inglês-espanhol, inglês-alemão e espanhol-francês). As medidas de similaridade semântica utilizadas foram a distância entre conceitos e a distância *Le-*

¹³<https://wordnet.princeton.edu/>

¹⁴<http://www.illc.uva.nl/EuroWordNet/>

acock and Chodorow's. A distância entre conceitos é apresentada na Equação 3.23 e estima quão próximos os conceitos aparecem na hierarquia.

$$ComSim(C1, C2) = \frac{2 * N3}{N1 + N2 + 2 * N3} \quad (3.23)$$

supondo que existe $C3$ que é o super conceito comum mais próximo entre os conceitos $C1$ e $C2$. $N1$ é o número de nós no caminho de $C1$ até $C3$. $N2$ é o número de nós entre o caminho de $C2$ até $C3$. $N3$ é o número de nós entre $C3$ e a raiz da estrutura hierárquica.

A distância *Leacock and Chodorow's* usa o comprimento do caminho normalizado entre dois conceitos $C1$ e $C2$ e é calculada segundo a Equação 3.24

$$LC(C1, C2) = -\log\left[\frac{len(C1, C2)}{2 * max}\right] \quad (3.24)$$

em que, len é o número de arestas do menor caminho entre os dois conceitos e max é a profundidade da estrutura.

A abordagem proposta por Mitkov et al. (2007) foi testada por meio de dois experimentos. No primeiro experimento, a classificação foi realizada utilizando taxonomias e no segundo utilizou-se um vetor de coocorrências das palavras para procurar similaridades entre os conceitos. Os autores mostraram que o melhor desempenho foi obtido no primeiro experimento e utilizando a distância entre os conceitos. Um aporte interessante desse trabalho é a seleção automática dos pares de palavras por meio das medidas de similaridade ortográficas.

Nakov (2009) apresenta um algoritmo não supervisionado que identifica automaticamente falsos cognatos, combinando técnicas estatísticas, que observam ocorrência e coocorrência de palavras em textos paralelos com técnicas que medem similaridade semântica usando a *Web* como um *cópus*. Segundo o autor, há poucos trabalhos que extraem os falsos cognatos de um *cópus* paralelo e a maioria deles emprega técnicas de alinhamento lexical para extraí-los. O autor propõe, então, a extração de falsos cognatos após um alinhamento sentencial dos textos paralelos. O método é aplicável para qualquer língua, mas foi testado com as línguas búlgara e russa. Nakov (2009) seleciona os pares de palavras candidatas de *cópus* paralelos ou comparáveis, utilizando medidas de similaridade ortográfica. Uma vez formada a lista de pares candidatos, o autor começa a tarefa de classificação. O método proposto para separar cognatos de falsos cognatos segue a hipótese de que se duas palavras são cognatas então as palavras que formam parte do contexto

em que a palavra esta inserida serão traduções mútuas. A partir disso, o autor propõe dado um par de palavras, extrair o contexto local (uma janela de comprimento três palavras), para criar dois vetores de palavras, neste caso a palavra búlgara com seu contexto local e a palavra russa, também com seu contexto local. Uma vez formados os dois vetores, o autor utiliza a distância cosseno para calcular a similaridade entre eles. A precisão para este algoritmo foi de 77%.

Recentemente, Gamallo e García (2012) propuseram um método para extrair cognatos de um córpus comparável. Após o estudo foi criado um léxico para os pares de línguas português e espanhol com aproximadamente 27.000 pares de lemas e palavras compostas extraídas da *Wikipedia*. Para identificar as palavras (substantivos, adjetivos e verbos) com ortografia similar foi empregada a distância DICE entre os lemas, e o método de aprendizado de máquina precisava de listas de palavras nas duas línguas para estabelecer a classificação. Para isso, foram utilizados dois dicionários português-espanhol. O primeiro de língua geral, chamado *OpenTrad-Apertium*¹⁵ e o segundo criado por meio da *Wikipedia*. A precisão do algoritmo proposto para extrair os pares de palavras foi aproximadamente de 92% e os maiores problemas foram identificados na extração de estrangeirismos, palavras com prefixos tais como anti-fascista/fascista e palavras compostas.

Nos trabalhos da literatura que identificam cognatos e falsos cognatos observa-se a aplicação de diversas técnicas, assim como de diversos recursos, tais como: dicionários, córpus paralelos e córpus comparáveis. Os métodos propostos seguem a mesma metodologia apresentada na Seção 3.2, em que métodos para compilar léxicos bilíngues foram apresentados. A metodologia adotada para validar as listas de cognatos e falsos cognatos é a mesma que é aplicada na geração de léxicos bilíngues, destacando-se a necessidade de análises manuais e o cálculo da precisão para avaliar os métodos propostos.

3.3.2 Verbos

Nesta pesquisa de doutorado, foi proposta a criação de um léxico bilíngue que mostre os verbos que são utilizados na escrita acadêmica em português com seus equivalentes em espanhol. Aprendizes de uma segunda língua têm dificuldades em se adequar aos léxicos de determinados domínios. Esse fato é observado na escrita acadêmica, em que às vezes itens da língua geral ou da língua falada são incluídos, quando termos científicos eram mais adequa-

¹⁵Disponível em: <http://sourceforge.net/projects/apertium/files/>

dos. Neste capítulo, diversas abordagens para estabelecer equivalências entre uma língua alvo e fonte foram apresentados; qualquer uma delas poderia ser utilizada para gerar o léxico bilíngue de verbos em português e em espanhol. Além disso, seria necessário considerar um anotador morfossintático que automaticamente identificasse os verbos das duas línguas. Seguindo essa linha, o problema que será tratado nesta seção está relacionado com as abordagens utilizadas para identificar os léxicos que são da escrita acadêmica.

Diversos estudos (Simpson-Vlach e Ellis, 2010; Morley, 2015; Gardner e Davies, 2013; Dos Santos, 2006; Baptista et al., 2010) foram propostos para levantar os itens mais comumente utilizados na escrita acadêmica. Em geral, esses itens são extraídos por meio de análises estatísticas dos textos, que contrastam *córpus* da língua geral ou falada com *córpus* da língua acadêmica para identificar os itens que são frequentes na escrita acadêmica (Simpson-Vlach e Ellis, 2010; Gardner e Davies, 2013). Indicar o que é ou não inadequado nos textos acadêmicos, em geral, é muito subjetivo, pois o que é frequente em uma área, pode ser inadequado em outra. Desta forma, os estudos que identificam os itens da escrita acadêmica fazem análises com *córpus* de diferentes áreas do conhecimento (Simpson-Vlach e Ellis, 2010; Morley, 2015; Dos Santos, 2006; Baptista et al., 2010). No entanto, compilar *córpus* de diferentes áreas do conhecimento não é uma tarefa fácil, sendo mais difícil ainda compilar *córpus* paralelos.

Gardner e Davies (2013) combinaram um grande *córpus* acadêmico de diversas áreas do conhecimento com um *córpus* da escrita geral para criar uma lista de vocabulário acadêmico, nomeada AVL (*Academic Vocabulary List*). A lista foi criada com o intuito de auxiliar os aprendizes de inglês acadêmico na produção escrita. A ideia do trabalho era disponibilizar o recurso com itens lexicais comuns à escrita acadêmica, em geral, sem distinção de área de conhecimento. Para isso, Gardner e Davies (2013) aplicaram estatísticas tais como: limiares de frequência; ocorrência do termo em pelo menos n áreas do conhecimento e limiar de frequência do termo em todas as áreas do conhecimento em que aparece. Os autores justificaram a necessidade desta lista e mostraram como o recurso conseguiu aumentar em aproximadamente 14% a cobertura de outra lista acadêmica, apresentada por Coxhead (2000).

Baptista et al. (2010) usaram como ponto de partida o vocabulário criado por Coxhead (2000) para obter uma lista de termos acadêmicos em português. A metodologia adotada foi manualmente estabelecer as equivalências entre os termos em inglês e em português até gerar o vocabulário em português. O

objetivo do trabalho era criar um recurso útil para apoiar a aprendizagem da língua portuguesa. O léxico é composto por 1.812 palavras, das quais 409 são verbos, categoria morfossintática de interesse nesta pesquisa. Porém o recurso criado por Baptista et al. (2010) contém verbos da escrita acadêmica do português europeu que muitas vezes não coincidem com os verbos usados nos textos acadêmicos em português do Brasil.

Dos Santos (2006) também levantou um léxico da escrita acadêmica em português. O léxico foi extraído por meio de uma análise manual de corpús científicos de diferentes áreas do conhecimento. Porém, a lista criada por Dos Santos (2006) é pequena, com somente 54 verbos da escrita acadêmica.

Os recursos disponíveis para o português que mostram os itens usados na escrita acadêmica não têm boa cobertura, mostrando a dificuldade em criar um recurso deste tipo de forma manual. A análise manual pode contribuir para criar uma lista de itens com alta precisão, mas a cobertura é baixa, pois são poucos os itens incluídos.

3.3.3 Marcadores Discursivos

Espera-se que um texto escrito corretamente não seja apenas um texto livre de erros lexicais e gramaticais, mas também um texto que atenda ao seu gênero e apresente as ideias de forma coesa e coerente. Por isso, é importante que o aprendiz utilize corretamente os marcadores discursivos na língua estrangeira. Marcadores discursivos são itens lexicais externos às orações, na sua maioria invariável, que relacionam os enunciados e desempenham um importante papel na coesão dos textos (Aidinlou e Mehr, 2012; Fernández, 2005). A simples presença dos marcadores nos textos não é, porém, garantia de que o texto seja coerente. Aliás, pode ocorrer o contrário: o uso excessivo e desnecessário dos marcadores pode resultar em um texto com problemas de coerência e coesão (Castele e Collewaert, 2013; Aidinlou e Mehr, 2012; Fernández, 2005).

Na literatura foram encontrados estudos (Castele e Collewaert, 2013; Aidinlou e Mehr, 2012; Jalilifar, 2008) que, por meio de uma análise linguística, compararam como os aprendizes de uma segunda língua e os nativos usam os marcadores discursivos. Em geral, esses estudos fazem análises contrastivas para obter desvios ou inadequações no uso de marcadores discursivos em corpús de aprendizes e em textos produzidos por nativos. Entre os problemas relacionados ao uso de marcadores discursivos relatados nesses trabalhos estão:

- Emprego incorreto de marcadores discursivos (erros de construção gramatical);
- Sobreuso de marcadores discursivos: os aprendizes aprendem apenas um marcador para cada função e fazem uso recorrente dele;
- Uso de marcadores informais na escrita formal e vice-versa (desconhecimento do registro de cada marcador)
- Uso de marcadores de uma função para outra função (erros de escolha lexical)
- Uso de marcadores inexistentes, gerados por transferência da língua materna para a língua-alvo.

Uma das sugestões apontadas por uma análise conduzida por Fernández (2005) aponta a necessidade de um estudo aprofundado das características dos textos de aprendizes de português e de espanhol no que se refere ao emprego dos marcadores discursivos. Embora algumas pesquisas (Castele e Collewaert, 2013) tenham realizado essa análise para os aprendizes de espanhol, até o momento não se conhece nenhuma pesquisa voltada aos aprendizes de português como língua estrangeira. Além disso, no cenário desta pesquisa não se conhece nenhuma iniciativa computacional para criar um recurso que dê suporte à escrita de marcadores discursivos em português, tratando erros produzidos pela interferência da língua nativa.

No português, os marcadores discursivos têm sido estudados como elementos que contribuem para a estruturação retórica dos textos (Feltrim et al., 2003), obtendo resultados mais úteis para os nativos da língua, pois pressupõem conhecimentos da língua que os aprendizes normalmente não têm. Adicionalmente, para o português, os marcadores discursivos também foram estudados para apoiar a tarefa de análise discursiva dos textos (Pardo e Nunes, 2006). Uma das iniciativas anteriores é o projeto *Dizer*, no qual foi disponibilizada uma lista de marcadores discursivos subcategorizados de acordo com sua função RST(*Rhetorical Structure Theory*).

Apesar da importância dos marcadores na produção escrita, principalmente na escrita acadêmica, esses itens lexicais não estão totalmente inventariados em português e espanhol. A subcategorização dos marcadores lexicais de acordo com sua função, embora reconhecida, varia muito na literatura das duas línguas (Fernández, 2005; Shanru, 2012). Devido ao fato de existirem

vários marcadores discursivos para cada função, é muito importante contar-mos com uma subcategorização precisa desses itens lexicais por função em ambas as línguas (língua materna do aprendiz e língua-alvo). Isso porque a tradução dos marcadores discursivos não costuma respeitar uma relação de equivalência item a item, mas sim a relação de equivalência entre as subcategorias de função.

Um estudo foi desenvolvido por Fernández (2005) com a finalidade de classificar, analisar e comparar os marcadores discursivos em espanhol e em português em 200 textos do gênero jornalístico a partir de uma perspectiva linguístico-pragmática. A autora fez um levantamento dos marcadores presentes no corpus e considerou os cinco grupos apresentados por Zorraquino e Portolés (1999):

- Estruturadores da informação;
- Conectores;
- Operadores Argumentativos;
- Reformuladores e
- Conversacionais.

O resultado desse estudo, porém, não possui as características necessárias para uso computacional. Para automatização de auxílios à escrita é necessário que esse léxico esteja inventariado extensivamente em cada língua, agrupado por funções e que cada grupo de uma língua tenha um grupo equivalente na outra língua, como é mostrado na Figura 3.5.

Figura 3.5: Exemplo de agrupamento dos marcadores discursivos com a mesma função

A tarefa de traduzir marcadores discursivos de uma língua alvo para uma língua fonte é uma tarefa árdua devido à diversidade funcional dos marcadores discursivos. Em uma língua, um marcador é capaz de assumir todo um conjunto de funções, enquanto seu equivalente em outra língua não se corresponde com toda a diversidade de funções na língua fonte. Por esse fato, os tradutores são obrigados a procurar na língua alvo um correspondente linguístico que tenha, no mesmo contexto, propriedades funcionais idênticas, ou pelo menos tão semelhantes quanto possível (Fernández, 2005).

Para aprimorar a tradução obtida com um algoritmo automático é necessário considerar informações semânticas e discursivas desses itens. Até onde se conhece da literatura, para as línguas espanhol e português ainda não há um estudo que apresente detalhadamente a relação dos marcadores discursivos das duas línguas. O estudo mais próximo foi o apresentado por Fernández (2005), porém as equivalências criadas pela autora foram aproximadas, pois além de analisar poucas instâncias, a autora não considera as inúmeras sutilezas do processo de tradução dos marcadores discursivos.

Os métodos analisados para gerar léxicos bilíngues (Seção 3.2) poderiam ser empregados na tarefa de gerar um léxico dos marcadores discursivos em português e em espanhol. O léxico poderia ser extraído de corpúis paralelos e, embora precise ser validado manualmente, a tarefa de extração automática sempre vai ser mais rápida que a extração manual. Porém, uma das dificuldades da tarefa seria o reconhecimento automático dos itens fazendo função de marcador discursivo, pois como comentado, esses itens podem assumir diversas funções; o "e", por exemplo, tanto pode ser um marcador que une orações como sintagmas nominais. Um recurso lexical com as equivalências dos marcadores discursivos em espanhol e português poderia representar uma alternativa para que os hispano falantes reconhecessem as transferências destes itens de uma língua para a outra e, como consequência, diminuir o número de inadequações produzidas pela interferência da língua nativa.

3.3.4 Expressões Formulaicas

Expressões formulaicas (ou *cue-phrases* em inglês (Abdalla e Teufel, 2006)) são muito importantes não somente na língua geral, mas também na escrita acadêmica. Wray (2002) define expressões formulaicas como sequências, contínuas ou descontínuas, de palavras que são ou parecem ser frases pré-fabricadas, sendo que elas expressam um conceito e contribuem à fluidez do texto, por isso a importância de saber usar essas expressões na escrita

acadêmica.

Segundo Ellis et al. (2008) os aprendizes de uma segunda língua têm dificuldade em empregar expressões formulaicas do gênero acadêmico; isso foi constatado até na produção escrita de aprendizes com níveis mais avançados da língua. Na análise realizada por Ellis et al. (2008) observou-se, com frequência, como os aprendizes geram expressões formulaicas erradamente via tradução literal ou por interferência da língua nativa. Porém, na literatura, poucos estudos foram desenvolvidos com a finalidade de auxiliar a aquisição de expressões da escrita científica em uma segunda língua ou língua estrangeira. Por outro lado, diversos estudos (Ellis et al., 2008; O'Donnell et al., 2013) concentraram-se no levantamento dos padrões que representam como nativos e aprendizes usam as expressões formulaicas.

Computacionalmente, diversos métodos têm sido propostos para identificar as expressões formulaicas em textos acadêmicos (O'Donnell et al., 2013; Leedham, 2012; Simpson-Vlach e Ellis, 2010). O intuito desses métodos é mostrar aos usuários essas expressões e como elas são utilizadas. No entanto, esses auxílios não são direcionados para os aprendizes de uma segunda língua e sim para os nativos da língua. Esses métodos acontecem em um contexto monolíngue e os léxicos gerados não são bilíngues e não mostram as equivalências entre uma língua fonte e alvo.

Como comentado na Seção 3.2, o fato das expressões formulaicas serem um tipo de multpalavras influencia na dificuldade de estabelecer um método para extrair automaticamente todas as expressões formulaicas de um corpus. Como consequência, alguns estudos realizaram a extração de forma manual e, nos estudos em que uma extração automática aconteceu, sugere-se a verificação da saída por um ou mais humanos, sendo recomendado estimar a concordância entre os anotadores envolvidos, porque ainda entre humanos há divergências na hora de identificar se uma dada expressão é ou não formulaica (Ramisch, 2012).

Para a língua portuguesa, expressões formulaicas foram categorizadas segundo os componentes retóricos presentes em um resumo acadêmico no projeto *SciPo* (Feltrim, 2004). No projeto levantaram-se manualmente listas de expressões formulaicas para cada componente retórico do resumo. A tarefa que se desejava desenvolver, com a anotação manual das expressões formulaicas no *SciPo*, era identificar automaticamente o componente retórico da sentença, considerando, entre outros atributos, a presença de expressões formulaicas. Posteriormente, o processo de classificação das sentenças dos resumos aca-

dêmicos foi automatizado com a ferramenta *JBow* (Machado, 2009), que classifica as sentenças, considerando as expressões contidas, porém o *JBow* não identifica automaticamente expressões formulaicas. A ferramenta implementa um modelo estatístico de *bag of words* para classificar sentenças segundo as diferentes categorias retóricas, mas a ferramenta precisa de um corpus com expressões formulaicas previamente anotado para classificar novas sentenças.

Na literatura foram identificadas as seguintes abordagens para se extrair automaticamente expressões formulaicas:

- frequência de n-gramas;
- medidas de associação;
- *phrase-frames*;
- informações sintáticas e semânticas;
- a informação dos itens empregados pelos nativos da língua.

Os métodos computacionais adotados, assim como a metodologia para extrair e avaliar expressões formulaicas se sobrepõe com os métodos para extrair multipalavras (Seção 3.2).

A abordagem baseada na frequência de n-gramas é um método simples que identifica sequências que acontecem com uma alta frequência. Ele identifica sequências tais como "*how are you*", "*nice day today*" e "*good to see you*", porém também identifica sequências tais como "*and of the*" ou "*but it is*" que não são expressões formulaicas. Por outro lado, expressões que acontecem com baixa frequência no corpus, tais como "*blue moon*", "*longitude and latitude*", que representam colocações, não serão levantadas com o emprego unicamente da abordagem estatística. Em contraste, medidas de associação conseguem identificar expressões fixas com baixa frequência. A medida de associação mais empregada para extrair expressões formulaicas é a *Pointwise mutual information*. A equação para calcular essa e outras medidas de associação aparecem no Apêndice A desta monografia. *Phrase-frames* agrupa sequências que são muito similares com outras, por exemplo, "*it is * to*" é um padrão que aceita diversas palavras no lugar do símbolo * (*interesting, good, useful*). Desta forma, extrai expressões com variações, essas variações podem ser recuperadas por meio de padrões morfossintáticos ou semânticos. No exemplo, todas as palavras são adjetivos e todas elas têm o mesmo significado. Outros autores aproveitam os padrões sintáticos e as similaridades semânticas para

levantar expressões formulaicas dos textos. As abordagens anteriores nem sempre mostram o uso correto da língua, em outras palavras, o uso que o nativo faz da língua, por isso é importante contrastar as expressões candidatas com as expressões utilizadas pelos nativos, como realizado nos trabalhos de Ellis et al. (2008) e O'Donnell et al. (2013). Além disso, as vezes é preciso extrair as expressões de um determinado gênero, porque uma expressão frequentemente empregada na fala, por exemplo, provavelmente não é usada na escrita acadêmica, que é um registro mais formal.

Leedham (2012) e Simpson-Vlach e Ellis (2010) trabalharam com a abordagem baseada em n-gramas para extrair expressões formulaicas de textos produzidos por aprendizes de L2. Todos os estudos consideraram n-gramas contidos por 3, 4 e 5 palavras. Os bigramas em geral são excluídos, pois aparecem com muita frequência e, além disso, sobrepõem-se com os n-gramas maiores. Todos os n-gramas incluídos na lista de candidatos são expressões que nem sempre representam expressões formulaicas, por isso, em ambos os estudos foi necessário implementar um filtro para manter na lista as expressões com maior probabilidade de serem formulaicas.

Prentice (2012) apresenta um método para auxiliar os professores em sala de aula na extração de expressões formulaicas de textos produzidos por aprendizes de L2. Desta forma, os professores podiam extrair as expressões e analisar como os aprendizes as usavam. O autor relata que um grande número de expressões formulaicas acontece no começo da sentença. Depois de extrair as expressões, o autor ordena a lista de candidatas segundo a frequência das expressões no corpus. Como as expressões foram extraídas do começo das sentenças, na lista inicial existiam muitas sobreposições entre as expressões, por exemplo: "*I think it is*" aparecia no topo da lista de 4-gramas, enquanto "*I think it*" aparecia no topo da lista dos 3-gramas. Além disso, como as expressões foram somente ordenadas segundo sua frequência, apareciam nas listas muitas expressões que não eram consideradas expressões formulaicas, pois careciam de significado, isto é, não tinham sentido sozinhas. Considerando os dois últimos fatos, o autor propôs uma verificação manual das listas geradas, mantendo na lista as expressões formulaicas. O método proposto é bastante simples, mas segundo o autor é melhor que analisar manualmente todas as sentenças do corpus de alunos.

Simpson-Vlach e Ellis (2010) criaram uma lista de termos acadêmicos que incluía expressões formulaicas. Os autores identificaram as expressões que aconteciam com maior frequência nos gêneros acadêmico e da língua geral,

considerando uma análise do logaritmo de probabilidade das expressões em ambos os gêneros. As expressões de ambos os corpúscos foram ordenadas segundo o logaritmo da probabilidade obtendo duas listas com aproximadamente 2.000 elementos cada. Os elementos que coocorriam nas duas listas simultaneamente pertenciam à escrita de língua geral, enquanto os que unicamente apareciam na lista de termos acadêmicos pertenciam a esse gênero.

Para identificar as expressões formulaicas, Simpson-Vlach e Ellis (2010) combinaram as frequências das expressões e o valor da informação mútua e selecionaram 108 expressões com diferentes faixas de valores de frequência e de informação mútua (alta, média, baixa). Uma inspeção manual determinou a classe das expressões (se formulaicas ou não). A partir da anotação manual foi estabelecido um limiar que combina a frequência e a informação mútua para ordenar as expressões. Segundo os autores, tanto a frequência quanto a informação mútua foram decisivas para determinar se as expressões são formulaicas ou não, mas a informação mútua foi mais determinante na hora de ordenar a importância dessas expressões. A conclusão anterior indica que as associações das palavras dentro da expressão são informações decisivas. Para calcular a frequência e a informação mútua, Simpson-Vlach e Ellis (2010) utilizaram a ferramenta *Collocate* (Barlow, 2004).

O método proposto por Simpson-Vlach e Ellis (2010) é mais robusto que o método apresentado por Prentice (2012), que somente usa a frequência como parâmetro para extrair as expressões formulaicas, porque além de usar a frequência das expressões utilizam a informação mútua para avaliar a associação entre as palavras que compõem a expressão. Segundo McDonough (2011) uma análise baseada unicamente na frequência das expressões não é válida porque muitas das expressões que acontecem frequentemente no corpúscos não são expressões formulaicas e por outro lado muitas das expressões formulaicas não acontecem frequentemente. Desta forma, na literatura observa-se uma combinação das abordagens anteriores.

Kermes (2012) propôs um método para levantar n-gramas com 4 palavras. O método de Kermes (2012) tinha a finalidade de reconhecer os usos das expressões formulaicas identificadas; para isso a extração aconteceu com textos de diferentes áreas do conhecimento. O autor mostrou que existem expressões formulaicas que são específicas de determinado gênero textual, enquanto outras são comuns para todos os gêneros textuais. Para extrair a lista de expressões candidatas o autor estabeleceu uma frequência de corte e para descrever o uso da expressão utilizou heurísticas que consideravam as posi-

ções das expressões nas sentenças (começo, meio e fim) e o contexto ao redor da expressão. Na descrição do método proposto por Kermes (2012) não há informação sobre a metodologia de avaliação adotada.

Morley (2015) disponibilizou um recurso¹⁶ na forma de *Phrase-frames* que contém inúmeras expressões da escrita acadêmica em inglês, utilizadas frequentemente em diversas áreas de conhecimento. O recurso foi projetado para auxiliar a escrita dos aprendizes de inglês, porém, também é de grande utilidade para os nativos. As expressões foram extraídas mediante análises de colocações, estimativas de frequência e foram analisadas por vários linguistas. As expressões foram classificadas segundo a seção no texto em que elas acontecem: introdução, metodologia, etc.. Embora o recurso seja de grande valor, ele é monolíngue e não mostra as correspondências entre uma língua nativa e o inglês.

O método proposto por Abdalla e Teufel (2006) levantou expressões formulaicas compostas por um verbo e um objeto direto. Com ajuda de um parser e a combinação de várias heurísticas os autores extraíram expressões, tais como: "*we present an approach for*" ou "*we introduce a method for*". Abdalla e Teufel (2006) empregaram uma abordagem *bootstrapping* para a partir de tuplas com verbos e substantivos, por exemplo: [*present, approach*] extrair de um corpus expressões com a mesma estrutura sintática e similares semanticamente. O método estabelecia similaridades semânticas entre os componentes das tuplas dadas como entradas, as quais foram compiladas manualmente. Desta forma, expressões que não foram observadas no conjunto de treinamento podem ser identificadas e todas elas estavam agrupadas pelo sentido, por exemplo: o objetivo do autor. O método proposto por Abdalla e Teufel (2006), embora simples, obtém bons resultados, no entanto, é um método monolíngue que não visa auxiliar a escrita dos aprendizes da língua.

Na Seção 3.2 foram discutidas as dificuldades em estabelecer correspondências entre as multipalavras de uma língua fonte e alvo. Esse problema foi analisado por Fernandez (2008) que realizou um estudo manual que mostra as dificuldades para traduzir expressões formulaicas. Fernandez (2008) utiliza um corpus paralelo inglês e espanhol para realizar a análise. Para traduzir as expressões do inglês (língua fonte) para o espanhol (língua alvo) foram identificadas, previamente, 232 expressões formulaicas de diversos tipos. Esse número representa aproximadamente o 6% cada 1000 palavras. A descoberta anterior mostrou que é significativo o número de expressões

¹⁶<http://www.phrasebank.manchester.ac.uk/>

no corpus e, como consequência, é importante realizar análises que aprimorem a tradução dessas expressões. Do total de expressões identificadas em inglês 62 delas foram traduzidas incorretamente para o espanhol. Segundo Fernandez (2008), a quantidade de erros está correlacionada com o tipo de expressões que foi traduzida. Além disso, um dos principais problemas identificados durante a tradução é que as expressões formulaicas em inglês nem sempre foram traduzidas para expressões formulaicas em espanhol, muitas vezes isso aconteceu devido à falta de uma equivalência na língua alvo. A partir desse resultado, a autora considera importante que o tradutor antes de traduzir reconheça que a expressão a ser traduzidas é uma expressão formulaica. Porém, a identificação dessas sequências é extremamente difícil porque são muitas as características que as definem. A análise guiada por Fernandez (2008) mostra a dificuldade em traduzir expressões formulaicas mesmo entre humanos, o que indica que os valores de precisão e cobertura obtidos por um método automático devem estar longe dos valores alcançados pelos humanos.

3.4 Considerações Finais

A revisão bibliográfica apresentada neste capítulo mostrou que a tarefa de geração de léxicos bilíngues é complexa, mas diversos métodos estão disponíveis. Em geral, todos os léxicos comentados foram criados para uma aplicação específica, por isso, comprovou-se que tanto a utilidade do léxico, quanto os itens incluídos determinam a abordagem computacional que será utilizada para gerar o recurso lexical.

Os métodos de tradução se mostraram eficientes para a criação destes recursos, no entanto a maioria deles depende de um corpus paralelo. Além disso, o sucesso do método depende do tamanho do corpus. Adicionalmente, se o desejo é extrair léxicos de um gênero específico, por exemplo, o acadêmico, é preciso que o corpus paralelo seja desse gênero. Corpus paralelos são recursos difíceis de compilar, o que representa uma das principais dificuldades na criação de um léxico bilíngue com boa cobertura. Para o português do Brasil, são poucos os corpus paralelos disponíveis, sendo menor ainda o número de corpus para o gênero acadêmico. Até o momento, o único projeto conhecido com interesse de disponibilizar corpus da área técnico-científica é o projeto *Per-Fide*¹⁷, que tem como principal objetivo a criação de recursos bilíngues entre a língua portuguesa e outras seis línguas: espanhol, russo,

¹⁷<http://per-fide.ilch.uminho.pt/site.pl/index.it>

francês, italiano, inglês e alemão. Além disso, o projeto prevê a compilação de corpus paralelos de diferentes áreas (Araújo et al., 2010).

As etapas propostas para compilar expressões multipalavras foram observadas na compilação de vários dos trabalhos analisados. Desta forma, as etapas podem ser adotadas para compilar os léxicos bilíngues entre o espanhol e o português. Além disso, alguns trabalhos primeiramente extraem os léxicos da língua fonte e depois estabelecem as equivalências, separando esse processo em duas etapas. Seguindo essa linha, o passo de obter as equivalências entre as línguas poderia ser inserido no conjunto de passos estabelecidos no procedimento de extrair multipalavras no contexto monolíngue.

A extração de multipalavras se mostrou como uma tarefa complexa, sendo necessária a avaliação humana na maioria dos estudos. Acredita-se que nesta pesquisa a extração de itens lexicais é uma das etapas mais críticas, em todos os recursos será necessário criar heurísticas para reconhecer: falsos cognatos, marcadores discursivos e expressões formulaicas. O léxico de verbos é o mais simples, pois um anotador morfossintático realiza a tarefa de forma automática, porém erros podem acontecer, sendo necessária a verificação da saída.

Observou-se que a etapa mais crítica dos métodos comentados é a avaliação dos recursos. Sendo necessário, em alguns casos ter conjuntos previamente avaliados ou construídos pelos humanos ou avaliar manualmente a saída obtida automaticamente. A maioria dos trabalhos utilizou a avaliação humana para validar os métodos, isso é, sem dúvida, um dos maiores problemas enfrentados nesta pesquisa. No caso das expressões é necessário primeiramente validar a etapa da extração, na língua fonte, para depois validar se a equivalência obtida foi válida ou não.

Ferramentas de Apoio à Escrita de Textos

Neste capítulo são apresentadas ferramentas e métodos, propostos para auxiliar a escrita de aprendizes (Seção 4.1) as quais foram classificadas, segundo Hayashibe et al. (2012), nas seguintes categorias: (i) sistemas de pós-edição de textos; (ii) sistemas de auxílio em tempo real; (iii) sistemas que procuram por frases; e (iv) sistemas que apoiam a tradução. Na Seção 4.2 comentam-se as principais abordagens utilizadas para tratar diversos tipos de erros gramaticais. Essas abordagens integram os sistemas de pós-edição de textos, cuja finalidade é detectar e corrigir erros. Nesta mesma seção também é apresentado o corretor gramatical ReGra, o qual é o mais utilizado para produzir textos em português brasileiro. Para finalizar, comentam-se as metodologias empregadas para avaliar as abordagens computacionais que detectam e corrigem erros.

4.1 Ferramentas e Métodos para Auxiliar a Escrita de Aprendizes

O processamento de línguas naturais (PLN) é relevante no contexto de ensino/aprendizado de língua, especialmente com relação à língua escrita. Neste cenário, o PLN tem sido utilizado para analisar a linguagem produzida pelos aprendizes de uma língua, via métodos dedicados à criação de aplicações

ICALL (*Intelligent Computer-Assisted Language Learning*), AES (*Automated Essay Scoring*) e ferramentas de auxílio à escrita. Além disso, o PLN também se ocupou da análise dos textos dos nativos com o intuito de apoiar a escrita dos aprendizes, por meio da apresentação de exemplos, ou criação de materiais, produzidos pelos nativos, que auxiliem a aprendizagem dos aprendizes (Meurers, 2013).

Pesquisas que criam ferramentas e recursos para auxiliar à escrita de textos utilizam abordagens da linguística computacional e da teoria de aprendizagem de língua para alcançar seus objetivos. Em geral, a primeira etapa desses estudos está focada na análise da língua produzida pelos aprendizes para identificar suas características. Na maioria das vezes, esta análise concentra-se na identificação de diversos tipos de erros, porém há muitos trabalhos que avaliam aspectos estilísticos dos textos que são considerados inadequações e não erros. Para a realização das análises da língua dos aprendizes geralmente um corpus de aprendizes é compilado. A partir de diversas análises e/ou anotações são levantadas as principais dificuldades. A seguir, são apresentados sistemas de auxílio à escrita, que seguiram a metodologia anterior para criar os auxílios. Esses sistemas estão categorizados segundo a classificação proposta por Hayashibe et al. (2012).

Sistemas de Pós-Edição de Textos

Os sistemas de pós-edição de textos recebem um texto como entrada e procuram por problemas pontuais, especificamente relacionados com inadequações gramaticais. Esses estudos integram corretores ortográficos, gramaticais e estilísticos e são úteis porque oferecem alternativas para melhorar aspectos de clareza e concisão nos textos (Aluísio, 1995). Para construir esses sistemas é necessário analisar grandes corpus de aprendizes para conhecer as características dos aprendizes. Além disso, diversas abordagens computacionais são combinadas para tratar automaticamente a gama de problemas gramaticais. Algumas dessas abordagens são apresentadas na Seção 4.2. Neste grupo estão incluídos vários sistemas, livres ou comerciais, que apoiam à escrita em inglês, tais como: *Grammarly*¹, *WhiteSmoke*², e *Ginger*³.

Leacock et al. (2009) apresentaram o sistema *ESL Assistant* que é uma ferramenta Web que auxilia os aprendizes de inglês a produzirem textos sem erros gramaticais. Além disso, também disponibiliza uma busca na Web que mostra o uso adequado de expressões, apresentando a expressão original e

¹<https://www.grammarly.com/>

²<http://www.whitesmoke.com/>

³<http://www.getginger.jp/>

possíveis substituições. *ESL Assistant* utiliza uma abordagem híbrida que combina métodos estatísticos e baseados em regra para detectar e corrigir, por meio de sugestões, problemas gramaticais. Métodos de aprendizado de máquina foram propostos para tratar erros mais complexos, tais como erros de uso de artigos e preposições. Por outro lado, expressões regulares são utilizadas para problemas mais simples, relacionados com verbos modais, por exemplo. Adicionalmente, a ferramenta utiliza um modelo de língua suficientemente grande para verificar a validade da sugestão. A sugestão não é apresentada ao usuário até que a probabilidade dela não seja maior que o trecho original. Mais detalhes da arquitetura do sistema *ESL Assistant* são apresentados nos trabalhos de Gamon et al. (2008, 2009). Alguns comentários referentes a esses trabalhos são apresentados na Seção 4.2.

Nesta categoria são incluídos os sistemas que usam técnicas de PLN para avaliar e pontuar a qualidade da escrita de um texto. Um exemplo desse tipo de sistemas é o *Criterion (Online Essay Evaluation Service)*⁴ (Burstein et al., 2004), um sistema *Web* que além de avaliar as habilidades da escrita dos usuários também fornece uma nota e um conjunto de sugestões para o texto. *Criterion* se baseia no sistema *e-rater* (Attali e Burstein, 2006; Burstein et al., 2003) para pontuar os textos e contém uma aplicação, chamada *Critique*, que inclui um grupo de programas para identificar erros gramaticais, de usos da língua, de estruturação do discurso e de inadequações de estilo nos textos (Marquiefável, 2007; Attali e Burstein, 2006; Dikli, 2006). A nota dada pelo *Criterion* é obtida após uma avaliação geral dos problemas identificados. Apesar de o sistema considerar uma pontuação holística, ele consegue prover correções para as deficiências encontradas. O sistema atende textos de diferentes gêneros, como o narrativo, expositivo e argumentativo, no entanto, não tratam o gênero científico.

O *Criterion* e *ESL Assistant* foram avaliados por Chodorow et al. (2010) na tarefa de detectar e corrigir erros de preposições e artigos. Ambos os sistemas mostraram-se eficientes na apresentação das sugestões, contribuindo para aprimorar a escrita dos aprendizes de inglês. Por isso, tanto *Criterion* como *ESL Assistant* são o estado da arte de sistemas de pós-edição de textos.

Sistemas de Auxílio em Tempo Real

Há um conjunto de ferramentas que auxiliam a escrita dos textos em tempo real, isto é, durante a tarefa de escrita. Essas ferramentas ajudam na composição do texto, indicando as palavras e/ou expressões em L2 mais adequadas

⁴<http://criterion.ets.org>

para determinando contexto. Segundo Hayashibe et al. (2012), o auxílio em tempo real pode ser uma solução mais atraente para detecção e correção de erros de aprendizes, pois consegue prever alguns erros enquanto o texto está sendo escrito. AI-type⁵ é um software de assistência as entradas em inglês, que ajuda os usuários a digitar mais rapidamente, apresentando possíveis palavras. Estas palavras são apresentadas considerando o contexto em que foram utilizadas.

Sistemas de auxílio em tempo real são propostos para aprendizes que não necessariamente têm um bom nível na escrita em L2. Esses sistemas integram diversos recursos, tais como: dicionários, métodos de busca de expressões, regras gramaticais e exemplos de sentenças usadas na língua. A integração dos recursos em uma ferramenta diminui o tempo de busca e facilita as consultas dos aprendizes. As ferramentas de auxílio em tempo real, apresentadas nesta seção, utilizam a língua nativa do aprendiz como ponto de partida e, para isso, empregam métodos de tradução automática. Além disso, oferecem auxílios no nível lexical. Diferentemente dos sistemas de pós-edição da escrita, que focam na detecção de erros, os sistemas de auxílio em tempo real tentam prever o erro, auxiliando na produção do texto.

As ferramentas *PENS* (*Perfect ENGLISH System*) (Huang et al., 2000) e *WINGS* (*Writing with INtelligent Guidance and Suggestions*) (Dai et al., 2014) apóiam a produção escrita de chineses, aprendizes de inglês. Ambas as ferramentas oferecem ajuda e sugestões em tempo real na escrita nos níveis lexicais e sentenciais. Tanto *PENS* quanto *WINGS* podem ser considerados sistemas de recuperação de informação que apóiam a escrita. Para isso, eles disponibilizam grandes bases de conhecimento e mecanismos de busca para auxiliar o aprendiz na tarefa de escrita de textos.

PENS é um sistema de consulta que, por meio da interação com o usuário, apresenta sugestões relacionadas com a grafia das palavras e a reescrita de sentenças a partir de sentenças de exemplos recuperadas automaticamente. O sistema *PENS* oferece sinônimos e antônimos de palavras na língua inglesa e, para isso, consulta léxicos disponíveis. Adicionalmente, *PENS* permite a consulta de palavras na língua chinesa, procurando a tradução mais adequada para o inglês. Para isso, o sistema utiliza um modelo de tradução estatística, dicionários bilíngues chinês inglês e um modelo da língua inglesa. O modelo de língua em inglês tem 247.238.396 trigramas e o dicionário bilíngue contém 115.200 entradas com anotação morfosintática e classificação semântica. O

⁵<http://www.aitype.com/>

modelo de tradução foi treinado a partir de um corpus paralelo composto por 96.362 pares de sentenças.

Huang et al. (2000) compilaram recursos lexicais, tais como: léxicos bilíngues e listas de colocações a partir do alinhamento de um corpus paralelo para gerar os auxílios no nível lexical. Esses recursos foram avaliados manualmente. Por outro lado, Dai et al. (2014) implementaram um modelo de representação vectorial das palavras, que usa como base a similaridade entre as palavras e estratégias baseadas em regras com padrões morfossintáticos, para mostrar ao usuário itens lexicais que podem substituir a palavra dada como entrada.

As ferramentas *PENS* e *WINGS* também apóiam a reescrita de sentenças. Para isso, Huang et al. (2000) empregam um algoritmo de recuperação de informação que, por meio de palavras ou expressões curtas, procuram no corpus paralelo as sentenças relacionadas. As palavras chaves são dadas como entrada pelo usuário e as sentenças recuperadas compartilham o mesmo sentido ou a mesma sintaxes das palavras de entrada. Dai et al. (2014) recomendam sentenças relacionadas com uma sentença dada como parâmetro, empregando o modelo de tópicos *Latent Dirichlet Allocation*⁶ (LDA).

Dai et al. (2014) avaliaram o desempenho do sistema proposto na apresentação de palavras similares por meio da comparação com outra ferramenta. O resultado da avaliação mostrou maior diversidade nas palavras apresentadas por Dai et al. (2014). Além disso, o sistema também foi avaliado na apresentação de sentenças relacionadas; sendo realizada uma verificação manual da saída por 12 estudantes aprendizes de inglês. Somente foram consideradas as saídas em que houve mais de 70% de concordância entre os avaliadores, conseguindo uma precisão de 76%. Já, a avaliação realizada por Huang et al. (2000) considerou a precisão do algoritmo de tradução proposto para obter as palavras em inglês a partir das palavras em chinês. A recuperação sentencial foi avaliada comparando a saída do sistema com um *gold standard*, criado de forma manual. A precisão do sistema foi de 83%, enquanto a cobertura foi de 55%. Embora os dois sistemas ofereçam auxílios similares, o sistema *WINGS* pode prover melhores sugestões na recuperação das sentenças similares, pois introduzem técnicas de PLN que avaliam o contexto das expressões para selecionar o conjuntos de sentenças.

⁶LDA é um modelo gerativo probabilístico, em que os documentos são representados como misturas aleatórias de tópicos latentes, de forma que cada tópico é caracterizado segundo a distribuição das palavras presentes no texto. Assim, cada documento é representado segundo a distribuição dos tópicos.

Hayashibe et al. (2012) apresentaram a ferramenta *phloat* (*PHrase LOokup Assistant Tool*), similar aos sistemas *PENS* e *WINGS*, porém *phloat* foi implementada para auxiliar nativos do japonês, em etapas iniciais de aprendizagem, nas escolhas lexicais em inglês. Para isso, o sistema usa uma grande base de conhecimento composta por palavras e expressões em japonês e suas traduções em inglês. O sistema foi validado por meio de tarefas desenvolvidas pelos aprendizes, com e sem a ferramenta *phloat*. Dois humanos avaliaram o desempenho dos aprendizes, concluindo que os textos produzidos com a ferramenta tinham mais qualidade.

*FLOW*⁷ (Chen et al., 2012) apresenta um modelo que combina tradução estatística com modelos de desambiguação sensíveis ao contexto para obter os n-gramas mais adequados dada a língua nativa (processo de composição) ou a segunda língua (processo de revisão). A análise se baseou em um corpus paralelo alinhado, por meio do Giza++. *FLOW* apresenta aos chineses aprendizes de inglês as expressões mais adequadas para um dado contexto. Por exemplo: dada a expressão "*We propose a method to*" o sistema prediz as expressões que vem depois "*solve the problem*", "*tackle the problem*". A vantagem da abordagem proposta é que ela somente depende do corpus paralelo, que embora seja um recurso difícil de se compilarr, caso disponível o recurso, o método pode ser adaptado para auxiliar em outros gêneros, como o científico, por exemplo. Os autores nada comentaram sobre a avaliação do recurso.

Buscadores de Expressões

Nesta categoria incluem-se um conjunto de estudos que visam criar recursos com expressões pré-fabricadas, que podem auxiliar na produção escrita da língua estrangeira. No entanto, em geral, estes recursos são utilizados por usuários com níveis avançados da língua, que conhecem a importância de usar expressões pré-fabricadas. Além disso, não consideram possíveis erros produzidos por interferências da língua nativa, como os sistemas de auxílio em tempo real. Diversos estudos têm sido desenvolvidos para criar recursos que levantem padrões lexicais ou gramaticais do inglês. Morley (2015) disponibilizaram um recurso⁸ na forma de *Phrase-frames* que contém inúmeras expressões da escrita acadêmica em inglês, utilizadas frequentemente em diversas áreas de conhecimento. Por outro lado, Wible e Tsao (2010) criaram a base de conhecimento léxico gramatical *StringNet*, baseada na extração de n-gramas híbridos, que consideram a coocorrência da forma, do lema e da eti-

⁷<http://flowa1demo.appspot.com>

⁸<http://www.phrasebank.manchester.ac.uk/>

queta morfossintática das sequências de palavras. Esses recursos apresentam os padrões do inglês, tais como: “*there is a tendency for [noun] to [verb]*”, por meio dos quais, os aprendizes podem fazer substituições e gerar expressões válidas na língua. Esse tipo de recurso pode ser muito útil na aprendizagem de língua estrangeira e em tarefas de tradução.

Sistemas de Suporte à Tradução

Nesta categoria são incluídos sistemas que interagem com o usuário para produzir um texto na língua alvo a partir de um texto na língua fonte. Esses sistemas, diferentemente dos tradutores tradicionais, conseguem apresentar uma tradução mais precisa, pois contam com a ajuda do usuário para aprimorar a saída. Eles surgiram com o intuito de substituir, durante a produção de textos em L2, as consultas a tradutores ou dicionários bilíngues que nem sempre apresentam a melhor tradução. Os sistemas também oferecem auxílio em tempo real, mas focando unicamente na obtenção de traduções adequadas.

Os sistemas *TransType2* (Esteban et al., 2004) e *TransAhead* (Huang et al., 2012) sugerem palavras ou frases candidatas para tradução na língua alvo, considerando uma sentença na língua fonte. *TransType2* é uma ferramenta que observa um tradutor automático e um modelo de língua para prever a palavra ou expressão em L2 que segue após uma digitação do usuário. O usuário pode aceitar as sugestões dadas pela ferramenta ou pode continuar escrevendo até construir o texto. As sugestões podem melhorar a produtividade na produção do texto, pois, caso aceitas, diminui o tempo de digitação e contribui para melhorar a qualidade do texto. *TransAhead* foi avaliada em sala de aula na tarefa de gerar textos em inglês por chineses. Os textos gerados com o uso da ferramenta apresentaram melhor qualidade do que os textos produzidos sem a ferramenta ou com auxílio do tradutor *Google*.

A seguir são apresentados métodos para detectar e corrigir diversos tipos de erros. A maioria desses métodos é utilizada para dar auxílio aos sistemas de pós-edição de textos que são os que focam na detecção e correção de erros, no entanto, os outros tipos de sistemas também podem incorporar módulos para detectar e corrigir erros.

4.2 Detecção e Correção de Erros Gramaticais

Para detectar e corrigir erros gramaticais têm sido propostos desde métodos simples, que tratam um único problema, até abordagens mais robustas que apoiam a escrita dos textos em diversos níveis. O objetivo dos trabalhos que

são apresentados aqui é detectar erros gramaticais produzidos por aprendizes da língua, por isso, será observado como nestes estudos foram combinadas análises com textos de aprendizes e nativos, seja para criar os auxílios ou para avaliar os métodos propostos.

Segundo Leacock et al. (2010), os erros gramaticais (ou sintáticos) estão relacionados com a gramática da língua, com os mecanismos ou usos da mesma. Os erros de ortografia podem ser considerados nesta categoria de erros, no entanto, na maioria das vezes eles são tratados em etapas de pré-processamento.

Os primeiros esforços para detectar erros gramaticais de nativos e não nativos foram baseados em estudos puramente linguísticos que usavam gramáticas e *parsers* sofisticados. Essas abordagens produziram análises sintáticas completas dos textos de entrada, sendo possível identificar diferentes tipos de erros. As gramáticas são recursos que contêm regras da sintaxe e da morfologia das línguas, mas é comum que elas não especifiquem aspectos de uso da língua, por exemplo, indicar que preposição é adequada em um determinado contexto. Esse é um dos fatos que indicam que as gramáticas não são a melhor alternativa para detectar erros de aprendizes, que na maioria das vezes cometem erros de uso da língua (Leacock et al., 2010). A maioria das abordagens iniciais envolveu processos manuais para criar tanto regras quanto gramáticas. Com o avanço de técnicas de processamento de línguas naturais, as abordagens baseadas em regras deram lugar a métodos baseados em aprendizado de máquina, usando *córpus* (Leacock et al., 2010).

Os estudos baseados em *córpus* utilizam algoritmos de classificação para determinar a palavra correta em um determinado contexto. Essa abordagem tem sido muito utilizada para detectar e corrigir erros de artigos e preposições. Adicionalmente, há outra linha de pesquisa que emprega modelos de língua com a finalidade de detectar diversas categorias de erros. Nesses trabalhos, os erros são localizados nas regiões em que a probabilidade do modelo de língua é muito baixa.

Estas técnicas serão exemplificadas no decorrer desta seção, em que vários estudos serão comentados. Entretanto, segundo Leacock et al. (2010), nenhuma destas abordagens consegue identificar sentenças com mais de dois ou três erros gramaticais. Comumente, os aprendizes produzem mais de um erro por sentença, enquanto o número de erros dos nativos é menor. Esse fato, em conjunto com a constatação de que os erros dos aprendizes não são os mesmos que os erros dos nativos determinam que um sistema criado para detectar erros de nativos, pode não ser suficientemente eficiente na detecção

de erros de aprendizes. Como alternativa de solução, segundo Leacock et al. (2010), propõe-se a construção de sistemas que focam nos principais erros que nativos de uma língua cometem, e esses sistemas são estendidos para tratar erros produzidos por aprendizes. Segundo Dahlmeier e Tou Ng (2011) e Nagata et al. (2011), são poucos os métodos que detectam e corrigem erros gramaticais baseados, unicamente, em corpus de aprendizes previamente anotados. Em geral, os autores apóiam-se em corpus de nativos para construir os modelos e depois combinam esses modelos com as anotações humanas realizadas no corpus de aprendizes.

Seguindo essa linha, a detecção automática de erros é guiada por anotações dos erros nos corpus de aprendizes. Contudo, esta tarefa é demorada e requer pessoal treinado. Além disso, para ser um recurso útil na identificação automática dos erros, o corpus precisa ser anotado com um grande número de erros, o que requer a anotação de um grande volume de dados. Em geral, os estudos que empregam modelos de língua para apoiar a tarefa de detectar e corrigir erros tentam driblar a necessidade do corpus de aprendizes. Alternativamente, para tratar a deficiência anterior aplica-se a geração artificial de erros.

Os erros mais estudados na literatura são: uso inadequado de artigos (Chodorow e Leacock, 2000; Dahlmeier e Tou Ng, 2011; De Felice, 2008; Andersen, 2007; Genoves et al., 2007; Nagata et al., 2011) e uso inadequado de preposições (Chodorow e Leacock, 2000; Chodorow et al., 2007; Dahlmeier e Tou Ng, 2011; Tetreault et al., 2010; Andersen, 2007; De Felice, 2008; Nagata et al., 2011). No caso das inadequações das preposições é comum avaliar um conjunto de preposições e não todas, por exemplo, Dahlmeier e Tou Ng (2011) trabalharam com as 36 preposições mais frequentes do inglês. Outros erros que também têm sido tratados são: concordância das palavras (concordância nominal e verbal) (Andersen, 2007; Foster e Vogel, 2004; Chodorow e Leacock, 2000); erros léxicos relacionados com palavras ausentes ou duplicadas; erros de grafia (Gagner et al., 2007; Andersen, 2007); colocações (Hua chen et al., 2010); tempos verbais (Andersen, 2007; Nagata et al., 2011); diversos problemas no uso dos substantivos, tais como: número e complementos (Nagata et al., 2011); erros na ordem das palavras (Andersen, 2007); uso incorreto da função morfossintática de palavras (Chodorow e Leacock, 2000) e sentenças mal formadas (Gagner et al., 2007; Andersen, 2007; Foster e Vogel, 2004). A maioria dos trabalhos concentram-se nos erros mais frequentes de aprendizes de inglês como língua estrangeira ou segunda língua. A seguir, são comenta-

dos alguns dos trabalhos que dedicam seus esforços à identificação e detecção de erros ortográficos e gramaticais.

A detecção de erros ortográficos consiste em identificar palavras incorretas que não pertencem ao vocabulário da língua ou que foram utilizadas em um contexto inadequado. Tradicionalmente, métodos para detectar erros ortográficos realizam buscas em dicionários para detectar os erros (Dale et al., 2012). Quando a palavra não aparece no dicionário é apontado o erro. Um método comumente aplicado é o algoritmo baseado em n-gramas, neste caso n-gramas são sequências de letras que compõem a palavra (Fossati e Di Eugenio, 2008). A análise procura por n-gramas que não são válidos na língua. Além disso, análises morfológicas (Shaan et al., 2010) e métodos de aprendizagem de máquina (Gale et al., 1992) têm sido propostos, ao mesmo tempo, em que a combinação de técnicas foram aplicadas para realizar a verificação ortográfica (Fossati e Di Eugenio, 2007).

Os métodos para corrigir erros de ortografia são categorizados em função de como a correção acontece: dependente do contexto ou independente do contexto. A correção dependente do contexto considera informações ao redor da palavra para corrigir os erros, enquanto os métodos independentes do contexto corrigem a palavra como uma palavra independente sem considerar o contexto em que aconteceu. A técnica mais comumente utilizada para corrigir erros de ortografia, independente do contexto, é a distância de *Levenshtein* (Levenshtein, 1966; Damerau, 1964a) que identifica a quantidade de operações (inserção, eliminação e substituição) que devem ser realizadas para transformar uma palavra errada em outra, incluída no dicionário da língua (Damerau, 1964b). Outras técnicas independentes do contexto são baseadas na criação de regras (Kukich, 1992) e na probabilidade de n-gramas (Kukich, 1992; Wilcox-O'Hearn et al., 2008). A abordagem baseada em regras utiliza algoritmos que representam padrões dos erros gramaticais em forma de regras, as quais são utilizadas para transformar uma palavra incorreta em uma correta. Além disso, os métodos baseados em n-gramas procuram os n-gramas mais adequados para corrigir o erro, considerando as probabilidades condicionais das letras no n-gramas.

Os métodos dependentes do contexto analisam o contexto do erro para sugerir a forma correta. Esta abordagem é comum para tratar erros de palavras reais, por exemplo, "esta" ou "está"; as duas palavras são corretas, mas o erro está associado ao contexto em que aconteceram. Algumas das técnicas uti-

lizadas são: conjuntos de confusão⁹; aprendizagem de máquina; semelhança fonética e distância semântica. O conjunto de confusão consiste em criar um conjunto de palavras, a partir de um dicionário, com palavras que podem ser utilizadas no lugar da outra. Por exemplo, "está", "esta" e "essa" estariam no mesmo conjunto (Carlson et al., 2001). Os métodos de aprendizagem de máquina são combinados com os conjuntos de confusão (Golding e Roth, 1999) para determinar a palavra mais adequada dentro do conjunto. Técnicas de semelhança fonética (Stuker et al., 2011) são baseados em erros fonológicos comuns, decorrentes da ortografia incorreta de pronúncia. O algoritmo *SOUNDEX* é o mais utilizado para estabelecer semelhanças fonéticas. Técnicas que avaliam a distância semântica (Budanitsky e Hirst, 2006) utilizam estruturas taxonômicas como a *WordNet*, por exemplo, para identificar, a partir da similaridade semântica, palavras que corrijam o erro.

Frequentemente, para detectar erros ortográficos usa-se um corretor ortográfico, tal como o GNU Aspell¹⁰, que baseado na distância de edição procura as palavras candidatas a correção (Dale et al., 2012). Essas palavras são ordenadas ou filtradas por meio de um modelo de língua. A probabilidade dos n-gramas contidos no modelo de língua é bastante usada para corrigir erros ortográficos, porque permite predizer a troca mais adequada para um erro em um contexto específico. Stehouwer e Van Zaanen (2009) analisaram diferentes estratégias para construir modelos de língua encarregados de determinar a melhor escolha de palavras incluídas em um conjunto de confusão, considerando o contexto. No entanto, o algoritmo não foi pensando para tratar erros de aprendizes. Por outro lado, Tsao e Wible (2009) apresentaram um método não supervisionado que, utilizando um modelo de n-grams e a medida de similaridade *Levenshtein* (Levenshtein, 1966; Damerau, 1964a), detectam e corrigem erros lexicais. Porém, a análise não foi testada com erros reais produzidos por aprendizes da língua.

Após conhecer as abordagens propostas para tratar erros ortográficos, a seguir serão comentados alguns trabalhos que detectam e corrigem erros gramaticais.

Os trabalhos comentados a seguir foram propostos para tratar erros produzidos por aprendizes de inglês de diversas línguas nativas. Um dos aspectos a destacar nestas pesquisas é o conjunto de dados utilizado para realizar as análises. Por exemplo, nas análises apresentadas por (Chodorow e Lea-

⁹do inglês *confusion sets*

¹⁰<http://aspell.net/>

cock, 2000), (Gagner et al., 2007), (Andersen, 2007) e (Foster e Vogel, 2004) utilizaram-se sentenças com erros e suas correções, correções que foram produzidas manualmente. Já, De Felice (2008) utilizaram um corpus de nativos e de aprendizes para avaliar o algoritmo. Tetreault et al. (2010) selecionaram sentenças produzidas por aprendizes da língua inglesa no TOEFL, as quais foram classificadas em três grupos: livres de erros; com um erro de preposição e com mais de um erro para testar o algoritmo. Finalmente, Dahlmeier e Tou Ng (2011) treinaram e testaram seu modelo em um corpus de aprendizes previamente anotado. O conjunto de dados utilizado mostra a necessidade de se criar conjuntos de avaliação que precisam ser analisados por humanos. As análises humanas são essenciais para construir algoritmos com bom desempenho, porém são demoradas e requerem pessoal treinado.

A seguir, são apresentadas algumas variações nos métodos computacionais e os atributos utilizados para tratar problemas gramaticais. No entanto, é observado em todos os estudos como o emprego de ferramentas computacionais, como *parsers* e anotadores morfossintáticos são decisivos na tarefa de detecção de erros, sendo preciso disponibilizar ferramentas capazes de lidar com erros de aprendizes, pois, caso contrário muitas das falhas dos algoritmos estão relacionadas com falhas dessas ferramentas (Tetreault et al., 2010).

Chodorow e Leacock (2000), De Felice (2008), Chodorow et al. (2007), Tetreault et al. (2010) e Dahlmeier e Tou Ng (2011) utilizaram o contexto local de uma palavra (uma janela equivalente a duas palavras a esquerda e a direita) para detectar os erros. O foco desses estudos consiste em extrair atributos que permitam identificar o erro.

De Felice (2008) utilizou os seguintes atributos para caracterizar o contexto da palavra: etiquetas morfossintáticas, medidas sintáticas e semânticas, relações semânticas entre palavras, obtidas a partir da *WordNet*¹¹ e entidades nomeadas. Todas essas características foram extraídas utilizando o *parser* CCG.¹² Os erros são identificados calculando-se a informação mútua entre os vetores de características da sentença analisada e uma sentença de exem-

¹¹É um grande banco de dados lexicais do Inglês. Substantivos, verbos, adjetivos e advérbios agrupados em conjuntos de sinônimos (*synsets*), cada um expressando um conceito distinto (FELLBAUM, 1998).

¹²O *parser* CCG (Clark e Curran, 2007), é um *parser* estatístico incluído no conjunto de ferramentas *CC Tools* (<http://svn.ask.it.usyd.edu.au/trac/candc/wiki>). A saída do *parser* CCG está no formato *RASP-style GRs* (ver (Briscoe et al., 2006)). A ferramenta *CC Tools* também inclui um etiquetador morfossintático, um analisador morfológico e um reconhecedor de entidades nomeadas. Todos os componentes da ferramenta *CC Tools* funcionam em *pipeline* permitindo a identificação e a extração de todos os predicados de interesse, que podem ser desde etiquetas morfossintáticas até dependências sintáticas.

plo. Avaliações no corpus de nativos e de aprendizes realizadas por De Felice (2008) indicaram que o método proposto é adequado para ser aplicado em um contexto de aprendizes da língua inglesa, mesmo que, segundo o autor, ele necessite de alguns refinamentos.

Chodorow et al. (2007), Tetreault et al. (2010) e Dahlmeier e Tou Ng (2011) trataram a tarefa de detecção de erros de preposições como um problema de classificação. Extraíram, com ajuda de um *parser*, um conjunto de atributos do contexto da palavra, para identificar erros de diversas categorias. As propostas apresentadas por Tetreault et al. (2010), De Felice (2008) e Chodorow et al. (2007) empregaram um corpus escrito por nativos para treinar o modelo responsável pela correção dos erros de preposições. O modelo é testado em um corpus de aprendizes, nos quais o número de erros é mais frequente. Por outro lado, Dahlmeier e Tou Ng (2011) apresentaram um método para correção de erros de artigos e preposições, treinando e testado unicamente no corpus de aprendizes *NUCLE*, que é anotado com diferentes tipos de erros.

O algoritmo de classificação empregado por Dahlmeier e Tou Ng (2011) é baseado no método de alternância da otimização da estrutura (ASO), do inglês *Alternating Structure Optimization*. Dahlmeier e Tou Ng (2011) definiram o problema de correção de erros gramaticais como um problema de classificação linear, em que procura-se identificar as posições com seus respectivos contextos, assim como a correção válida, em um corpus de nativos. Os autores representam os artigos e as preposições além de seus contextos em um vetor de características $x \in X$, enquanto as correções são classes $y \in Y$. Dahlmeier e Tou Ng (2011) combinaram os métodos de extração de características empregados por Tetreault et al. (2010), De Felice (2008), Chodorow et al. (2007) e Han et al. (2006). Obtiveram uma precisão de 87% na identificação dos artigos com o método de extração proposto por Han et al. (2006) e uma precisão de 68.25% com o método de extração de Tetreault et al. (2010).

Gagner et al. (2007), Andersen (2007) e Foster e Vogel (2004) apresentaram abordagens similares para detectar se uma sentença é gramatical ou não. Andersen (2007) e Foster e Vogel (2004) utilizaram sentenças com erros e suas correções para avaliação do método, enquanto Gagner et al. (2007) geraram erros artificiais, pois existia muita dispersão dos erros anotados no corpus. No método de Foster e Vogel (2004), o *parser* analisava os textos a partir de uma gramática convencional, caso aparecia uma sentença mal formada que não se correspondia com nenhuma regra da gramática, o *parser* utilizava um conjunto de regras, construídas manualmente após a correção das sentenças

erradas. Apesar do desempenho do algoritmo ser de 84%, ele não consegue detectar sentenças mal formadas com mais de um erro. Devido a esse problema, De Felice (2008) considera que esta abordagem não é útil para detectar erros em textos escritos por aprendizes.

Por outro lado, Gagner et al. (2007) combinaram uma análise superficial, baseada na frequência dos n-gramas e na função morfossintática dos componentes das sequências, com uma análise sintática mais profunda, realizada por meio do *parser* XLE¹³. As características extraídas foram utilizadas para treinar um algoritmo de aprendizado baseado em árvore de decisão. Andersen (2007) utilizou o algoritmo *Naive Bayes* para identificar sentenças com ou sem erros, representando as sentenças com os seguintes atributos: etiquetas morfossintáticas de cada palavra; lematização das palavras e relações gramaticais entre os itens da sentença. Esses atributos foram extraídas dos textos de treinamento empregando o *parser* RASP (*Robust Accurate Statistical Parsing*)¹⁴.

Além dos métodos comentados, há outro conjunto de estudos que detectam e corrigem erros utilizando um modelo de língua. Atwell (1997) empregou um modelo de língua baseado na anotação morfossintática dos textos para detectar erros. Gamon et al. (2009) combinaram um grande modelo de língua com um classificador para identificar erros de preposições, enquanto Hermet e Désilets (2009) criaram um modelo de língua a partir da contagem de n-gramas da web e o recurso foi empregado para corrigir erros. Hermet e Désilets (2009) propuseram um método monolíngue que baseado na probabilidades das n-gramas detectava erros de preposições. Além disso, apresentaram um método que combinava um modelo de língua com tradução automática, que detectava erros produzidos por interferência da língua nativa. O modelo baseado em tradução implementou a abordagem *Roundtrip*, que tenta identificar erros no processo de traduzir sentenças em L2, candidatas a erros, em sentença em L1 e de L1, voltar para a sentença em L2. O método foi testado em um corpus de aprendizes, alcançando uma precisão de 82% com a combinação dos dois modelos. Já, os métodos de Gamon et al. (2009) foram treinados e testados em um corpus de nativos e, além disso, os erros identificados pelo sistemas em três corpus de aprendizes diferentes foram avaliados manualmente. O modelo proposto por Gamon et al. (2009) é responsável pela detecção e correção de

¹³O *parser* XLE consiste em um algoritmo que analisa um texto baseando-se na gramática léxico-funcional (<http://www2.parc.com/isl/groups/nlft/xle/>).

¹⁴É um anotador sintático de textos que inclui uma representação semântica do texto e um etiquetador morfossintático (Briscoe et al., 2006).

erros na ferramenta *ESL Assistant*, apresentada na Seção 4.

Outra alternativa utilizada para detectar diversas categorias de erros é a geração artificial de erros. Essa solução surge para suprir a ausência de um corpus com erros anotados ou quando o número de erros não é suficiente. Foster e Oistein (2009) apontam que “erros artificiais devem se adequar aos erros anotados manualmente”. Se não for considerado esse ponto, o desempenho dos métodos de classificação pode cair quando aplicado a textos reais produzidos por aprendizes. Foster e Oistein (2009) apresentam uma ferramenta, chamada *GenERRate*, que é empregada para produzir diferentes tipos de erros sintáticos. Os erros gerados foram inspirados no corpus de aprendizes *Cambridge Learner Corpus* (CLC), que não está disponível livremente e que é um dos corpus referentes ao projeto ICLE. Os erros artificiais foram usados na tarefa de detecção de erros a fim de classificar sentenças em gramaticais ou não gramaticais.

A mesma abordagem foi empregada para a língua russa por Dicknson (2010), entretanto, o autor focou na combinação de lexemas e sufixos com o objetivo de criar dados reais para um sistema de aprendizado de máquina. Como a combinação aleatória de sufixos e lexemas pode resultar em inúmeros erros improváveis, o autor guiou as combinações, considerando a probabilidade dos erros para assegurar que os erros gerados sigam uma distribuição adequada. Dicknson (2010) propôs gerar erros similares aos dos aprendizes, mas não especificou que aprendizes cometem os erros gerados, segundo o autor, que não se aprofundou em questões relacionadas com erros decorrentes de interferências da língua nativa. Embora Dicknson (2010) não tenha considerado nenhum conjunto de erros como referência para a tarefa de geração de erros artificiais, esse mesmo fato infringe a regra básica da criação de erros artificiais, pois um dos principais problemas dessa abordagem é que os erros artificiais devem representar erros reais. No contexto de aprendizagem de uma segunda língua, os erros gerados devem representar os erros que aprendizes da língua estrangeira cometem.

Outro problema que tem sido estudado para auxiliar a escrita de aprendizes de uma segunda língua são os erros de colocações. Erros de colocações são frequentes até em etapas mais avançadas da aprendizagem, porém ainda representa um desafio para PLN o tratamento desse tipo de erros (Ferraro et al., 2014). Os modelos automáticos do estado da arte das colocações predizem, com uma precisão razoável, quando uma dada colocação é uma colocação válida ou não, porém são poucos os métodos que oferecem uma su-

gestão adequada e com uma boa precisão. Muitas vezes é mostrada uma lista de opções para o usuário escolher (Ferraro et al., 2014; Chang et al., 2008; Hua chen et al., 2010). Para detectar erros de colocações os autores usaram técnicas baseadas em frequências e em medidas de associação (Ferraro et al., 2014; Chang et al., 2008; Hua chen et al., 2010). Por outro lado, para selecionar a correção do erro de colocação combinaram-se medidas baseadas na similaridade ortográfica (Liu et al., 2009), medidas de associação (Ferraro et al., 2014; Chang et al., 2008; Hua chen et al., 2010; Liu et al., 2009), análises contextuais (Liu et al., 2009) e emprego de outros recursos como listas de n-gramas com anotação morfossintática e estruturas taxonômicas como a *EuroWordNet*(Chang et al., 2008).

A Tabela 4.1 apresenta um resumo com os estudos comentados nesta seção, propostos para detectar e corrigir diferentes tipos de erros.

Tabela 4.1: Resumo dos trabalhos apresentados na Seção 4.2

Autores	Método Computacional	Erro	Atributos	Precisão	Treinamento e Avaliação
Foster e Vogel (2004)	<i>Parser</i>	Erros gramaticais	---	84%	Sentenças com e sem erros
Gagner et al. (2007)	Frequência dos n-gramas e <i>Parser</i>	Palavras ausentes, palavras duplicadas, erros de ortografia e de concordância	Etiquetas morfossintáticas	66%	Erros artificiais
De Felice (2008)	<i>Parser</i>	Preposições e artigos	Etiquetas morfossintáticas, Relações semânticas entre palavras, e Entidades nomeadas	70%-92%	Cópus de nativos e aprendizes
Andersen (2007)	<i>Parser</i> e <i>Naive Bayes</i>	Erros ortográficos, Palavras desnecessárias, tempos verbais, ordem das palavras	Etiqueta morfossintática, lemas, relações gramaticais e concordância	70%	Cópus de nativos
Chodorow et al. (2007)	Classificadores	Preposições	Contexto da palavra e conjunto de regras	80%-88%	Cópus de aprendizes
Tetreault et al. (2010)	Classificadores e <i>parser</i>	Preposição e complemento da preposição	Extensão das características usadas em Chodorow et al. (2007)	(1) 86% - 95%; (2) 87% - 96%; (3) 70% - 87%	Cópus de aprendizes (sentenças sem erro (1) / com erros (2) / mais erros (3))
Dahlmeier e Tou Ng (2011)	Classificadores	Erros gramaticais	Igual que Chodorow et al. (2007) e De Felice (2008)	68%-87%	Cópus de nativos
Hermet e Désilets (2009)	Modelo de língua e Tradução	Preposições	---	82%	Cópus de aprendizes
Gamon et al. (2009)	Classificação e modelo de língua	Preposições e artigos	---	---	Cópus nativos e aprendizes
Foster e Oistein (2009)	classificação	sentenças gramaticais ou não gramaticais	---	Erros artificiais	Cópus aprendizes
Dickinson (2010)	classificação	erros sufixos	---	Erros artificiais	---

4.2.1 O Revisor Gramatical ReGra

Na década dos 90, foram vários os pesquisadores interessados em criar um revisor gramatical para o português do Brasil. Essas iniciativas culminaram nos seguintes corretores para o tratamento gramatical da língua portuguesa do Brasil: o DTS¹⁵ (Dts-Software-Ltda, 1995), a Gramática Eletrônica (Informática, 1999), o revisor SOS (Moura, 1996; Lins et al., 1999) e o ReGra (Martins et al., 1998), comercializado nos produtos RLP Itautech¹⁶ e *Microsoft Office 2000*. Destes corretores, atualmente, o único que se mantém no mercado é o ReGra que coexiste com outros corretores, tais como CoGrOO¹⁷, que é o corretor gramatical acoplado ao BrOffice.org; o FLiP¹⁸, que é um pacote de ferramentas de revisão e auxílio à escrita para o português europeu e brasileiro, compatível com *LibreOffice* e *OpenOffice.org* e o Revisor Gramatical¹⁹. Apesar da existência dessas três últimas ferramentas, o ReGra, segundo Ribeiro (2012) é o precursor de todo o desenvolvimento dos estudos ligados ao processamento computacional do português do Brasil e desde o ano 2000 está incorporado na ferramenta *Office*, que é a mais vendida no Brasil e no mundo, o que, o converte em um revisor robusto e estável. Embora o revisor ReGra não foi criado para auxiliar os aprendizes de português na produção escrita dessa língua, atualmente é a ferramenta mais utilizada pelos hispanofalantes com essa finalidade. Por isso, merece destaque nesta pesquisa.

O revisor ReGra foi construído pela equipe do NILC, na forma de regras compiladas. Esse recurso apresenta um padrão de eficiência e robustez de destaque para o tratamento da língua portuguesa falada no Brasil. O revisor conta com mais de 1000 regras sintáticas e seu bom desempenho deve-se, em grande parte, ao fato da validação e construção das regras terem sido baseadas na compilação de um grande corpus de textos de diversos domínios que contém mais de 39 milhões de palavras. O ReGra conta também com um léxico com mais de 1.5 milhão de entradas, incluindo nomes próprios, acrônimos e abreviações (Nunes et al., 1996; Pinheiro et al., 2001; Hasegawa et al., 2002), com informações morfossintáticas sobre essas entradas. Desde sua criação houve trabalhos de aumento do seu dicionário e seu léxico e, no ano 2009, foi incorporada as regras referentes à nova ortografia da língua portuguesa.

¹⁵<http://www.dts.com.br/>

¹⁶<http://www.itautech.com.br/produtos/soft/rlp/>

¹⁷<http://www.broffice.org/projetos/cogroo>

¹⁸<http://www.flip.pt/>

¹⁹<http://www.katalogo.com.br/Produtos/IdProduto=CAT12807-7descricao>

O ReGra é dividido em quatro módulos e trata diversos tipos de erros que vão desde a escrita incorreta das palavras, erros de digitação e até imperfeições quanto à estruturação das sentenças de acordo com as normas de escrita da língua portuguesa. O sistema também produz medidas estatísticas do texto sendo analisado. A seguir, são explicados brevemente cada um dos módulos:

Módulo Estatístico: Oferece uma análise estatística do texto, indicando aos escritores um conjunto de informações, tais como número total de parágrafos, sentenças, palavras, letras, número médio de sentenças por parágrafos, de palavras por sentenças e de letras e símbolos por palavra.

Módulo Mecânico: Este módulo é responsável por erros que aparecem durante a digitação, tais como escrita incorreta de palavras ou símbolos, palavras repetidas em uma sentença, delimitadores desbalanceados, capitalização incorreta, uso de aspas e símbolos de pontuação, entre outros.

Módulo Linguístico: Este módulo está dividido em três partes: a de revisão pontual, a de análise sintática e a de revisão gramatical. A revisão pontual tenta identificar erros frequentes que são facilmente detectados por meio de padrões pré-definidos. Nestes problemas estão: uso da crase, uso de prefixos, gênero das palavras, posicionamento dos pronomes, uso das preposições em expressões, uso do particípio de alguns verbos e verificação da pontuação em expressões. Após o etiquetamento morfosintático de todas as palavras e uma vez corrigidos os desvios mais comuns pelas regras pontuais, o *parser* parte para a identificação da estrutura sintática das sentenças. O *parser* checa erros mais complexos, tais como relações de regência e concordância, que dependem das funções sintáticas. O *parser* empregado na análise sintática utiliza uma Gramática Livre de Contexto com técnicas de restrições baseadas em relaxamento (*relaxation constraints-based techniques*) para permitir o casamento de sentenças não gramaticais. O *parsing* é feito utilizando uma tabela de predições de probabilidades sintáticas. Se as regras corretas não casarem, o *parser* procura um casamento com estruturas supostamente incorretas; isso é feito com ajuda das técnicas de relaxamento. A análise gramatical intercala o *parsing* e a checagem que consiste na aplicação de regras de concordância nominal e verbal entre os termos da sentença (Bonfante e Nunes, 2000).

O corretor gramatical ReGra também trata problemas específicos e de difícil solução na língua portuguesa. Um deles é o tratamento da ambiguidade categorial. Um exemplo claro deste fenômeno é o uso da partícula “se”pronomine e “se”conjunção, por ser esta uma palavra empregada frequentemente nos textos da língua portuguesa, sendo a ela atribuídas diversas categorias sintá-

ticas.

Hasegawa et al. (2002) apresenta uma avaliação das principais características incluídas na versão final do revisor gramatical ReGra (Maio/2002). Após a comparação do desempenho desta última versão com as versões anteriores, o autor confirma o aumento da precisão do revisor gramatical, ao mesmo tempo em que atualmente identifica alguns erros que o revisor não conseguia tratar. Alguns desses problemas são listados a seguir, junto com exemplos apresentados por Hasegawa et al. (2002).

- **Acentuação Gráfica**

- 1- V'a para casa e n ao volte mais aqui. (Inadequado) -> Vá para casa e não volte mais aqui. (Adequado)
- 2- Canta-lo-ei. (Inadequado) -> Cantá-lo-ei. (Adequado)
- 3- Tu vais compra-lo. (Inadequado) -> Tu vais comprá-lo. (Adequado)

- **Uso do Adjetivo**

- 4- O paciente necessitou de uma intervenção médica-cirúrgica. (Inadequado) -> O paciente necessitou de uma intervenção médico-cirúrgica. (Adequado)
- 5- Ela usa vestidos cinzas. (Inadequado) -> Ela usa vestidos cinza. (Adequado)
- 6- Os rapazes estavam usando camisas verdes-águas. (Inadequado) -> Os rapazes estavam usando camisas verde-água. (Adequado)

- **Uso de Artigos e Determinantes**

- 7- A Sua Majestade encontra-se descansando. (Inadequado)-> Sua Majestade encontra-se descansando. (Adequado)
- 8- Uma certa vez, o encontrei na rua. (Inadequado) -> Certa vez o encontrei na rua. (Adequado)
- 9- Ele está errado ao meu ver. (Inadequado) -> Ele está errado a meu ver. (Adequado)

- **Concordância entre Modos e Tempos**

- 10- Eles desejariam que as meninas ficassem. (Inadequado) -> Eles desejariam que as meninas ficassem. (Adequado)

- 11- Ele talvez vai. (Inadequado) → Ele talvez vá. (Adequado)

- **Inadequação Lexical**

- 12- Indústrias de automóveis (Inadequado) → Indústrias de automóvel (Adequado)
- 13- Departamento pessoal (Inadequado) → Departamento de pessoal (Adequado)

- **Regência Nominal**

- 14- Bacharel de Direito (Inadequado) → Bacharel em Direito (Adequado)
- 15- Isto é fácil de se resolver. (Inadequado) → Isto é fácil de resolver. (Adequado)

- **Regência Verbal**

- 16- Os pedidos são entregues a domicílio. (Inadequado) → Os pedidos são entregues em domicílio. (Adequado)
- 17- Vamos dar entrada no processo amanhã. (Inadequado) → Vamos dar entrada ao processo amanhã. (Adequado)

Após a última versão do revisor gramatical ReGra, reportada por Hasegawa et al. (2002), o sistema continua sendo aprimorado pela empresa *Techno Software*²⁰ (parceria da Itautech), para oferecer um maior auxílio ao usuário. O fato anterior é confirmado após verificar que a maioria dos erros comentados anteriormente já são detectados e corrigidos pelo revisor. O ReGra, atualmente conhecido como o corretor gramatical do *Word 2010* para o português, não conseguiu identificar os problemas apresentados nos itens 3, 4, 6 e 16, mostrando que ainda há problemas relacionados com acentuação gráfica, uso de adjetivo, uso de substantivos compostos e regência verbal que não apontados pela ferramenta.

Apesar disso, atualmente a ferramenta mais utilizada para apoiar a escrita em português brasileiro é o corretor gramatical ReGra, inclusive entre os aprendizes do português. No entanto, essa ferramenta não foi construída considerando os erros dos aprendizes, por isso, não está preparada para tratar os problemas que esses usuários enfrentam, aceitando inúmeros erros produzidos pelos aprendizes. Muitas vezes, a ferramenta consegue apontar o

²⁰<http://www.technosoftware.com.br/>

erro, mas devido ao fato do erro ser produzido por uma interferência da língua nativa, o ReGra não consegue oferecer uma correção adequada.

4.2.2 Avaliação da Tarefa de Detecção de Erros Gramaticais

Inúmeros esforços aparecem na literatura para detectar e corrigir erros gramaticais, sendo várias as abordagens utilizadas para avaliar os métodos propostos. Além disso, há muita disparidade nos conjuntos de dados utilizados e os tipos de erros tratados. Esses fatos, influenciam na difícil tarefa de mensurar o desempenho dos sistemas, sendo mais difícil ainda comparar um sistema com outro (Chodorow et al., 2012; Leacock et al., 2010). Desta forma, ainda não existe um consenso sobre qual é a melhor alternativa para avaliar sistemas de detecção e correção de erros gramaticais.

A precisão, cobertura e F-média têm sido as medidas usadas, tradicionalmente, para avaliar sistemas de detecção e correção de erros. Para ilustrar como elas são calculadas emprega-se como exemplo a tarefa de detectar erros de uso da vírgula, apresentado por Chodorow et al. (2012). Imaginem que as vírgulas foram eliminadas de um texto original e o sistema precisa predizer onde colocá-las novamente. Desta forma, tem-se um texto original, em que aparecem as vírgulas e o texto em que elas foram eliminadas, e a comparação entre a saída do sistema e texto original (*gold standard*) é representada como na tabela de contingência (Tabela 4.2).

Na tarefa, a presença da vírgula é etiquetada como a classe positiva, enquanto a ausência dela é a classe negativa. As instâncias em que o sistema e o *gold standard* coincidem na presença da vírgula são consideradas verdadeiros positivos (VP); quando ambos indicam que não deve aparecer a vírgula são verdadeiros negativos (VN); quando o sistema indica o uso da vírgula e o *gold standard* não, são falsos positivos (FP) e quando o sistema indica que a vírgula não deve ser usada e o *gold standard* aponta que deve, são falsos negativos. O cálculo da precisão, cobertura e F-média é realizado considerando-se os valores da Tabela 4.2 como aparece nas Equações 4.1, 4.2 e 4.3, respectivamente.

Tabela 4.2: Avaliação de sistemas de detecção e correção de erros gramaticais

		<i>Gold Standard</i>	
		Presença Vírgula	Ausência Vírgula
Predição	Presença Vírgula	VP	FP
	Ausência Vírgula	FN	VN

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (4.1)$$

$$\text{Cobertura} = \frac{VP}{VP + FN} \quad (4.2)$$

$$F - \text{media} = 2 * \frac{P * C}{P + C} \quad (4.3)$$

A precisão mede a quantidade de vezes que o sistema aponta um erro quando realmente é um erro. Por outro lado, a cobertura representa a proporção de erros que foram detectados. A F-média é a média harmônica entre a precisão e a cobertura. Porém, Leacock et al. (2010) recomendam não relatar os resultados da avaliação considerando a F-média, porque ela dá o mesmo nível de importância para a precisão e a cobertura. Vários pesquisadores consideram que em um sistema de detecção de erros a precisão é mais importante que a cobertura, pois tem um efeito adverso sobre o aluno se o sistema produz um falso positivo por sinalizar uma construção bem formada como um erro.

O *gold standard* utilizado para avaliar os sistemas pode ser criado de diversas formas. Uma das formas mais simples é empregar textos produzidos por nativos, que têm poucos erros, para avaliar os métodos (Leacock et al., 2010). Esse tipo de recurso é fácil de achar e há diversos disponíveis na Web. Nessas avaliações acontece uma classificação binária: correto ou incorreto. Quantificam um erro quando foi predito um *token* diferente ao achado no texto, caso contrário a predição é correta. Uma deficiência dessa avaliação é que unicamente é aceito como correto o termo que aparece no texto; sabe-se que a correção pode ter diversas formas. Essa avaliação é totalmente automática e estima-se a precisão instantaneamente. Avaliar usando textos sem erros permite comparar os sistemas, porém o fato de que o contexto ao redor de um erro seja livre de erros, coisa que não acontece em textos de aprendizes, influencia positivamente no cálculo da precisão. Seguindo essa linha, a avaliação em textos livres de erros diz pouco sobre o comportamento do sistema em textos de aprendizes.

Outra forma de criar o *gold standard* é considerando textos produzidos por aprendizes, previamente anotados com todas as categorias de erros. O principal desafio dessa abordagem é a compilação do corpus de aprendizes e a anotação manual das diversas categorias de erros e suas correções. A maioria dos corpus de aprendizes com anotação, não são públicos ou são pequenos,

de tal forma que não oferecem uma análise estatística válida.

Para anotar erros em um *córpus* de aprendizes, em vários casos, é necessário identificar as intenções do escritor para categorizar o erro e oferecer a melhor correção. Para diminuir as ambiguidades nesse processo recomenda-se a participação de mais de um anotador. A concordância entre os anotadores é uma medida que indica quão difícil é a tarefa e, além disso, é um teste para estimar a confiabilidade da anotação. Para avaliar a concordância entre os anotadores, é comum o uso do coeficiente *Kappa*. Atualmente, o único *córpus* anotado por mais de um anotador, com os valores de concordância *Kappa* é o *córpus* NUCLE (Dahlmeier et al., 2013).

Outro grande desafio da avaliação em *córpus* de aprendizes é que a tarefa não consiste em classificar o erro em correto ou incorreto. Nos textos dos aprendizes aparecem diversas categorias de erros (ortográficos, gramaticais, semânticos) e o sistema precisa identificar essas categorias para conseguir oferecer a sugestão correta. Além disso, no contexto do erro identificado podem aparecer outros erros, fato que dificulta a tarefa de correção do texto.

Por outro lado, quando disponível um *córpus* de aprendizes, mas que não foi anotado com os erros, a avaliação consiste em verificar manualmente os erros detectados. Esta abordagem foi utilizada por Gamon et al. (2008) que, evitando a anotação das diversas categorias de erros, assim como suas correções, propôs os critérios apresentados na Tabela 4.3 para avaliar a saída do sistema, que detectava e corrigia erros de artigos e preposições.

Tabela 4.3: Verificação das categorias de erros proposta por Gamon et al. (2008)

Categorias de verificação de erros propostas por Gamon et al. (2008)
1) correção é válida e corrige o problema
2) o erro é identificado corretamente, mas a sugestão não corrige o problema
3) o trecho original e a correção são igualmente adequados
4) um erro ocorre na sugestão ou perto dela, mas o erro é de outro tipo
5) há erros ortográficos na correção
6) a correção está errada, a original é correta

A verificação apresentada na Tabela 4.3 consegue levantar os mesmos aspectos apontados na avaliação do sistemas em textos livre de erros (categorias 1 e 6), mas também consegue fazer uma análise de alguns pontos que complicam a tarefa de detecção de erros, como é caso da presença de erros nas sugestões, ou a dificuldade em categorizar os diferentes tipos de erros. Embora a tarefa de verificação da saída do sistema seja uma tarefa custosa e demorada,

recomenda-se que mais de um anotador realize a verificação do sistema, para depois se estabelecer os níveis de concordância entre os anotadores e obter mais confiabilidade na análise.

4.3 Considerações Finais

Neste capítulo, foram apresentadas ferramentas e métodos com a finalidade de dar apoio à escrita de aprendizes de uma língua estrangeira ou segunda língua. Os trabalhos apresentados mostraram que existem inúmeras abordagens para criar os auxílios e que a área de auxílio à escrita é uma área ativa, em que ainda muitos problemas precisam ser tratados.

Pela revisão da literatura, é possível confirmar que os métodos dedicados à identificação de erros englobam um conjunto de técnicas relacionadas com a criação de regras; métodos baseados em *parser* e métodos estatísticos baseados em *cópus*. Após conhecer as diferentes abordagens propostas, é possível confirmar que para construir um detector e/ou corretor de erros eficiente, precisa-se empregar recursos caros de desenvolver, como é o caso de *parser* robustos, léxicos com informações morfossintáticas, estruturas taxonômicas, reconhecimento de entidades nomeadas, etc.. Além disso, é necessário compilar *cópus* de aprendizes que mostrem os diferentes tipos de erros, sendo um recurso útil para treinar e avaliar os métodos automáticos. Caso um *cópus* de aprendizes não esteja disponível pode-se utilizar *cópus* de nativos, no entanto, esse tipo de recurso não mostra as características dos aprendizes. A escolha do conjunto de dados para treinar, testar ou avaliar os métodos de detecção de erros não é uma tarefa trivial e é um processo que deve ser realizado de forma muito criteriosa.

A tarefa de avaliação dos métodos computacionais é um dos maiores desafios enfrentados na área, pois em todos os trabalhos consultados foi necessária a avaliação humana ou a criação de recursos por humanos. Além disso, os métodos precisam ser avaliados em diferentes momentos, por exemplo: é preciso saber a qualidade do *cópus* de aprendizes; a eficiência dos métodos propostos para solucionar problemas específicos; o desempenho da integração de vários métodos e finalmente a utilidade dos métodos, em forma de ferramenta, para auxiliar os aprendizes da língua.

Observa-se que alguns trabalhos são continuidade dos outros, mostrando a integração de pesquisadores de diversas áreas, para ter sucesso no desempenho dos métodos propostos. Por exemplo, os trabalhos de De Felice (2008)

e Tetreault et al. (2010) foram propostos para aprimorar os resultados obtidos nos trabalhos de Chodorow e Leacock (2000) e Chodorow et al. (2007), respectivamente. Além disso, a maioria dos métodos propostos combinam técnicas de PLN e são dependentes de língua. Desta forma, sua migração para outra língua é custosa, por depender de ferramentas básicas, recursos linguísticos ou computacionais e aprendizes de determinada língua.

Projeto HABLA: Compilação dos Córpus e Análises

Nesta pesquisa, foram criados léxicos bilíngues para auxiliar os hispano falantes na produção escrita em português. Como os léxicos bilíngues contêm itens lexicais de diferentes tipos, a forma de compilar cada léxico e os tipos de erros associados a eles foram diferentes.

A compilação dos léxicos em forma de auxílios é apresentada no Capítulo 6, contudo, na Seção 5.1 deste capítulo, são apresentados os corpúscos selecionados como referência. Alguns dos corpúscos apresentados na Seção 5.1 foram compilados no escopo do projeto HABLA, mas outros foram selecionados de outras pesquisas da literatura.

Além disso, foram realizadas diversas análises, apresentadas na Seção 5.3, com o objetivo de caracterizar a produção escrita em português dos hispano falantes. As análises foram realizadas considerando textos em português produzidos por nativos e por hispano falantes. Tanto as análises realizadas quanto os corpúscos utilizados fazem parte da base de conhecimento do projeto HABLA. As análises confirmaram que erros lexicais, produzidos pela interferência da língua nativa, o espanhol, são recorrentes nos textos dos hispano falantes. Desta forma, verificou-se a importância de propor auxílios que apoiem alguns destes problemas lexicais.

A Figura 5.1 traz uma representação visual da base de conhecimento do projeto HABLA, mostrando o seu processo de criação. Nesta pesquisa, léxicos

bílingues foram extraídos de corpus e por meio de diversas análises em textos de aprendizes foram levantados alguns dos problemas que hispano falantes enfrentam na escrita acadêmica em português.

Figura 5.1: Base de conhecimento do projeto HABLA

5.1 Cópus de Referência

Nesta pesquisa foram compilados os seguintes léxicos bilíngues: (i) Falsos Cognatos/ Cognatos; (ii) Verbos usados frequentemente na escrita acadêmica; (iii) Expressões Acadêmicas e (iv) Marcadores discursivos. Para construir os léxicos, cuja finalidade é apoiar a escrita acadêmica em português, uma abordagem baseada em corpus foi adotada. Para isso, foi necessário selecionar os corpus mais adequados para guiar a tarefa. Nas duas seções seguintes, são apresentados estes corpus, suas principais características e como eles foram utilizados nesta pesquisa.

5.1.1 Cópus Paralelo espanhol-português

Como comentado na Seção 3.1, os corpus paralelos são recursos úteis em diversas aplicações. Neste projeto corpus paralelos aparecem como base de conhecimento para guiar métodos dedicados a extrair léxicos bilíngues para

apoiar uma ferramenta de auxílio à escrita. Apesar da utilidade desses recursos para os pares português e espanhol, não há muitos corpúis paralelos disponíveis na literatura, sendo menor ainda o número de corpúis paralelos com textos do gênero acadêmico. Projetos desenvolvidos no NILC têm disponibilizado corpúis paralelos dos pares de línguas inglês-português e português-espanhol. A seguir, é descrito o corpúis criado para os pares de línguas português-espanhol, apresentado em Aziz e Specia (2011), pois é o mais adequado para compilar os léxicos bilíngues para o projeto HABLA.

O corpúis paralelo disponibilizado por Aziz e Specia (2011) é composto por textos provenientes da edição online da Revista Pesquisa FAPESP¹. Esta revista divulga as notícias e reportagens científicas de diferentes áreas do conhecimento, tais como Meio Ambiente, Ciência, Humanidades, Política e Tecnologia. Uma versão inicial deste corpúis foi utilizada no projeto PESA², porém o recurso era pequeno se comparado com a versão atual de Aziz e Specia (2011).

Aziz e Specia (2011) apresentaram um método para compilar de forma automática os textos publicados na Revista Pesquisa FAPESP, nas línguas português, inglês e espanhol. Como resultado final desse estudo, um corpúis com 2.669 textos paralelos foram obtidos para os pares de línguas português e espanhol e 2.675 para os pares português-inglês. Esse corpúis já vem tokenizado e separado por sentenças, uma em cada linha dos seus respectivos arquivos. O corpúis está alinhado por sentenças pela ferramenta *TCAling*, descrita na Seção 3.1.1 e possui 152.238 pares de sentenças em português e espanhol e 156.712 pares de sentenças em português e inglês. Aziz e Specia (2011) avaliaram a qualidade do corpúis paralelo analisando 900 alinhamentos do tipo 1-1 nos textos português e inglês, o que resultou em 98% de precisão. Os autores não avaliaram manualmente os alinhamentos dos textos português e espanhol, porque devido à proximidade entre as línguas eles consideraram que o desempenho do algoritmo seria inclusive melhor. A parte em espanhol do corpúis contém aproximadamente 4.246.410 palavras, enquanto a parte em português contém 3.903.138.

Entretanto, o corpúis paralelo, disponibilizado por Aziz e Specia (2011), para as línguas português-espanhol, não é suficientemente grande se comparado com outros recursos disponíveis para a língua inglesa, que foram utilizados para compilar léxicos bilíngues (Tiu e Roxas, 2008). Adicionalmente, os textos

¹<http://revistapesquisa.fapesp.br>

²<http://www.nilc.icmc.usp.br/nilc/projects/pesa.htm>

que integram o corpus da Revista Pesquisa FAPESP são do gênero jornalístico, mas descrevem trabalhos científicos, caracterizando o gênero de divulgação científica. Esse fato indica que o corpus não é o recurso ideal para a construção dos recursos do projeto HABLA, pois os recursos devem apoiar a escrita acadêmica. Desta forma, o ideal era dispor de um corpus que mostrasse as particularidades da escrita deste gênero. Entende-se como particularidades do gênero científico tanto a estrutura dos textos quanto os principais termos e expressões linguísticas utilizadas. Dada essa falta, foi proposta a criação de um corpus paralelo português-espanhol composto por textos acadêmicos que descrevam as características do gênero científico.

O processo de criação do corpus paralelo português-espanhol começou com uma busca minuciosa no Portal de Periódicos Capes/MEC³, que disponibiliza, livremente, artigos científicos de revistas nacionais, internacionais e estrangeiras. O Portal da Capes foi selecionado para procurar os textos, porque nele são indexadas revistas que, em sua maioria, exigem qualidade na escrita acadêmica. Como resultado da busca, foi comprovada a inexistência de revistas científicas que fornecem textos acadêmicos completos, escritos em português e em espanhol. A maioria das revistas consultadas disponibilizam resumos em português e em inglês ou em espanhol e em inglês. Contudo, foram identificadas nove revistas, em que alguns resumos apareciam em espanhol e em português. Essas revistas são apresentadas na Tabela 5.1; somente as revistas *Interciência* e *Ibero-americana de Educación*, foram selecionadas para obtenção dos resumos paralelos, pois nas outras revistas o número de resumos disponíveis era pequeno. Para extrair os resumos das revistas candidatas, foi empregada a biblioteca *BeautifulSoup*⁴ em *Python* para implementar um *crawler* que automaticamente compilou todos os textos. Algumas estatísticas dos corpus extraídos aparecem na Tabela 5.2.

O revista Venezuelana *Interciencia* é multidisciplinar, e seus temas prioritários são agronomia e bosques tropicais. Os resumos dessa revista são de aproximadamente 250 palavras e a revista disponibiliza em media 100 textos por ano, tendo um fator de impacto de 0,0674. A *Revista Iberoamericana de Educación* é da área de Ciências Sociais Aplicadas, avaliada como B1 segundo a Capes. O periódico é uma publicação mensal e inclui pesquisas da área da educação; os resumos são de aproximadamente 100 a 300 palavra. Nas duas revistas os resumos foram escritos em espanhol e em português, não

³<http://www-periodicos-capes-gov-br.ez67.periodicos.capes.gov.br/index.php>

⁴<http://www.crummy.com/software/BeautifulSoup/bs4/doc/>

Tabela 5.1: Revistas que disponibilizam textos ou resumos em português e espanhol

Ciências Exatas e da Terra
Interciência
Ciências Sociais Aplicadas
Interações
Revista Ibero-americana de Educación
Innovar
Revista Brasileira de Estudos de População
Katálisis
Linguística Letras Artes
Documentação de Estudos em Linguística Teórica e Aplicada – DELTA
Ciências Biológicas
Revista de Saúde Pública
Rev.LatinoAm.Enfermagem

são traduções mútuas, pois encontramos diferenças no conteúdo dos textos paralelos; esse fato se evidencia nos números das estatísticas apresentadas na Tabela 5.2.

Na busca por revistas que aceitam artigos em espanhol e português para a área de Ciência da Computação, por exemplo, não foi identificada nenhuma revista. Considera-se que isso esteja relacionado com o pequeno número de estudos publicados em português e em espanhol nesta área do conhecimento.

Tabela 5.2: Estatísticas do corpus paralelo compilado no projeto HABLA

	Estatísticas	Espanhol	Português
Revista Ibero-americana de Educación	Textos	186	186
	Tokens	33.725	29.561
	Palavras	29.557	28.906
	Sentenças	849	798
	Palavras por Sentenças	37	38
	Estatísticas	Espanhol	Português
Interciencia	Textos	876	876
	Tokens	199.794	193.518
	Palavras	160.642	151.251
	Sentenças	6.332	6.412
	Palavras por Sentenças	26	29

As duas revistas selecionadas são de origem hispânica e, como consequência, foi verificada a qualidade dos textos em português. Após analisar uma parte desses textos comprovou-se a presença de marcas do espanhol nos tex-

tos em português, sendo posteriormente confirmado que os autores dos textos eram hispano falantes. Devido ao fato anterior, os corpúis paralelos compilados de forma automática e apresentados na Tabela 5.2 não representam a língua portuguesa escrita por nativos do português e sim hispano falantes produzindo textos em uma língua estrangeira ou segunda língua. Desta forma, esses corpúis não foram utilizados na construção dos léxicos bilíngues, mas eles representam um recurso valioso no processo de identificação de erros de hispano falantes, como comentado na Seção 5.3.3. A seguir, apresentam-se exemplos de erros extraídos dos corpúis *Interciencia* e *Revista Iberoamericana de Educación*. Os trechos com erros aparecem em itálico e sublinhado, já as correções em negrito e entre parênteses.

1. Exemplo extraído da Revista *Interciencia*: Não há comunidades humanas dentro da reserva, mas na sua zona aledanha (**ao seu redor**) a maior parte das terras pertencem a camponeses ejidatarios (**herdeiros de terras**), e na região colindante com o (**próxima ao**) mar existem hoteles (**hotéis**), casas de praia e oturas (**outras**) facilidades turísticas (**atrações turísticas**).
2. Exemplo extraído da *Revista Iberoamericana de Educación*: Este trabalho contempla os aspectos mais significativos da situação atual da acreditação (**valorização**) da formação universitária no Peru.

Concluindo, o único corpúis paralelo que foi utilizado para compilar os léxicos bilíngues é o corpúis da Revista Pesquisa Fapesp, disponibilizado por Aziz e Specia (2011). Porém, como mostrado no Capítulo 6 o corpúis paralelo foi combinado com outros recursos lexicais, por exemplo, com o corpúis do gênero acadêmico, apresentado na Seção 5.1.2.

5.1.2 Corpúis do Gênero Acadêmico

Como o corpúis paralelo disponibilizado por Aziz e Specia (2011) e apresentado na Seção 5.1.1 não é do gênero acadêmico foi necessário compilar um recurso com textos específicos desse gênero. A ideia era obter um corpúis em espanhol e outro em português que mostrassem algumas das características do gênero acadêmico, principalmente os vocabulários utilizados.

Para compilar o corpúis em português foi selecionado o sub-corpúis Lacio-Web Exatas, construído no projeto Lácio-Web⁵ e que contém textos acadêmicos

⁵<http://www.nilc.icmc.usp.br/lacioweb/colaboradores.htm>

da área de exatas. Os textos deste sub-córpus foram produzidos por nativos da língua portuguesa e publicados em exames de qualificação ou defesa e em revistas científicas. O córpus com os textos acadêmicos em espanhol é composto dos resumos selecionados da Revista *Interciencia e Iberoamericana de Educación*. Esses resumos foram escritos por nativos do espanhol e para revistas científicas. As estatísticas de ambos os córpus aparecem na Tabela 5.3.

Tabela 5.3: Estatísticas do córpus de referência do projeto HABLA

Córpus Referência recopilado no projeto HABLA					
	Textos	Tokens	Palavras	Sentenças	Palavras por Sentenças
acadêmico português	543	1.303.220	1.069.314	46.873	23
acadêmico espanhol	1.062	233.519	190.199	7.181	31,5

Como ilustrado na Tabela 5.3, os córpus acadêmicos são pequenos, mas acredita-se que eles são uma boa referência para mostrar algumas das características do gênero acadêmico, principalmente as palavras e expressões mais frequentemente utilizadas. O número de textos do córpus em português é menor do que os em espanhol, porém os textos em português são textos completos, enquanto os textos em espanhol são resumos. Dessa forma, o número de palavras e sentenças em português é maior.

Para obter um recurso que combinasse as características do córpus paralelo da revista Fapesp e do córpus do gênero acadêmico, foi criado um modelo de língua baseado em n-gramas. Como comentado na Seção 3.1.3, modelos de língua são muito utilizados em diversas áreas do PLN, sendo um recurso que tem auxiliado diversas pesquisas que detectam e corrigem erros de aprendizes de uma língua. Neste projeto, o modelo de língua aparece como um recurso que mostra o uso adequado da língua portuguesa. Seguindo essa linha, foi criado um modelo de língua usando como entrada os textos do córpus Fapesp e do gênero acadêmico.

O modelo baseado em n-gramas foi criado utilizando a ferramenta SRLIM, citada na Seção 3.1.3. Para isso, todas as sentenças dos córpus foram incluídas em um único arquivo. No arquivo há uma sentença por linha e as etiquetas de começo e fim da sentença foram inseridas. As sentenças foram tokenizadas e o método *Kneser-Ney smoothing* (Jurafsky e Martin, 2009) foi

utilizado para estimar a contagem dos n-gramas. Um tokenizador baseado em expressões regulares foi desenvolvido em Python. O modelo treinado de tamanho 5-gramas é composto por, aproximadamente, 400 mil palavras, cinco milhões de 2-gramas, quatro milhões de 3-gramas e 4-gramas e três milhões de 5-gramas. Esse modelo de língua foi empregado com diferentes finalidades, como mostrado ao longo desta pesquisa.

5.2 Córpus de Aprendizes Espanhol-Acadêmico-Br

O *cópus de aprendizes*, em sua **versão inicial**, era composto por teses e dissertações de diferentes áreas do conhecimento. Todas as teses e dissertações foram produzidas por 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. Os autores dos textos são 20 alunos matriculados nos diferentes programas, sendo que em alguns casos são textos com um número de páginas entre 2 e 20 e em outros são textos maiores produzidos para exames de qualificação e defesa. Alguns alunos produziram várias versões de um mesmo texto, e nesse caso cada versão dentro do *cópus* foi quantificada como um texto diferente. A maioria desses textos é composta por todas as seções que um texto acadêmico possui, por exemplo: resumo, introdução, revisão da literatura, materiais e métodos, resultados e conclusões.

Os textos foram produzidos de forma digital e com ajuda de corretores ortográficos e/ou gramaticais. Os autores enviaram os textos por e-mail e muitos deles em formato PDF, sendo necessário realizar uma conversão para o formato TXT. Essa conversão foi realizada cuidadosamente, para manter o texto igual a sua versão inicial. No decorrer deste documento, este sub-*cópus* do *cópus Espanhol-Acadêmico-Br* é chamado de sub-*cópus USP*, mas cada uma das seções dos textos foram consideradas como *cópus* independentes, como mostra a Tabela 5.4. Na Tabela 5.4 apresentam-se as estatísticas desse sub-*cópus* separado por cada seção das incluídas: Introdução, Resumo e Conclusões.

Os autores desses textos são estrangeiros aprendizes de português. Devido ao fato dos alunos estarem matriculados em programas de pós-graduação, que exigem níveis avançados de proficiência da língua portuguesa, a maioria dos textos dos alunos tem nível de proficiência avançado ou intermediário. Isso foi comprovado por meio dos experimentos apresentados na Seção 5.3.1, em que

alguns textos foram classificados segundo a qualidade da escrita. Os alunos que produziram os textos do sub-cópus USP responderam a um questionário apresentado na Seção 5.3.2. Por meio do questionário levantaram-se informações relacionadas com nacionalidade, sexo, tempo de residência no Brasil, entre outras.

Tabela 5.4: Estatísticas do cópus Espanhol-Acadêmico-Br

Sub-Cópus USP					
	Textos	Tokens	Palavras	Sentenças	Palavras por Sentenças
Resumos	19	5.315	5.037	161	32
Introduções	20	21.178	26.054	840	31
Conclusões	10	5.941	4.973	187	26
Sub-cópus Revista Ibero-americana de Educación					
Resumos	186	29.561	28.906	798	38
Sub-cópus Interciencia					
Resumos	876	193.518	18.131	6.412	29
Total	1111	255.513	83.101	8.398	31

Além do sub-cópus USP, o cópus de aprendizes Espanhol-Acadêmico-Br contém, na sua **versão final**, resumos em português escritos por hispano falantes e submetidos à *Revista Ibero-Americana de Educación* e à *Revista Interciencia*. Como comentado na Seção 5.1.1, inicialmente os pares de textos paralelos, disponibilizados pelas duas revistas, formariam parte da base de conhecimento para criar os léxicos bilíngues, porém, devido à presença de marcas de espanhol nos textos, eles foram incluídos no cópus de aprendizes Espanhol-Acadêmico-Br.

O cópus paralelo extraído das revistas *Ibero-Americana de Educación* e *Interciencia* é um recurso que mostra como os hispano falantes realizam transferências do espanhol para o português, resultando em transferências corretas, mas muitas vezes gerando erros. Esse recurso é muito útil para compreender como ocorrem alguns erros na etapa de transferência e pode ser utilizado em pesquisas futuras. Os textos em português desse cópus paralelo integram o cópus de hispano falantes e suas estatísticas também aparecem na Tabela 5.4.

Esse sub-cópus foi compilado da internet e se sabe que os autores dos textos são hispano falantes escrevendo em português, mas não há informações exatas do país em que residem. Porém, supõe-se que eles não residam no Brasil, segundo a afiliação dos autores, que aparece na página da revista. Considerando esse fato, acredita-se que esses textos foram produzidos por aprendizes de L2, mas também não há informações sobre os autores serem

aprendizes ou não.

Na Tabela 5.4 observa-se que o *córpus Espanhol-Acadêmico-Br* é composto por 1.111 textos e aproximadamente 8.400 sentenças. A maioria dos textos que compõe o *córpus Espanhol-Acadêmico-Br* são resumos acadêmicos curtos que possuem em média entre quatro e cinco sentenças. No entanto, as introduções contidas no sub-*córpus USP* são textos maiores que, em média, contêm 20 ou 30 sentenças. Os textos do *córpus Espanhol-Acadêmico-Br* são do gênero acadêmico, especificamente, pertencem a três áreas de conhecimento diferentes: Engenharia/Ciência da Computação; Agronomia e Educação. O fato anterior pode ser um problema para análises automáticas no *córpus Espanhol-Acadêmico-Br*, porque os léxicos utilizados nas três áreas são diferentes. Por exemplo, nos textos de agronomia aparecem diversos termos científicos para denominar nome de plantas e de animais.

5.3 Hispano Falantes Aprendizes de Português Acadêmico: Diversas Análises

Um dos objetivos desta pesquisa foi identificar os erros que nativos do espanhol cometem quando produzem textos acadêmicos em português. A revisão da literatura apresentada na Seção 2.1 mostrou que, embora existam diversos estudos (Rojas (2006); Grannier e Carvalho (2001); Durão (1999)) que levantaram os problemas dos hispano falantes na escrita em português, as análises foram baseadas em *córpus* pequenos que não podem ser utilizados como base de conhecimento para guiar processos automáticos. Esses estudos disponibilizaram tipologias de erros, mas não foi descrito o processo de anotação dos erros e nem informado o número de anotadores envolvidos e as concordâncias da anotação. Por outro lado, as tipologias criadas nessas análises são muito detalhadas, sendo complexa sua reutilização em processos de anotação. Outro ponto a destacar é que não existe nenhum estudo que mostre o desempenho dos hispano falantes na produção escrita de textos científicos em português e nenhuma análise automática que tente diferenciar a escrita dos hispano falantes aprendizes de português da escrita dos nativos dessa língua.

A partir destas lacunas, e para analisar as dificuldades dos hispano falantes na produção escrita em português na Figura 5.2 são apresentadas as análises que envolveram textos do *córpus Espanhol-Acadêmico-Br*. Todas as análises da Figura 5.2 têm a finalidade de caracterizar a escrita dos hispano falantes, assim como identificar algumas formas de auxílio para aprimorar a

escrita em português desses pesquisadores.

Figura 5.2: Análises para caracterizar a produção escrita dos hispano falantes

Como indicado na Figura 5.2, primeiramente, foi proposta uma avaliação holística, em que textos de aprendizes e nativos foram classificados em níveis de proficiência. Essa avaliação é apresentada na Seção 5.3.1 e foi realizada empregando uma ferramenta de classificação iterativa de dados. O funcionamento da ferramenta será detalhado em conjunto com a avaliação, na qual os textos foram classificados em níveis de proficiência. Como resultado da avaliação holística uma rubrica para avaliar a qualidade dos textos foi proposta. Além disso, a avaliação apontou a necessidade de uma análise mais detalhada para conseguir caracterizar a escrita dos hispano falantes. O ponto de partida da análise mais detalhada é apresentado na Seção 5.3.2, que traz o estudo piloto que levantou as principais dificuldades e preocupações que os hispano falantes têm quando escrevem em português. Como resultado do estudo piloto foi obtida uma tipologia inicial de erros que foi utilizada na análise apresentada na Seção 5.3.3, na qual diversas categorias de erros foram anotadas em textos produzidos por nativos e aprendizes.

Como os textos de nativos do português foram incluídos nas análises apresentadas nesta seção, foi necessário compilar textos acadêmicos desse grupo de alunos. Para isso, foram selecionados resumos compilados por Feltrim

et al. (2004) no projeto SciPo (*Scientific Portuguese*) e resumos já publicados, disponíveis na internet. Os resumos disponíveis no projeto SciPo foram escritos por alunos nativos do português matriculados nos programas de Pós-graduação da USP de São Carlos e foram anotados seguindo a estruturação retórica dos componentes de um resumo científico. Mais informações relacionadas ao projeto SciPo estão disponíveis em Feltrim et al. (2004).

A seguir são apresentadas, em três seções, todas as análises desenvolvidas para caracterizar a escrita dos hispano falantes.

5.3.1 Avaliação Holística da Qualidade da Escrita

Etiquetar manualmente erros de diversas categorias em textos de aprendizes é uma tarefa demorada e complexa que requer treinamento. Em uma tentativa de simplificar a tarefa anterior, dado que uma das metas desta pesquisa é detectar as características que marcam a produção escrita dos aprendizes, foi selecionado um conjunto de textos de nativos e de aprendizes para uma avaliação holística dos textos. Embora essa tarefa requeresse anotação manual, ela era mais simples que a tarefa de anotar erros e suas correções.

O objetivo da avaliação holística foi fazer uma análise geral da qualidade dos textos que permitisse classificá-los em três níveis de proficiência: Iniciante, Intermediário e Avançado. Esperava-se que existisse uma correspondência entre os níveis de avaliação e os autores dos textos, de forma que, os textos dos nativos se concentrassem no grupo Intermediário ou Avançado, e textos dos aprendizes nos grupos Iniciante ou Intermediário. Confirmando a hipótese anterior seria possível traçar um perfil dos hispano falantes aprendizes de português.

A tarefa de avaliação automática da qualidade dos textos foi abordada como uma tarefa de classificação, em que os textos deviam ser classificados nos três níveis de proficiência: Avançado, Intermediário e Iniciante. Uma abordagem similar foi adotada por Tsvetkov et al. (2013) para classificar textos em inglês produzidos por nativos de 11 línguas diferentes. Tsvetkov et al. (2013) obtiveram bons resultados classificando os textos segundo as línguas dos não nativos.

A abordagem dominante para realizar a classificação de textos é baseada em algoritmos de aprendizado de máquina que automaticamente classificam os textos a partir de um conjunto de treinamento. A saída dos algoritmos de classificação depende de diversos fatores, dentre eles o espaço de características e o conjunto de treinamento empregado. Por isso, a qualidade do conjunto

de treinamento é fundamental para o desempenho do classificador. Visando superar a dificuldade de criar um conjunto grande de treinamento, que representasse adequadamente as classes do modelo, dado que a anotação manual é um processo caro que exige especialistas, foi utilizada a ferramenta de classificação visual de conjuntos, *Visual Classification System* (VCS) (Paiva et al., 2012). Essa ferramenta auxiliou na: (i) definição do conjunto de treinamento; (ii) construção e verificação do modelo de classificação e (iii) atualização iterativa e iterativa do modelo criado. A seguir, é detalhado o processo para selecionar os textos classificados manualmente e para construir e atualizar o conjunto de treinamento.

O sistema VCS disponibiliza um ambiente baseado em técnicas de visualização, que implementa o processo completo de classificação iterativa de dados. A ferramenta exibe a estrutura do conjunto de textos, como mostrado na Figura 5.3 e permite que sejam selecionadas instâncias de treinamento para criar modelos de classificação. Esses modelos de classificação podem ser usados para classificar outros conjuntos de dados, assim como analisar os resultados da classificação, quando há um esquema de classificação ideal, ou para compreender os motivos de eventuais falhas na classificação. Além disso, de forma iterativa pode-se atualizar os modelos de classificação e realizar uma nova classificação, com uma sequência de passos que levam ao desempenho esperado do algoritmo de classificação.

Na Tabela 5.5 é apresentada a rubrica de avaliação proposta para classificar os resumos nos níveis de proficiência. Essa rubrica é uma combinação das rubricas do exame CELPE-Bras e da rubrica apresentada em Aluísio et al. (2005), para avaliar a qualidade de textos acadêmicos. Os critérios de avaliação (coluna 1 da Tabela 5.5) foram escolhidos considerando os principais problemas que hispanofalantes aprendizes de português enfrentam quando escrevem nessa língua, relatados na Seção 2.3.

Utilizou-se um conjunto de 73 atributos para caracterizar os textos. O primeiro grupo de atributos foi derivado da ferramenta AZPORT (Feltrim et al., 2004) e foi utilizado para avaliar o critério de adequação ao gênero. O AZPORT é um sistema que auxilia na estruturação retórica de resumos científicos em português, classificando as sentenças dos resumos segundo os componentes que devem estar presentes em um resumo acadêmico: Contexto, Lacuna, Propósito, Metodologia, Resultado e Conclusões. Para classificar as sentenças do

⁵Para o sistema SciPo os componentes principais de um resumo acadêmico são Propósito, Metodologia, Resultado e Conclusões. A metodologia é opcional.

⁶Para o SciPo os componentes não principais são Contexto e Lacuna

Tabela 5.5: Rubrica de avaliação da qualidade da escrita

CrITÉRIOS de Avaliação			
Adequação ao gênero	Adequado (Presença dos componentes principais ⁵ , ausência de algum componente não principal ⁶)	Parcialmente Adequado (Pode faltar um componente principal e até dois não principais)	Inadequado (Falta a maioria dos componentes principais)
Clareza e Coesão Sentencial	Adequado (Sentenças sem problemas de coesão e coerência)	Parcialmente Adequado (Pode ter uma sentença com problema)	Inadequado (Mais de duas sentenças com erros)
Adequação Lexical e Gramatical	Adequado (Pode ter um ou dois erros)	Parcialmente Adequado (Pode ter entre três e quatro erros)	Inadequado (cinco ou mais erros)
Interferências da língua nativa	Adequado (Não tem interferência)	Parcialmente Adequado (Uma ou duas interferências)	Inadequado (Mais de 3 interferências)
Níveis Proficiência	Avançado	Intermediário	Iniciante

resumo, o SciPo emprega os seguintes sete atributos: tamanho da sentença, localização, presença de citações, presença de expressões formulaicas, tempo verbal e voz da sentença. Os sete atributos foram empregadas para descrever a primeira e a última sentença de cada resumo. Além disso, foi criada uma heurística que avalia os componentes que devem aparecer no resumo científico. A heurística é baseada na classificação realizada pelo AZPORT para cada sentença do resumo, resultando em 15 atributos, que representam a adequação ao gênero de cada texto.

O segundo grupo de atributos foi derivado do modelo de língua, apresentado na Seção 5.1.2. A partir do modelo de língua foram computados os seguintes dez atributos: probabilidade e perplexidade de unigramas, bigramas e trigramas, além de quantificar as palavras fora do vocabulário. Os atributos extraídos do modelo de língua tentam caracterizar alguns textos nos quais há expressões ou palavras do espanhol. Esse grupo de atributos foi empregado por Aluísio et al. (2010) para prever os níveis de inteligibilidade de textos.

O terceiro grupo de atributos foi obtido por meio da ferramenta Coh-Metrix-Port (Scarton, 2010), desenvolvida para o português e que disponibiliza atributos que vão desde métricas simples (como contagem de palavras) até medidas mais complexas, que automaticamente mensuram a complexidade de um texto. Essa ferramenta combina diversas medidas lexicais, sintáticas e semânticas com a finalidade de indicar a adequação de um texto a seu público-alvo ou gênero. Os 48 atributos implementados no Coh-Metrix-Port e explicadas detalhadamente em Scarton (2010) foram empregados para caracterizar os textos, em uma tentativa de caracterizar resumos muito simples (produzidos por aprendizes em níveis iniciais de aprendizagem, ou até textos muito complexos que mostram usos rebuscados da língua portuguesa). Esses atributos foram inseridos no processo como uma alternativa de identificar problemas relacionados com os critérios de Clareza e Coesão Sentencial, assim como com a Adequação Lexical e Gramatical.

Após calcular os 73 atributos para cada um dos textos, as instâncias obtidas são a entrada do sistema VCS. A Figura 5.3, mostra o *layout* do conjunto de textos. Esse *layout*, que foi produzido empregando a técnica de visualização baseada em árvore de similaridade, foi utilizado para extrair as instâncias inseridas no conjunto de treinamento. Veja que todas as instâncias do *layout* aparecem com uma cor única, o que significa que elas ainda não foram classificadas. A árvore de similaridade oferece informações sobre o conjunto de dados, por exemplo, o fato da árvore ter sete ramos poderia estar indicando

que as instâncias podem ser separadas em sete classes diferentes. Porém, as instâncias também poderiam ser separadas nas três regiões da árvore: acima, médio e baixo. A ideia é que a estrutura da árvore mostre possíveis distribuições das classes do conjunto, considerando as similaridades entre as instâncias. Segundo Paiva et al. (2012), as instâncias situadas nas partes mais externas das folhas da árvore são as que melhor caracterizam a classe à qual pertencem. Por outro lado, as instâncias posicionadas no centro da árvore são instâncias cujas características não se encaixam em nenhuma classe ou se encaixam em mais de uma classe ao mesmo tempo. Essa interpretação foi utilizada para selecionar as instâncias do conjunto de treinamento e avaliação.

Figura 5.3: Layout de textos de nativos e aprendizes

Durante a seleção das instâncias, a ferramenta VCS também permite que o usuário analise seu conteúdo, sendo possível classificar as instâncias. Como

mostrado na Figura 5.4, esta funcionalidade foi empregada para classificar manualmente os resumos selecionados nos três níveis de proficiência: Avançado, Intermediário e Iniciante. No total, foram anotados manualmente 144 resumos em português, 72 extraídos do corpus dos nativos, apresentado no começo desta seção e outros 72 extraídos do corpus Espanhol-Acadêmico-Br. Os textos classificados manualmente foram divididos em dois conjuntos disjuntos: Treinamento e Avaliação. Os resumos foram distribuídos de forma balanceada, sendo que cada conjunto continha 24 resumos de cada classe.

Figura 5.4: Classificação manual de um texto.

A Figura 5.5a representa o *layout* com as 72 instâncias que compõem o conjunto de treinamento. Esse *layout* também foi produzido empregando-se a técnica de visualização baseada em árvore de similaridade. A Figura 5.5a apresenta a distribuição, no espaço de características, das três classes do conjunto de treinamento, na qual as bolinhas vermelhas representam os textos classificados como Avançado, as azuis são os textos com classe Intermediário e as verdes são os textos da classe Iniciante. Analisando essa representação visual, pode-se prever um possível comportamento do classificador. Por exemplo, as instâncias da classe Intermediário aparecem misturadas com as instâncias das classes Avançado e Iniciante, essa distribuição representa uma

difficuldade para o modelo de classificação, ao mesmo tempo em que o fato dos textos Avançado e Iniciante aparecerem melhor distribuídos nas regiões mais externas da árvore representa um facilitador para o classificador, indicando que elas foram mais bem caracterizadas pelo conjunto de atributos.

Figura 5.5: Estrutura, Modelo de Classificação e *Class Matching* do conjunto de treinamento

As instâncias do conjunto de treinamento foram selecionadas para construir o modelo de classificação empregando o método *Partial Least Squares*⁷ (PLS) com fator⁸ 6. Como o conjunto de treinamento foi classificado previamente, o sistema VCS disponibiliza outra ferramenta chamada *Class Matching* que, por meio de um *layout* visual, mostra as instâncias classificadas correta e incorretamente. Esse *layout* é construído empregando-se as instancias vizinhas, de forma que ajuda o usuário a entender as razões pelas quais o modelo tomou uma determinada decisão.

Na Figura 5.5b, é mostrado o resultado visual do modelo de classificação no conjunto de treinamento. A representação do *class matching*, que apresenta os erros e os acertos do classificador, é mostrada na Figura 5.5c e os resultados numéricos desta classificação aparecem na coluna 2 da Tabela 5.6. As instâncias classificadas incorretamente são identificadas na Figura 5.5c (quatro bolinhas de cor vermelha) e comparando-se os resultados das Figuras 5.5a e 5.5b é possível observar que todas elas pertencem à classe Intermediária.

Após criar o modelo de classificação, ele foi testado na classificação de novas instâncias que integram o conjunto de avaliação. A representação visual do conjunto de avaliação (Figura 5.6a) mostra mais claramente como os

⁷PLS representa uma classe de métodos estatísticos utilizados para modelar relações entre conjuntos de variáveis multidimensionais através de um espaço latente de baixa dimensionalidade (Wold, 1985).

⁸O método PLS tem um parâmetro fator, que indica o número de componentes que melhor predizem a classe do elemento que está sendo analisado. Este parâmetro precisa ser ajustado até conseguir o classificador com melhor desempenho.

textos Intermediários aparecem misturados com os textos classificados como Avançados e Iniciantes. Esta análise visual confirma o resultado observado na classificação das instâncias de treinamento e pode-se prever que a maioria dos erros do modelo de classificação estará concentrada na classificação dos textos Intermediários. A análise visual, também, sugere que os atributos empregados não são adequados para categorizar os textos Intermediários.

Figura 5.6: Estrutura, atualização do modelo de classificação e *class matching* do conjunto de avaliação

Na Figura 5.6b, ilustra-se a classificação visual do conjunto de avaliação empregando-se o modelo criado a partir do conjunto de treinamento. Os resultados numéricos desta classificação aparecem na coluna 3 da Tabela 5.6, enquanto a Figura 5.6c mostra a representação do *class matching*, que mostra as 32 instâncias classificadas incorretamente. Considerando que os resultados da classificação de novas instâncias não foram bons (coluna 3 da Tabela 5.6) decidiu-se atualizar o modelo de treinamento, seguindo o princípio de classificação incremental. Para isso, o modelo de classificação foi atualizado por meio da inserção de instâncias do conjunto de avaliação. A intenção, a partir desta atualização, é obter um modelo que agregue informações de ambos os conjuntos e, conseqüentemente, melhore o desempenho do classificador.

A partir do *class matching* (Figura 5.6c), foram selecionadas 6 instancias,

Tabela 5.6: Resultados numéricos da classificação do conjunto de treinamento e do conjunto de avaliação.

Conjuntos	Treinamento	Avaliação	Atualização		
Modelo	Inicial	Inicial	1	2	3
Classificações Corretas	68-94,4%	40-55,6%	48-66,7%	52-72,2%	54-75%
Classificações In-corretas	4-5,6%	32-44,4%	24-33,3%	20-27,8%	18-25%
Acurácia	96,4%	74,13%	80,01%	83,15%	84,64%
Precisão	94,64%	55,15%	65,5%	72,54%	74,33%
Recall	94,44%	55,56%	66,67%	72,22%	75%
F1	0,95	0,55	0,66	0,72	0,75

duas instâncias de cada classe, que foram classificadas incorretamente. As instâncias foram selecionadas de forma balanceada, porque experimentalmente este critério de seleção manteve um bom desempenho do modelo. Na Figura 5.6d, mostram-se as instâncias selecionadas; dessa vez foram incluídas aquelas que aparecem mais distantes do centro da árvore, porque segundo Paiva et al. (2012) representam instâncias cujos atributos são bem característicos das classes nas quais estão inseridas.

O modelo atualizado com as 6 novas instâncias (iteração 1) foi testado com o conjunto de avaliação. A classificação obtida nesta iteração é mostrada na Figura 5.6e, enquanto os resultados numéricos aparecem na coluna 4 da Tabela 5.6. Comparando-se os resultados do primeiro modelo com os resultados da primeira iteração há uma diminuição da taxa de erro após a inserção de 6 novas instâncias.

Após os resultados da primeira iteração e considerando-se que ainda não é adequado o desempenho do classificador, foram realizadas outras duas iterações, cujos resultados visuais são ilustrados na Figura 5.6g a 5.6l e os resultados numéricos aparecem na coluna 5 e 6 da Tabela 5.6. Em cada uma das iterações realizadas foram inseridas 6 instâncias, mantendo-se o balanceamento do conjunto. Finalmente, o modelo de classificação foi construído com mais 18 instancias do conjunto de avaliação e o desempenho inicial foi de $f1=0,55$ para $f1=0,75$.

Embora observada uma melhora no desempenho do classificador após a atualização do modelo (resultados da Tabela 5.6), comparando-se o *layout* da Figura 5.6a (conjunto original) com os *layouts* nas Figuras 5.6b, 5.6e, 5.6h e 5.6k, obtidos nas classificações de cada iteração, é possível perceber que,

como previsto, a classificação das instancias da classe Intermediário possuem o maior percentual de erros. A interpretação dada para esse resultado está relacionada com o fato dos atributos usados não serem suficientes para caracterizar os textos da classe Intermediário.

Para verificar esta hipótese, foi construído um classificador binário que separa os textos das classes Avançado e Iniciante. Para isso, foram eliminadas as instâncias da classe Intermediário dos conjuntos de treinamento e avaliação. O modelo inicial foi construído, com PLS e fator três, a partir do conjunto de treinamento. Os resultados da classificação do modelo neste próprio conjunto são apresentados na coluna 2 da Tabela 5.7. A coluna 3, mostra o desempenho do modelo inicial na classificação das instâncias do conjunto de avaliação. Analisando-se esses resultados é possível confirmar que o desempenho do classificador binário é melhor que o classificador que precisa classificar as instâncias da classe Intermediária. Tentando melhorar a classificação do conjunto de avaliação, o modelo foi atualizado inserindo duas instâncias (uma de cada classe) deste conjunto. O resultado da atualização é mostrado na coluna 4 da Tabela 5.7. O resultado confirma que a inclusão de novos atributos é necessária para caracterizar os textos do nível intermediário e, conseqüentemente, aprimorar os resultados do classificador. Essa análise representa um dos trabalhos futuros desta pesquisa.

Tabela 5.7: Resultados numéricos da classificação do conjunto de treinamento e do conjunto de avaliação com duas classes: Avançado e Iniciante.

Classificador Binário			
Conjuntos	Treinamento	Avaliação	Atualização
Modelo	Inicial	Inicial	1
Classificações Corretas	48-100%	41-85,4%	43-89,6%
Classificações Incorretas	0-0%	7-14,6%	5-10,4%
Acurácia	100%	86,38%	90,09%
Precisão	100%	87,02%	90,21%
Recall	100%	85,42%	89,58%
F1	1	0,86	0,9

5.3.2 Análise Inicial de Erros de Hispano Falantes

Nesta seção, é apresentado um estudo inicial com base na análise de respostas a um questionário proposto para conhecer a opinião dos aprendizes de português e levantar suas principais preocupações na escrita acadêmica dessa língua.

No estudo foram selecionadas 13 introduções para avaliar a ocorrência de alguns erros; testar um processo de anotação e formalizar uma tipologia inicial com os erros identificados. Os resultados obtidos com a tipologia inicial, produzida no estudo piloto, foram utilizados para refinar a própria tipologia e formalizar a tipologia utilizada no processo de anotação posterior, que será explicado na Seção 5.3.3. Os resultados do estudo piloto foram publicados em Torres et al. (2014).

A primeira etapa do estudo inicial foi aplicar o questionário apresentado no Apêndice B aos hispano falantes engajados em programas de pós-graduação na USP de São Carlos. Um total de 66 estudantes, incluindo os autores dos textos do sub-cópus de Introduções, respondeu o questionário. O questionário é composto por 17 perguntas relacionadas aos problemas identificados por Grannier e Carvalho (2001) e Durão (1999), comentados nas Seções 2.3 e 2.3, respectivamente, do presente documento. Além disso, o questionário também considera aspectos associados à proficiência da língua. Na Figura 5.7, mostram-se os erros incluídos no questionário, assim como seu nível de dificuldade, segundo os hispano falantes aprendizes de português.

Figura 5.7: Quantificação das respostas apresentadas no questionário

Todos os problemas indicados na Figura 5.7 foram apontados como dificuldades na escrita. Os de maior preocupação foram: falsos cognatos, palavras compostas, pronomes complementos de verbos, 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. Problemas como o uso de artigos, confusão de grafemas, uso da crase, marca de nasalização e confusão de fonemas também foram considerados preocupantes para os hispano falantes. Entretanto, talvez porque sejam tratados pelos corretores ortográficos mais utilizados em processadores de textos, eles não apareceram no grupo de maior preocupação.

Após a geração das estatísticas das perguntas do questionário, o sub-cópus de Introduções foi analisado com base nos erros do questionário. O sub-cópus é composto por 617 sentenças e 17.795 palavras. A Tabela 5.8 mostra as estatísticas do sub-cópus de Introduções, assim como as áreas de pesquisa dos alunos que produziram os textos.

Tabela 5.8: Estatística do cópus de escrita acadêmica por aprendizes do português

Cópus de Introduções 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	I3	617	17795

Dois anotadores nativos do português analisaram cada texto e, conforme novos tipos de equívocos foram sendo identificados, novas categorias foram consideradas, incluindo o tipo de erro e uma sugestão para corrigi-lo. Alguns erros poderiam ter sido 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. Segundo Rojas (2006), a gravidade dos erros é medida em função de quanto o erro afeta a compreensão do texto; esse foi o mesmo critério seguido nesta pesquisa. Também foram considerados erros graves aqueles exemplos que mostraram interferência da língua materna. A partir da análise anterior, foi criada uma tipologia inicial de erros, apresentada na Figura 5.8. Na figura aparecem as categorias identificadas, assim como a ocorrência de cada tipo de erro no sub-cópus de Introduções.

No começo da análise dos textos, esperava-se que o sub-cópus 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 pro-

Figura 5.8: Quantificao dos erros dos aprendizes no crpus de Introdues

ficincia da lngua portuguesa. No entanto, aps a anotao, identificaram-se erros que corretores ortogrficos e/ou gramaticais tratam e outros que ainda no foram atendidos; em sua maioria porque so erros especficos de aprendizes de uma lngua. Desta forma, a anotao piloto mostrou autora desta pesquisa a necessidade de realizar um processo de anotao de novos textos, para obter um levantamento mais detalhado dos erros dos hispano falantes. A anotao piloto tambm foi o ponto inicial para formalizar a tipologia de erros apresentada na Figura 5.9. Essa tipologia de erros guiou o processo de anotao explicado na Seo 5.3.3 e representa uma das principais contribuies desta pesquisa. A tipologia de erros tambm detalhada por meio de exemplos, nesta seo.

Como pode ser observado na Figura 5.8, os erros mais recorrentes nem sempre correspondem aos indicados no questionrio, como o **uso de falsos cognatos**, que foi o mais inquietante na enquete e o de menor ocorrncia no crpus. Acredita-se que esse problema seja mais comum na lngua falada, enquanto na escrita no muito frequente. As listas de cognatos e falsos cognatos so recursos muito teis na aprendizagem de qualquer lngua, especialmente entre lnguas prximas, porque mostra semelhanas e diferenas entre as lnguas. Torres e Alusio (2011) apresentam um estudo para criar de forma

Figura 5.9: Tipologia de erros

automática dicionários de cognatos e falsos cognatos das línguas português e espanhol. Esse estudo é detalhado na Seção 6.2.1 deste documento. Embora os erros de falsos cognatos sejam importantes nesta pesquisa, a anotação manual desta categoria não foi considerada na tipologia final (apresentada na Figura 5.9), dada a baixa ocorrência desse tipo de erro no corpús. Além disso, acredita-se que a tarefa de identificação automática de erros de falsos cognatos seja uma tarefa complexa que requer recursos que não estão disponíveis para essa pesquisa. Mais detalhes são apresentados na Seção 6.2.1.

Na **categorias difusas** foram identificados erros relacionados com a composição de ideias claras e concisas; a construção de sentenças longas, com ausência de conectivos, de sinais de pontuação e com palavras repetidas. Esse tipo de problema foi frequentemente observado na escrita em português por hispano falantes, produzindo partes de um texto difícil de se ler e/ou entender. Embora os corretores ortográficos e gramaticais mais difundidos não corrijam esse problema, o uso de modelos de língua permite identificar automaticamente os problemas relacionados com a organização das palavras dentro de uma expressão. Além disso, com essa análise estatística dos textos é também possível identificar sentenças longas e com falta de conectivos.

Considerando-se que essa categoria de erros é subjetiva e que conhecer como os hispano falantes empregam os conectivos é de interesse nesta pesquisa, essa categoria foi dividida em duas na tipologia final: **Incoerente** e **Conectivos Errados**, como mostrado na Figura 5.9.

A categoria **separação e união de palavras** mostra que os nativos do espanhol apresentam problemas para produzir palavras compostas. Essa categoria também não foi incluída na tipologia final (apresentada na Figura 5.9), porque com o acordo ortográfico de 2010 a categoria mudou e os corpú de referência desta pesquisa não têm exemplos de palavras seguindo a nova mudança ortográfica.

A categoria **uso de vocabulário informal para o gênero científico** também foi eliminada da nova tipologia, pois esses vocabulários podem ser facilmente extraídos por meio de uma análise estatística de corpú de referência. Para tratar esse tipo de problema, o ideal seria compilar um corpú da língua falada e incluí-lo nas análises desta pesquisa, sendo possível indicar para o aprendiz quando um item lexical é mais comum na língua falada ou na língua escrita do gênero acadêmico.

Como mostrado na Figura 5.9, as categorias de erros foram agrupadas em três grandes grupos: Clareza e Coesão; Erros Gramaticais e Interferências. Essa tipologia final foi consolidada para anotar os textos dos aprendizes. Na Figura 5.9, observa-se que além das categorias que foram eliminadas no processo de refinamento da tipologia, incluíram-se novas categorias, como **Colocações**. Essa inclusão aconteceu porque foi observado, na análise do corpú de Introduções, que algumas das interferências do espanhol na escrita em português eram colocações da língua espanhola.

Para apresentar a tipologia de erros do projeto HABLA, usamos as Tabelas 5.9 e 5.10, as quais mostram cada categoria de erros e exemplos dos erros com suas correções. Todos os exemplos foram retirados do corpú e a correção aparece entre parênteses.

As tipologias apresentadas nas Tabelas 5.9 e 5.10 serviram de referência principal para dois anotadores humanos etiquetar os erros detectados no projeto HABLA. O processo completo da anotação, assim como os resultados obtidos são apresentados na Seção 5.3.3.

5.3.3 Anotação de Erros de Hispano Falantes e Nativos

Diferentemente do estudo piloto apresentado na Seção 5.3.2, na segunda anotação manual foram selecionados textos de nativos e de aprendizes. Con-

Tabela 5.9: Tipologia de Erros para anotar manualmente o corpus Espanhol-Acadêmico-Br

Categoria Clareza-Coesao
Conectivos-Errado Ex: ... estes estudos se escasseam , e por tanto (portanto) continuam silenciando todo um conjunto de aspectos ...
Incoerente Ex: De fato, SLSM constituem uma classe de sistemas lineares estocásticos que são relevantes para aplicações (vide por exemplo algumas referências) e apresentam um número de características paralelas àquelas dos sistemas lineares determinísticos, para uma abordagem com teoria de operadores para SLSM, como uma ilustração de trabalhos recentes e artigos inéditos; porém, controlabilidade foi considerada desde um ponto de vista único da trajetória.
Erros Gramaticais
Erro de Ortografia (Acentuação ortográfica equivocada; Marca equivocada de nasalização; Confusão de fonemas; Confusão de grafemas para o mesmo fonema) Ex: ... tem a tarefa de analisar um paragrafo (parágrafo) ou uma sentença de texto ...
Ausência ou uso inadequado da crase Ex: ... um processo altamente complexo quanto ao aspecto microbiológico e extremamente sensível as (às) condições ambientais...
Colocação pronominal (Posição dos pronomes oblíquos) Ex: Podese (pode-se) perceber a grande quantidade de interações ...
Uso incorreto de artigos Ex: Em particular, quanto mais alta é _ (a) taxa de compressão, ...
Conjugações verbais Ex: O Desenvolvimento Sustentável é definido como um modelo econômico, político, social, cultural e ambiental equilibrado, que satisfaz (satisfaz) as necessidades das gerações atuais. Ex: - Contudo, a extensão destes recursos não é tão significativa como se desejasse (deseja).
Concordância verbal Ex: A repartição agrária e as políticas públicas tem sido (têm sido) os principais motores da transformação dos bosques.
Uso incorreto de preposições Ex: - O recurso está sendo desenvolvido na base de (com base em) um pequeno corpus, em comparação ao corpus original.
Concordância nominal Ex: - ... conjuntos de técnicas de Mineração de Dados e Textos têm sido usadas (usados) para auxiliar ...
Uso incorreto da função da palavra Ex: - A digestão anaeróbia sofre grande influencia (influência) do regime hidráulico. - As técnicas de análise (análise) de DNA tais como

Tabela 5.10: Tipologia de Erros para anotar manualmente o corpus Espanhol-Acadêmico-Br

Interferências

Uso de palavras em espanhol

Ex: ... analisar e tornar visíveis las (as) múltiplas indetidades , la (a) diversidade de fins e a heterogeneidade ...

... com diferente grau de incidência prática na concreção (materialização) das propostas pedagógicas que oferecem ...

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).

Colocações

Ex: ... queremos pôr de manifesto (evidenciar) que não há unanimidade nas práticas ...

Convencionalidade

Ex: Como debilidades (pontos fracos), percebeu-se, principalmente, que o tempo de desenvolvimento de atividades não foi suficiente para o fomento (desenvolvimento) das competências desejadas.

siderando a produção escrita dos nativos e dos aprendizes é possível apontar os erros comuns e focar nos erros que são próprios dos hispano falantes. A análise permitiu traçar um perfil dos aprendizes, com base na identificação dos erros produzidos pelas interferências da língua materna: o espanhol.

Seleção dos Textos para Anotação Manual

Para a anotação manual foram selecionados 60 textos de aprendizes e 40 textos de nativos. Foram selecionados mais textos de aprendizes, porque o principal objetivo era identificar os erros desse grupo. Na Tabela 5.11 apresenta-se estatísticas do corpus compilado para anotação dos erros. Os textos de aprendizes anotados representam 5% do total dos textos que aparecem na Tabela 5.4. Nesse conjunto foram incluídos 29 textos do sub-corpus USP, sendo o resto do sub-corpus *Educación e Interciencia*.

Os textos selecionados para anotação manual representam um conjunto disjunto dos textos considerados no estudo piloto, porém todos eles foram incluídos na avaliação holística e são de diferentes níveis de proficiência. Como mostrado na Tabela 5.11, o conjunto de textos de aprendizes selecionado é composto por 484 sentenças e 12.964 palavras, enquanto o grupo com os textos dos nativos contém 187 sentenças e 5.106 palavras. No total, foram anotados 100 textos com 671 sentenças e 18.070 palavras.

Desbalanceamentos são observados nas tarefas de anotação de erros, sendo

sempre minoritário o número de textos com erros. Devido ao fato de não ter conhecimento prévio da ocorrência de erros nos textos, nem da diversidade dos tipos de erros, a anotação realizada para a avaliação holística foi usada como referência para selecionar o conjunto de textos que aparecem na Tabela 5.11.

Tabela 5.11: Estatísticas do corpus anotado manualmente

Cópus Anotado Manualmente				
	Textos	Tokens	Palavras	Sentenças
Nativos	40	5.944	5.106	187
Aprendizes	60	15.322	12.964	484
Total	100	21.266	18.070	671

Detalhes do Processo de Anotação

Os textos foram anotados via ferramenta Brat⁹. A ferramenta permite personalizar o processo de anotação, a tipologia de erros e a distribuição dos textos para cada anotador. Os textos foram anotados de forma independente usando um navegador Web. Na página do projeto HABLA⁶ está disponível a anotação apresentada nesta seção. Na Figura 5.10, mostramos um texto anotado com diferentes categorias de erros. Cada categoria de erro aparece em uma cor diferente, pré-selecionada, facilitando a análise após a anotação.

A anotação incluiu as seguintes tarefas:

- **Seleção dos erros:** selecionar trechos de textos contíguos usando o cursor para identificar o erro;
- **Categorização dos erros:** categorizar o erro segundo o tipo de erro, apresentado na tipologia da Figura 5.9;
- **Correção dos erros:** propor uma correção para o erro.

Duas anotadoras¹⁰, especialistas na área de ensino de português para não nativos, foram treinadas para realizar a tarefa de anotação. Ambas as anotadoras são falantes nativas do português e uma delas possui amplo conhecimento da língua espanhola. O fato de dispor de uma anotadora com conhecimento de L1 e L2 foi uma vantagem porque, em geral, ela conseguiu reconhecer melhor a natureza do erro produzido pela interferência da língua materna

⁹<http://brat.nlplab.org/>

⁶disponível em <http://nilc.icmc.usp.br/habla/>

¹⁰Aline Evers e Roana Rodrigues

Figura 5.10: Anotação de um texto com a ferramenta Brat

e, desta forma, teve maior facilidade para categorizar e corrigir o erro. As duas anotadoras receberam um treinamento antes de começar a tarefa de anotação que foi apoiada por um manual de anotação, apresentado no Apêndice B, com vários exemplos do processo.

Devido à similaridade entre o português e o espanhol, muitas vezes foi difícil identificar erros produzidos pela interferência do espanhol. Para garantir a qualidade da anotação, antes da etiquetagem dos erros, o manual inclui o passo de leitura completa do texto, sem focar em elementos específicos. Desta forma, o anotador consegue ter uma visão geral do texto. Além disso, foi sugerido fazer uma leitura detalhada de cada sentença e só depois começar a anotação dos erros. Cada categoria de erros foi anotada de forma independente; em outras palavras, a etiquetagem ocorreu por categorias.

Uma das decisões mais difíceis do processo de anotação foi identificar o começo e o fim do erro. Frequentemente, diferentes trechos de um texto correspondiam ao mesmo erro. Na Figura 5.11 mostra-se como os anotadores, referindo-se ao mesmo erro, sublinharam trechos de textos diferentes, indicando a mesma categorização para o erro. Para minimizar esse problema,

foi sugerido apontar como erro o trecho mínimo de texto, isto é, considerar a modificação mínima do texto para corrigir o erro.

Figura 5.11: Anotação manual de trechos diferentes de textos

No exemplo da Figura 5.11, a anotação mais adequada, considerando as indicações do manual, foi realizada pelo Anotador 1, mas a marcação do Anotador 2 não está errada, porque a expressão "reforma agrária" é uma colocação tanto no espanhol, quanto no português. O exemplo da Figura 5.11 mostra a complexidade da tarefa de comparar as anotações manuais de dois ou mais anotadores. No exemplo, é possível observar como a divergência na seleção do erro provoca divergências na concordância dos anotadores, embora as anotações do erro estejam corretas.

Nesta anotação de erros, o coeficiente *Kappa* foi utilizado para estimar a concordância dos anotadores. Desta forma, a concordância da anotação foi calculada para as três anotações requeridas: (i) seleção do erro; (ii) categorização do erro e (iii) correção do erro. Os detalhes são apresentados abaixo.

A ferramenta Brat quantifica a posição de cada caractere dentro do texto, e para cada erro anotado, atribui o intervalo de valores referente aos caracteres incluídos na anotação. No exemplo da Figura 5.11, à palavra *repartição* foi atribuído o intervalo de valores [871 - 881], enquanto que à expressão *repartição agrária* foi atribuído o intervalo [871 - 889]. O valor do *Kappa* da tarefa de seleção de erros foi calculado considerando os intervalos de anotação. Desta forma, a máxima concordância foi alcançada quando os dois intervalos eram idênticos ou compartilhavam o valor inicial ou o valor final, como aconteceu no exemplo da Figura 5.11. A confiabilidade da tarefa de seleção dos erros foi estimada considerando os intervalos com e sem erros, isto é, o cálculo do *Kappa* foi realizado considerando-se as palavras anotadas e as não anotadas. O valor do *Kappa* na seleção do erro foi de 0,77.

Considerando a complexidade da tarefa de anotar diferentes categorias de erros e comparando o *Kappa* obtido nesta anotação com o valor obtido por Dahlmeier et al. (2013) e por Leacock et al. (2010), o valor de *Kappa* de 0,77 indica um alto grau de concordância entre os anotadores, sendo um resultado aceitável no presente estudo. O resultado teria sido pior se só fossem consideradas as marcações com intervalos idênticos, mas devido à subjetividade da tarefa de seleção dos erros a autora considerou válido o critério comparativo.

O cálculo do *Kappa* na categorização dos erros foi realizado nos pares em que os dois anotadores concordaram na identificação do erro. O valor obtido foi 0,80, sendo a tarefa de maior concordância entre os anotadores. Resultado semelhante foi observado no estudo de Dahlmeier et al. (2013).

Na Figura 5.12 é apresentado um exemplo que indica discordância entre os anotadores na hora de categorizar o erro. O Anotador 1 categorizou o erro como um erro de ortografia, enquanto o Anotador 2 considerou como um erro de um conectivo usado de forma errada. Devido ao fato da expressão "*por tanto*" estar conectando duas ideias e considerando que no espanhol existe o marcador discursivo "*por tanto*" a anotação mais adequada, segundo o indicado no manual de anotação, era apontar o erro como sendo da categoria "Conectivo Errado". Entretanto, o erro também pode ser considerado como um erro de ortografia em que o marcador discursivo em português "*portanto*" foi separado. O exemplo anterior indica que a tarefa de categorizar o erro, também, é subjetiva. Além disso, confirma a importância do anotador conhecer a língua nativa do autor do texto. No exemplo da Figura 5.12, o Anotador 2 era o especialista em espanhol e reconheceu que "*por tanto*" é um marcador em espanhol. É importante apontar que a sentença da Figura 5.12 foi produzida por um hispano falante e que os anotadores não tinham essa informação quando estavam anotando os textos. Essa informação era detectada caso alguma marca de espanhol acontecesse no texto.

Por último, foi calculado o valor de *Kappa* para a tarefa de correção do erro. Foi estimado o *Kappa* para aqueles casos em que os anotadores tivessem concordado tanto na identificação do erro quanto na categorização. Como na maioria das vezes a correção do erro está relacionada com o tipo de erro, o exemplo apresentado na Figura 5.12 não foi considerado nesse cálculo, mesmo tendo os anotadores feito a mesma correção. Porém, o exemplo da Figura 5.11, no qual não houve concordância entre os anotadores na hora de corrigir o erro, foi considerado no cálculo. O valor do *Kappa* na tarefa de correção do erro foi de 0,68. Esse valor, também, é considerado adequado

Figura 5.12: Anotação manual de trechos diferentes de textos

partindo do princípio que esta tarefa é complexa, pois um único erro pode ter mais de uma correção.

Estatísticas da Anotação

Nesta seção, as estatísticas da tarefa de anotação são apresentadas. Especificamente, são apresentados os erros mais frequentes, tanto dos aprendizes como dos nativos, assim como os erros que são comuns para ambos. Adicionalmente, é verificada a ocorrência dos erros detectados no projeto HABLA com os erros identificados em outras pesquisas. Finalmente, são comentados os erros detectados que não foram cobertos pelo corretor ortográfico e gramatical Regra (Nunes et al., 1996; Martins et al., 1998; Bonfante e Nunes, 2000; Pinheiro et al., 2001; Hasegawa et al., 2002), que é o mais utilizado para corrigir textos na língua portuguesa.

As estatísticas apresentadas a seguir foram calculadas considerando o total de erros identificados pelos dois anotadores, que foi 370 erros. Para tornar mais clara a apresentação, as estatísticas dos erros dos aprendizes são mostradas separadamente das estatísticas dos erros dos nativos. É importante apontar que os totais de sentenças e palavras dos textos dos nativos e dos aprendizes eram diferentes, desta forma os valores foram normalizados para permitir uma comparação dos resultados.

Do total de 370 erros, só 33 erros foram produzidos pelos nativos do português, enquanto 337 erros foram produzidos pelos hispano falantes. Considerando que o total de *tokens* nos textos dos nativos é de 5.944 e o mesmo total nos textos dos aprendizes é de 15.322, os erros anotados representam 0,55% e 2,19% respectivamente. Esses valores indicam que o número de erros do

cópus anotado é muito pequeno, mas devemos levar em consideração que os textos, provavelmente, já foram revisados por corretores gramaticais.

A Figura 5.13a mostra a quantidade de erros por texto dos hispano falantes, enquanto a Figura 5.13b mostra a quantidade de erros por resumo dos nativos. Como mostra a Figura 5.13b há 25 resumos sem erros no grupo dos nativos e nenhum dos resumos com erros têm mais de dez erros. Já no grupo dos aprendizes, somente quinze resumos não têm erros, mas há onze textos com mais de dez erros. Entre os textos dos aprendizes há 22 textos que apresentam entre três e dez erros, enquanto doze textos têm entre um e três erros. Esses números indicam que, em geral, quando o aprendiz comete o primeiro erro, a tendência é que ele vai cometer outros erros. Esses resultados, também, mostram que a taxa de erros dos resumos dos hispano falantes é maior que a taxa de erro dos nativos.

(a) Hispano falantes: Total de erros por documento

(b) Nativos: Total de erros por documento

Figura 5.13: Total de erros por documento nos textos dos nativos e dos aprendizes

Na Tabela 5.12 aparecem os erros identificados nos resumos dos nativos. Os erros estão ordenados segundo a sua ocorrência, sendo que o erro mais frequente foi o uso incorreto de preposições, seguido de erros de ortografia, colocação pronominal e conectivos errados. Outros tipos de erros que também aconteceram nos textos dos nativos, desta vez com uma frequência tão baixa que podem ser considerados erros que aconteceram ao acaso, foram: concordância nominal, convencionalidade, uso incorreto de artigos e concordância verbal. É importante apontar que a ocorrência desses erros não é muito representativa, mas pode indicar que alguns nativos cometem, com frequência, erros gramaticais e podem indicar, também, que ainda há alguns erros que não são cobertos pelos corretores gramaticais mais utilizados. Na Figura 5.14 apresenta-se um gráfico de barras com a mesma informação da Tabela 5.12.

Na Figura 5.15 apresentam-se exemplos de erros produzidos pelos nativos

Tabela 5.12: Erros cometidos pelos nativos e etiquetados na anotação manual

Lista ordenada dos erros cometidos pelos nativos		
Uso Incorreto de Preposições	8	29.62%
Erro Ortográfico	5	18.51%
Colocação Pronominal	4	14.81%
Conetivos Errados	3	11.11%
Ausência Uso Inadequado da Crase	2	7.41%
Concordância Nominal	2	7.41%
Convencionalidade	1	3.70%
Uso Incorreto de Artigos	1	3.70%
Concordância Verbal	1	3.70%

Figura 5.14: Erros cometidos pelos nativos e etiquetados na anotação manual

de português. Cada um desses exemplos foi testado no corretor gramatical Regra e o único exemplo que a ferramenta conseguiu corrigir foi o Exemplo 3. Esses exemplos mostram que ainda há problemas que os corretores gramaticais não conseguem corrigir.

Na Tabela 5.13, apresentam-se os erros etiquetados nos textos produzidos pelos hispano falantes. Nessa tabela é possível observar, de forma ordenada, a frequência com que os erros aconteceram. Nos resumos dos hispano falantes todas as categorias de erros foram identificadas, exceto a categoria "uso incorreto da função da palavra", que não apareceu no cópús selecionado neste estudo. Na Figura 5.16 mostra-se um gráfico de barras com a percentagem de

Exemplo 1:	Convencionalidade:
O plano de gerenciamento de projeto de software proposto refere-se ao desenvolvimento de um sistema de informação na pós-graduação, <u>em termos de</u> mestrado, em uma Universidade pública.	
↓	
Correção: como pesquisa de	
Exemplo 2:	Uso Incorreto de Proposições:
A outra contribuição, caracterizada por aspectos abstratos e abordagem inovadora, é a definição de uma arquitetura de software reflexiva e um conjunto de padrões de projeto relacionados <u>para a</u> implementação de mecanismos de tratamento de exceções.	
↓	
Correção: com a; à	
Exemplo 3:	Colocação Pronominal:
Este é um problema presente em uma grande variedade de aplicações científicas e de engenharia e cuja solução <u>tornase</u> uma tarefa ...	
↓	
Correção: torna-se	
Exemplo 4:	Conetivos Errados:
A detecção de anomalias é uma abordagem de análise usada na detecção de intrusão <u>onde</u> se assume que a presença de anomalias no tráfego ...	
↓	
Correção: em que	

Figura 5.15: Erros cometidos pelos nativos e etiquetados na anotação manual

erros nos textos dos aprendizes.

Após a anotação, verificou-se que há uma intersecção entre os erros detectados nos resumos dos aprendizes e dos nativos, porém na maioria dos casos a natureza dos erros é diferente. Sabe-se que a natureza do erro influencia diretamente no sucesso da correção do erro. Além disso, segundo Carrió-Pastor e Mestre-Mestre (2013) a maioria dos erros em L2 poderiam ser evitados caso existisse conhecimento sobre como acontece a comunicação dos aprendizes em L2. Seguindo essa linha, nesta pesquisa houve um crescente interesse por levantar os erros lexicais que representam interferências do espanhol na escrita em português.

Considerando a classificação das causas dos erros lexicais, comentada na Seção 1.2, nos resumos dos aprendizes foram identificados erros interlinguais e intralinguais. Os erros produzidos pela interferência da língua nativa foram observados em todas as categorias de erros; os erros lexicais mostraram:

1. palavras e/ou expressões em espanhol que não existem em português;
2. palavras e/ou expressões que não existem nem em espanhol e nem em

português;

3. palavras e/ou expressões escolhidas de forma errada.

Outros erros identificados apresentaram falhas na semântica das expressões, tais como colocações ou escolhas erradas. O corretor gramatical Regra não detecta alguns dos erros anteriores, mas o que é pior ainda, caso detecte o erro, na maioria das vezes não consegue corrigi-lo adequadamente. Por exemplo, os erros das categorias (1) e (2) provavelmente são detectados pelo Regra, porque não existem no léxico do português, mas como são produzidos pela interferência da língua nativa, o corretor Regra, que foi projetado para auxiliar nativos do português, não vai conseguir oferecer a correção adequada. Por outro lado, os erros incluídos na categoria (3) são mais difíceis de serem identificados por abordagens baseadas em dicionários porque existem no léxico em português e, conseqüentemente, não é possível oferecer nem sequer um alerta do possível erro – esse tipo de erro é chamado de sensível ao contexto ou *real-word error* (Fossati e Di Eugenio, 2007, 2008; Mays et al., 1991; Wilcox-O’Hearn et al., 2008) em contraste com os chamados de *non-word errors*.

A Figura 5.16 mostra os erros de maior ocorrência: erros de ortográfica, uso inadequado de preposições, convencionalidade e conjugações verbais. Da mesma forma como aconteceu com os nativos, o uso de preposições e os erros de ortografia foram categorias frequentes. Esses erros, junto aos erros de conjugações verbais, também se destacaram na anotação do estudo piloto, apresentado na Seção 5.3.2.

Na Figura 5.17 mostramos exemplos das quatro categorias de erros mais frequentes. No Exemplo 1 da 5.17 observa-se como o advérbio "historicamente" aparece com acento. Esse problema é muito frequente entre os hispano falantes e é uma interferência clara do espanhol, devido ao fato de que em espanhol os advérbios terminados em "-mente" levam acento se os adjetivos sem o sufixo "-mente" também levam. O segundo erro do Exemplo 1 é um erro de digitação que muitas vezes está associado à pressa na hora de escrever, porém ele também pode ser considerado um erro de pronúncia. Ambos os exemplos são detectados e corrigidos pelo Regra, porque são palavras não incluídas no vocabulário do português.

O Exemplo 2 da Figura 5.17 mostra como os hispano falantes erram no uso de expressões fixas da língua portuguesa que são importantes para aumentar a fluência dos textos. Erros desse tipo poderiam ser evitados se os aprendizes tivessem acesso a um léxico com as expressões fixas utilizadas em textos científicos em português, para observar exemplos de uso.

Tabela 5.13: Erros cometidos pelos aprendizes e etiquetados na anotação manual

Lista ordenada dos erros cometidos pelos aprendizes		
Erro Ortográfico	55	21.15%
Uso Incorreto de Preposições	35	13.46%
Convencionalidade	34	13.07%
Conjugações Verbais	31	11.92%
Uso de Palavras em Espanhol	17	6.53%
Incoerente	15	5.76%
Concordância Nominal	14	5.38%
Conetivos Errados	13	5%
Uso Incorreto de Artigos	13	5%
Concordância Verbal	13	5%
Colocação Pronominal	9	3.46%
Colocações	5	1.92%
Ausência Uso Inadequado da Crase	4	1.53%
Mudança de Gênero	2	0.76%

Figura 5.16: Erros cometidos pelos aprendizes e etiquetados na anotação manual

O exemplo 3 da Figura 5.17 não é um erro do português, mas sim uma inadequação que marca desvios na escrita. No exemplo 3, o aprendiz usou o verbo em português "outorgar"; acredita-se que a escolha desse verbo foi guiada pela similaridade entre as duas línguas. Em espanhol existe o verbo "otorgar", que é utilizado frequentemente, porém o verbo "outorgar" em português é mais comum na linguagem jurídica. Como indicado na correção, é preferível usar "conceder". Na Seção 6.2.2 é apresentado um recurso que

auxilia na escrita de verbos usados frequentemente na escrita acadêmica, ao mesmo tempo que evita equívocos como o exemplificado anteriormente.

No Exemplo 4 da Figura 5.17, observa-se como os hispano falantes costumam usar com muita frequência e de forma inadequada verbos no subjuntivo, fenômeno que não acontece entre os nativos. Esse último exemplo não é um erro lexical, representa uma interferência do espanhol tão recorrente que é considerada uma marca do discurso dos hispano falantes. Os exemplos 2,3 e 4 não são detectados pelo corretor Regra.

Figura 5.17: Erros cometidos pelos aprendizes e etiquetados na anotação manual

Após as quatro categorias mais frequentes, as seguintes quatro categorias foram identificadas nos resumos dos aprendizes: uso de palavras em espanhol, incoerente, concordância nominal e conetivos errados. A categoria incoerente aconteceu unicamente nos textos produzidos pelos hispano falantes e, logicamente, a mesma coisa aconteceu com a categoria uso de palavras em espanhol. Devido ao fato da categoria concordância nominal ser uma categoria tratada no corretor gramatical Regra, ela não será comentada nesta seção. A categoria incoerente, também não será tratada, pois é uma categoria que foge do escopo desta pesquisa, dado que os recursos para detecção de um texto sem coerência não recaem no nível lexical somente. No entanto, sentenças in-

coerentes aconteceram com frequência nos textos analisados, de maneira que é um problema que deve ser aprofundado em estudos futuros porque afeta a compreensão dos textos e, por isso, é um problema "grave", que marca o discurso em português dos hispano falantes.

Na Figura 5.18 são apresentados exemplos que mostram a presença de palavras e expressões em espanhol nos textos. Além disso, na figura aparecem exemplos de conectivos errados que, em alguns casos, representam interferências do espanhol.

Uso de palavra em espanhol:

Exemplo 1:
 Não há comunidades humanas dentro da reserva, mas na sua zona aledanha a maior parte das terras pertencem a camponeses ejidatarios, e na região colindante com o mar existem hotéis, casas de praia ...

↓
Correção: herdeiros de terras

↓
Correção: próxima ao

↓
Correção: limítrofe

Exemplo 2:

Conectivos Errados:

Embora que num princípio o impacto ambiental desta atividade fosse de acordo com a resiliência do ambiente ...

↓

Correção: embora

Exemplo 3:
 ... abordagem que tem a vantagem de modelar facilmente padrões complexos de entidades relacionadas e, **por tanto**, evitar o crescimento intratável de variáveis.

↓

Correção: portanto

Exemplo 4:
 Sobre os convênios tecnológicos, estes podem ser outorgados em quanto o projeto está em desenvolvimento. Exemplos de este tipo são:

↓
Correção: deste

↓
Correção: enquanto

Figura 5.18: Erros cometidos pelos aprendizes e etiquetados na anotação manual

Os erros do Exemplo 1 da Figura 5.18 são interferências do espanhol que não são corrigidas pelo Regra, mas o corretor consegue detectá-los. A palavra "aledanha" não existe nem em espanhol nem em português, porém a expressão "zona *aledaña*" é uma colocação do espanhol que no português significa "zona limítrofe" ou "ao redor". O erro aconteceu após o aluno trocar a "ñ" pela "nh" e gerar uma nova palavra que não existe em português. Em alguns casos, trocar a "ñ" pela "nh" funciona bem, por exemplo "araña" em espanhol é "aranha" em português. O fato anterior, mostra o raciocínio do aprendiz e uma possível causa do erro, mas não ajuda na sua correção. O aluno fez uma tradução literal de uma expressão em espanhol e, além disso, gerou uma nova

palavra em português. O exemplo mostra a dificuldade da tarefa de oferecer sugestões corretas, pois para esse erro seria necessário analisar o contexto do erro em português e relacioná-lo com sua forma em espanhol. Os outros erros que aconteceram no Exemplo 1 são similares, com a diferença das duas palavras serem do espanhol. O Exemplo 1 mostra como é difícil corrigir erros de aprendizes de uma segunda língua, pois mais de um erro pode acontecer em um único trecho.

No corpus Espanhol-Acadêmico-Br, foram numerosos os erros relacionados com a produção de adjetivos e substantivos com interferência da língua nativa; como consequência, foram identificadas algumas das estratégias de transferências entre o espanhol e o português aplicadas pelos hispano falantes. Em alguns casos as estratégias eram bem sucedidas, mas em outros provocavam erros. Esses problemas também foram incluídos no grupo de problemas que marcaram o discurso em português dos hispano falantes. Um método para descobrir e corrigir palavras erradas produzidas pela troca de sufixos entre o espanhol e o português é apresentado na Seção 6.1.3.

Os erros ilustrados nos Exemplos 2,3 e 4 da Figura 5.18 são erros relacionados com conectivos. Nenhum desses erros é identificado pelo corretor Regra, pois todas as expressões etiquetadas como erros existem em português, mas nunca funcionando como conectivos. No exemplo 4, mostra-se como os hispano falantes têm problemas no uso das contrações em português. Os outros exemplos mostram como existem interferências do espanhol na produção escrita de marcadores discursivos, um grupo especial de conectivos. O erro do Exemplo 3 apresenta o marcador "por tanto" do espanhol, que em português é "portanto". O hispano falante pode não saber ou não lembrar que em português se escreve junto; um erro similar acontece no Exemplo 4.

Na produção dos marcadores discursivos em português, foi observada a estratégia dos hispano falantes em "arriscar" traduções literais desses itens lexicais. Essa estratégia funciona em grande parte das vezes, mas há casos em que ela falha, gerando erros. É extremamente comum, por exemplo, encontrarmos marcadores discursivos como, "*Em concreto...*", "*De uma parte..., de outra parte, ...*" quando os marcadores equivalentes em português são, respectivamente, "Na verdade", "De/Por um lado..., de/por outro lado, ...". Esses problemas estão relacionados com a interferência da língua nativa e na Seção 6.2.3 são propostos métodos computacionais para automaticamente tratar esse tipo de erro. Equívocos relacionados com a semântica dos marcadores discursivos não foram observados nos textos anotados manualmente.

Juntamente com a anotação manual de erros de marcadores discursivos, foi realizada uma análise para comparar o uso dos marcadores discursivos entre nativos e aprendizes. Na análise foram selecionados seis marcadores não ambíguos e com a mesma função em português para análise de suas ocorrências no corpus de aprendizes Espanhol-Acadêmico-Br, Tabela 5.4 e no corpus Fapesp, apresentado na Seção 5.1.1. Na Figura 5.19 mostra-se a distribuição desses marcadores por tipo de uso (no início da sentença e precedido de vírgula) separados pelos grupos de textos de nativos e aprendizes.

Figura 5.19: Nativos e Aprendizes: Distribuição da posição em que são usados os marcadores

Comparando-se a posição de uso de cada marcador, observa-se que os aprendizes utilizam prioritariamente os marcadores em posição inicial, pois é a construção mais simples, ao contrário dos nativos, que usam mais os marcadores entre as duas orações relacionadas. Já na Figura 5.20, comparou-se a frequência de uso dos mesmos marcadores nos grupos de textos de nativos e aprendizes.

Observa-se que os aprendizes utilizam mais os marcadores “todavia” e “não obstante” do que os nativos, provavelmente porque ambos possuem cognatos em espanhol. Por outro lado, aprendizes usam muito menos “porém” que os nativos. Embora não constituam erros, essas diferenças de uso “marcam” o discurso dos aprendizes, distanciando-o do discurso produzido por nativos.

Finalizando, as últimas categorias identificadas foram: uso incorreto de artigos, concordância verbal, colocação pronominal, colocações, ausência ou uso inadequado da crase e mudança de gênero. Para tornar a apresentação dos resultados mais clara, na Figura 5.21 só serão apresentados exemplos das

Figura 5.20: Nativos e Aprendizes: Frequência de uso de marcadores por nativos e aprendizes

categorias de erros que aconteceram unicamente nos textos dos não nativos, como foi o caso da categoria colocações e mudança de gênero.

É importante apontar que a maioria dos problemas relacionados com a concordância verbal foi na conjugação de verbos, tais como: ter e conter, os quais na sua forma no plural levam o acento circunflexo e no singular o acento agudo. Acredita-se que muitos hispano falantes não conheçam a regra anterior ou não se lembrem dela quando precisam usá-la. Já, com relação à categoria colocação pronominal, os hispano falantes parecem não saber ou não se lembrar que no português nenhuma sentença começa com um pronome, por exemplo, os pronomes "se" ou "lhe", sendo esse o erro de maior frequência nessa categoria. Esses dois tipos de erros são identificados pelo Regra, contudo acreditamos que apresentar essas regras para os hispano falantes, em forma de alerta, seja um bom caminho para a correção desses problemas.

Os exemplos apresentados na Figura 5.21 mostram interferências do espanhol que não são identificadas pelo corretor gramatical Regra. Nos Exemplos 1 e 2, os autores geraram frases em português a partir de colocações da língua espanhola: "*hacer eco*" e "*poner de manifesto*", respectivamente. A ocorrência dessas expressões fixas do espanhol nos textos em português mostra que os aprendizes querem empregar expressões fixas, mas não conhecem as equivalências em português.

Na Figura 5.22, apresenta-se um gráfico que ilustra os erros comuns entre nativos e hispano falantes aprendizes de português nos textos anotados nesta pesquisa. Por meio da figura é possível observar como todos os erros identificados nos textos dos nativos também foram identificados nos textos

Figura 5.21: Erros cometidos pelos aprendizes e etiquetados no processo de anotação

produzidos pelos aprendizes. Porém, há um grupo de erros que somente foram identificados nos resumos escritos por aprendizes; nesse grupo há erros que foram produzidos por interferências do espanhol na escrita em português e não são tratados pelo corretor gramatical Regra. Mas há outros exemplos de erros que são identificados e corrigidos pelo corretor mais utilizado na escrita em português.

Figura 5.22: Erros comuns e não comuns entre os nativos e os aprendizes nos textos anotados

5.4 Considerações Finais

Este capítulo foi separado em três grandes seções que formalizam a base de conhecimento utilizada neste projeto para gerar os recursos computacionais.

Na Seção 5.1 foram apresentados os *corpus* empregados para gerar os léxicos bilíngues que apoiam o auxílio à escrita em português dos hispano falantes. Os *corpus* selecionados são úteis para mostrar como os nativos usam a língua portuguesa, da mesma forma em que os textos em espanhol do *corpus* paralelo, descrevem a escrita nessa outra língua. Embora o *corpus* paralelo seja do gênero jornalístico científico, ele é de grande utilidade nesta pesquisa, pois os seus textos descrevem pesquisas acadêmicas e neles aparecem os léxicos mais comuns desse gênero. Além do *corpus* paralelo, foi compilado um *corpus* com textos acadêmicos para auxiliar a construção dos léxicos bilíngues. Esse *corpus* aparece como alternativa para trazer palavras e/ou expressões que não aparecem no *corpus* paralelo. Esse último recurso só inclui textos em português e, conseqüentemente, pode provocar algumas limitações na aplicação dos métodos propostos, devido ao fato de não conter as equivalências em espanhol. Para finalizar, como será mostrado no Capítulo 6, o tamanho dos *corpus* utilizados influencia diretamente no desempenho dos métodos propostos, porém, os recursos apresentados nesta seção são os mais adequados para apoiar as tarefas desenvolvidas nesta pesquisa, considerando os *corpus* disponíveis na literatura para as línguas espanhola e portuguesa.

Na Seção 5.2 foi apresentado o *corpus* de aprendizes de português Espanhol-Acadêmico-Br. Esse *corpus* vai estar disponível publicamente na página do projeto HABLA⁷ e poderá ser utilizado em outras pesquisas. Além disso, por meio do *corpus* foram levantadas características da escrita dos hispano falantes no nível lexical, um dos objetivos desta pesquisa. As características da produção escrita dos hispano falantes foram identificadas a partir de diversas análises, apresentadas ao longo da Seção 5.3.

A Seção 5.3 foi dividida em várias sub-seções, que apresentaram diversas análises realizadas com a finalidade de caracterizar a escrita dos hispano falantes. A primeira das análises proposta, foi uma avaliação holística, que classificava textos de aprendizes e nativos em três níveis de proficiência. Os resultados apresentados nessa análise mostraram a necessidade de se criar novos atributos para caracterizar os textos Intermediários. No entanto, por meio dessa análise não foi possível separar textos de nativos e de aprendizes,

⁷<http://nilc.icmc.usp.br/habla/>

concluindo que ambos os grupos produzem textos com diferentes níveis de proficiência. Os resultados mostraram que no corpus Espanhol-Acadêmico-Br há textos muito bem escritos que foram produzidos por aprendizes e classificados como Avançados. Ainda assim, manualmente, foi comprovado que esses textos classificados como Avançados também tem alguns problemas, como erros de preposição, que não são detectados por meio da avaliação holística proposta. Isso sugere a necessidade de se realizar uma análise mais aprofundada. A análise anterior também é importante para caracterizar o grupo dos textos Intermediários. Após os resultados observados com o algoritmo de classificação, surge a suspeita de que a classe Intermediária provavelmente não existe no conjunto de textos e esses textos poderiam ser classificados como Avançados ou Iniciantes.

Contudo, os experimentos apresentados mostraram como a ferramenta *Visual Classification System* pode auxiliar no processo de anotação e construção do conjunto de treinamento e do modelo de classificação de textos. A ferramenta foi testada em cenários em que os conjuntos de treinamento tinham diferentes características, como poucas instâncias ou desbalanceamentos e mostrou um bom desempenho, sem a necessidade de criar dados ou distribuições irreais que não respeitam a distribuição dos dados. Os resultados desses experimentos foram publicados em Torres et al. (2013). Além disso, o processo de classificação de forma iterativa e interativa, disponível na ferramenta VCS, mostra-se promissor em aplicações de PLN, pois é possível, em etapas iniciais do processo, conhecer as deficiências e os pontos positivos do conjunto analisado.

Concluindo, a avaliação holística mostrou-se inadequada para separar os textos dos nativos e dos aprendizes. O conjunto de atributos propostos indicou que, da mesma forma que os aprendizes têm algumas dificuldades, os nativos também têm e por meio da análise proposta não é possível identificar o conjunto de dificuldades. Seguindo essa linha, foi decidido realizar uma análise baseada nos erros cometidos pelos nativos e aprendizes. Primeiramente, mostrou-se os principais erros que preocupam e cometem os hispano falantes (Seção 5.3.2) e depois foi apresentado o processo de anotação manual de textos de nativos e aprendizes (Seção 5.3.3), que mostrou os principais erros que eles comentem.

Como consequência do processo de anotação foi criada uma tipologia de erros. A proposta dessa tipologia se deu devido ao fato de que as tipologias que existem na literatura, apresentadas na Seção 2.3, são muito detalhadas,

sendo complexa sua reutilização em processos de anotação manual em larga escala. Tanto a tipologia de erros como os resumos anotados também estão disponíveis para outras pesquisas na página do projeto HABLA.

O processo de anotação de erros permitiu contrastar erros identificados na escrita acadêmica com erros levantados em outras pesquisas. Até onde se sabe, esse estudo foi o primeiro realizado para analisar a escrita acadêmica dos hispano falantes, aprendizes de português. O fato dos erros de aprendizes terem sido anotados por dois anotadores é um destaque desta pesquisa. Um dos poucos trabalhos que se conhece, em que todos os erros dos aprendizes foram anotados por mais de um humano foi apresentado por Dahlmeier et al. (2013). Na literatura, encontraram-se estudos que anotaram várias categorias de erros, porém por um único anotador. Os estudos em que vários anotadores estiveram envolvidos na anotação dos erros focaram em uma única categoria de erro.

Os erros dos hispano falantes, identificados no *córpus Espanhol-Acadêmico-Br*, coincidem com alguns dos erros apontados na Seção 2.3. No presente estudo foi confirmada tanto a ocorrência de diversas categorias de erros, quanto a natureza desses erros. Da mesma forma como aconteceu no levantamento realizado por Durão (1999), na presente análise, os erros relacionados com grafia incorreta foram os mais frequentes. Esses tipos de erros foram agrupados por Durão (1999) na categoria Erros fonológicos e gráficos e aparecem na Tabela 2.6. Esse tipo de erro também foram identificados por Grannier e Carvalho (2001), os quais classificaram esses erros dentro da categoria lexicais (Veja a Tabela 2.3).

Por outro lado, tanto Grannier e Carvalho (2001) quanto Durão (1999) identificaram erros produzidos pela interferência da língua materna. Grannier e Carvalho (2001) chamou esses erros de "uso de formas mistas"(Veja a Tabela 2.3), em que os hispano falantes combinaram regras do espanhol com regras do português, e Durão (1999) os nomeou de casos de "fronteira"(veja a Tabela 2.7). A geração desses erros foi observada nos textos anotados. Os exemplos apresentados nas Figuras 5.18 e 5.21 comprovam a ocorrência desses erros no *córpus Espanhol-Acadêmico-Br*. Além disso, as palavras e/ou expressões com esses erros podem ser itens lexicais de diversas categorias, como foi o caso de marcadores discursivos, expressões do português geral e do português acadêmico, verbos, dentre outras. Em alguns casos, esses erros somente afetam a forma da palavra, mas em outros casos afetam a semântica e, como consequência, afetam o sentido da frase.

A segunda categoria de erros mais frequentes, identificados por Grannier e Carvalho (2001) foram os erros morfossintáticos (Veja a Tabela 2.3). Nessa categoria, os autores incluíram erros relacionados com o uso das preposições e com regência verbal. Os erros de regência verbal, devido a sua complexidade não foram considerados na presente anotação; em contraste, os erros de preposições foram os segundos mais frequentes no presente estudo. Além disso, na análise de Durão (1999) os erros gramaticais foram os menos frequentes. Nesse grupo, a autora detectou que os hispano falantes têm dificuldades para contrair as preposições. Esse fenômeno foi observado muitas vezes nos textos do *cópus Espanhol-Acadêmico-Br*, sendo anotado como conectivo errado, devido à função que a contração tinha no trecho.

O fato dos erros gramaticais serem os menos frequentes no estudo de Durão (1999), não acontecendo o mesmo no presente estudo e na análise de Grannier e Carvalho (2001), pode estar relacionado com o fato do estudo de Grannier e Carvalho (2001) ter sido realizado com aprendizes de português como língua estrangeira, enquanto a análise de Durão (1999) aconteceu em textos produzidos por aprendizes de L2. Os aprendizes de língua estrangeira devem ter mais contato com a língua e, por isso eles cometem com maior frequência erros que são comuns aos nativos da língua.

Um fenômeno comum dos três estudos que aconteceu com alta frequência foram os erros produzidos pela interferência da língua materna, neste caso o espanhol. Seguindo essa linha, a proximidade entre o espanhol e o português torna o problema de interferência da língua materna ainda maior, pois a existência de muitos cognatos entre as duas línguas faz com que os aprendizes “arrisquem” traduções literais de itens lexicais do espanhol para o português. Muitas vezes a estratégia anterior gera palavras e/ou expressões corretas em português, porém outras vezes erros são produzidos. Esse tipo de erro acontece em todos os níveis da escrita, mas como o foco desta pesquisa são os erros no nível lexical, eles foram mais detalhados nesta análise. No Capítulo 6 são apresentados diversos métodos para compilar léxicos bilíngues criados para auxiliar os hispano falantes na escrita acadêmica em português.

Nas anotações apresentadas nesta seção comprovou-se que há marcas que distinguem a produção escrita dos hispano falantes. No entanto, hispano falantes cometem os mesmos erros identificados em textos de nativos do português. Essa descoberta faz com que a tarefa de separar os textos dos hispano falantes dos textos dos nativos não seja trivial, como foi comprovado na tarefa de classificação, em que se avaliou a qualidade da escrita. A classificação de

textos em três níveis de proficiência não foi bem sucedida, pois os atributos utilizados não foram suficientes para caracterizar os textos do nível Intermediário. Um resultado importante da avaliação holística foi o fato de detectar textos muito bem escritos pelos hispano falantes. Porém, marcas do espanhol foram observadas em textos com diferentes níveis de proficiência. A descoberta anterior indicou a necessidade de realizar uma análise mais detalhada que considera os erros detectados nos textos. Essa análise deve ser realizada como parte de um trabalho futuro.

A anotação dos textos dos aprendizes foi uma tarefa árdua que ajudou a identificar os principais problemas que enfrentam os aprendizes de português, assim como levantar os erros que diferem do conjunto de erros que os nativos da língua comentem. No entanto, a anotação dos erros não foi suficiente para ser utilizada como referência de um sistema computacional, dado o número de erros ser pequeno para tal tarefa. Contudo, a detecção manual de erros lexicais confirmou a necessidade de criar auxílios que automaticamente tratem esses problemas. Esses auxílios são detalhados no Capítulo 6.

Projeto HABLA: Resultados e Avaliação dos Métodos de Auxílio

Neste capítulo, são apresentados os métodos de auxílios propostos para aprimorar a escrita acadêmica em português de hispano falantes. Esses auxílios são sustentados por léxicos bilíngues do espanhol e o português, criados para tratar erros e/ou inadequações que ocorrem no processo de transferência de itens lexicais do espanhol para o português. Análises realizadas no corpus de aprendizes Espanhol-Acadêmico-Br, apresentado no Capítulo 5, comprovaram a importância de se investir na criação desses léxicos bilíngues. Finalmente, neste capítulo apresenta-se uma arquitetura que inclui os léxicos bilíngues em uma ferramenta computacional para auxiliar os hispano falantes aprendizes de Português, chamada HABLA.

Este capítulo trata de um dos objetivos deste doutorado: a aplicação dos métodos e recursos léxico-sintático-semânticos baseados em corpus paralelos alinhados, na área de Sistemas de auxílio à Escrita. O uso destes se destina a auxiliar na pós-edição de um texto científico escrito em português por um aluno pesquisador cuja língua materna é o espanhol.

6.1 Métodos de Auxílios à Escrita

Como observado na Figura 6.1, nesta pesquisa empregou-se métodos de tradução automática, baseados em corpus paralelos para disponibilizar os

três auxílios à escrita:

1. Equivalências entre o espanhol e português;
2. Escolhas lexicais em português e
3. Detecção e correção de erros.

Para materializar esses auxílios, combinaram-se a tradução automática com diversas heurísticas para compilar os quatro léxicos bilíngues:

- listas de cognatos e falsos cognatos das línguas português e espanhol;
- listas de pares de marcadores discursivos;
- listas das expressões formulaicas que aparecem nos textos científicos;
- listas de verbos relacionados com pesquisa científica em português, recorrentemente utilizados em textos acadêmicos.

Os léxicos bilíngues foram gerados em função do tipo de auxílio que oferecem. Por exemplo, o léxico de cognatos e falsos cognatos tem a função de indicar que um determinado item é um falso cognato, como consequência, o recurso alerta os aprendizes sobre possíveis enganos na procura de equivalências entre as duas línguas. Desta forma, ele oferece o primeiro auxílio da Figura 6.1. Por outro lado, o léxico de marcadores discursivos é um recurso que oferece os três auxílios apresentados na Figura 6.1, ele mostra as equivalências entre o espanhol e o português; apresenta sinônimos de marcadores em português e identifica alguns erros produzidos pela interferência do espanhol nos marcadores discursivos em português.

A hipótese que sustenta a criação dos recursos lexicais é a de que as vezes um hispano falante, aprendiz do português, utiliza o conhecimento que tem da sua língua nativa, o espanhol, para produzir os textos em português. A partir disso, na transferência dos léxicos do espanhol para o português ocorrem erros ou inadequações. O auxílio, (1) Equivalências entre o espanhol e português, parte do princípio que o hispano falante conhece o termo em espanhol, mas não conhece o termo adequado em português. Assim apresentam-se as melhores equivalências e exemplos em que são utilizadas. Embora técnicas de tradução automática tenham sido utilizadas para obter as equivalências, os algoritmos se combinaram com outras heurísticas e as saídas dos métodos de tradução foram preparadas para auxiliar a produção escrita. O segundo auxílio, (2) Escolhas lexicais em português, é útil para aprendizes que conhecem

Figura 6.1: Auxílios oferecidos pelos léxicos bilíngues

o termo em português, mas procuram por termos mais adequados, que podem ser obtidos a partir da apresentação de sinônimos. Finalmente, o terceiro auxílio, (3) Detecção e correção de erros, identifica erros lexicais nos textos e tenta corrigi-los.

Os léxicos bilíngues contêm itens lexicais de diferentes tipos. Alguns léxicos contêm palavras e/ou multipalavras que são de categorias lexicais diferentes. Como consequência, a compilação de cada léxico é apresentada em seções diferentes deste capítulo para mostrar cada uma de suas especificidades. No entanto, métodos gerais foram aplicados para construir os auxílios apresentados na Figura 6.1. O projeto e avaliação desses métodos gerais são explicados a seguir e suas aplicações são evidenciadas na compilação de cada léxico bilíngue.

6.1.1 Equivalências entre o Espanhol e o Português

Os léxicos bilíngues desta pesquisa incluem itens lexicais com características diferentes. Por exemplo, o léxico de cognatos e falsos cognatos considera palavras com ortografia similares; os léxicos de marcadores discursivos e expressões formulaicas contêm palavras e multipalavras que são difíceis de extrair automaticamente; por último os verbos são facilmente reconheci-

dos com uma anotação morfossintática, mas indicar sua adequação à escrita acadêmica é um processo mais subjetivo. Os exemplos anteriores ajudam a refletir sobre a complexidade de construir esses léxicos. Como mostrado no Capítulo 3 a escolha da técnica utilizada na compilação depende dos itens incluídos nos léxicos e da aplicação para a qual o léxico bilíngue está sendo compilado. Considerando a diversidade nas características dos léxicos, foram testados três métodos para estabelecer as equivalências entre o espanhol e o português:

1. o baseado em similaridade ortográfica;
2. o baseado em tradução automática estatística e
3. o apresentado em Bai et al. (2009).

Por meio desses três métodos foram compilados os léxicos bilíngues propostos nesta pesquisa.

Os métodos propostos para obter as equivalências recebem como entrada um ou vários itens em espanhol e como saída apresentam a melhor equivalência em português. A metodologia adotada foi observada em alguns dos trabalhos apresentados na Seção 3.2, em que a etapa de extração do léxico na língua fonte foi separada da etapa de obtenção da equivalência na língua alvo. Como será mostrado ao longo deste capítulo, a metodologia utilizada requer a validação do recurso nas diferentes etapas, o que representou um dos maiores desafios desta pesquisa. Para avaliar os léxicos utilizou-se uma avaliação intrínseca em que a saída dos algoritmos foi comparada com um conjunto de traduções criadas por humanos. Além disso, alguns dos recursos criados também foram avaliados extrinsecamente na tarefa de corrigir erros lexicais. A seguir são apresentados os métodos para obter as equivalências entre os léxicos em espanhol e em português.

6.1.1.1 Método baseado em similaridade

O primeiro método utilizado para obter as equivalências entre palavras do espanhol e do português foi o método baseado em similaridade ortográfica. Esse método empregou a medida de similaridade *Longest Common Subsequence Ratio* (LCSR) para identificar palavras equivalentes. A medida LCSR foi empregada porque é uma medida muito utilizada para avaliar similaridade entre palavras (Kondrak e Dorr, 2004; Frunza e Inkpen, 2009; Koehn e Knight,

2002) e, além disso, em Torres e Aluísio (2011) ela foi a melhor ranqueada dentre várias medidas testadas na identificação de cognatos entre o português e o espanhol.

A similaridade entre as línguas viabiliza o uso do método. O algoritmo para estabelecer as equivalências por similaridade precisa de uma lista de palavras para procurar as similaridades a partir de uma entrada específica. Desta forma, o método é um algoritmo de busca que precisa de um critério de parada para selecionar a palavra mais similar. Comumente, o critério de parada é o número total de caracteres diferentes entre duas sequências. Quando identificadas as expressões com tamanho maior que o critério de entrada o algoritmo para e retorna a sequência com menor número de caracteres diferentes. Caso nenhuma sequência similar tenha sido encontrada, a saída do método é vazia.

O método baseado em similaridade é útil para se obter equivalências entre palavras das duas línguas que são ortograficamente similares. Por exemplo, para a palavra em espanhol "*necesidad*" o algoritmo será bem sucedido mostrando sua equivalente: "necessidade". Já, o algoritmo é incapaz de obter o equivalente da palavra "*vecindad*", pois a equivalência correta é "vizinhança" que não tem similaridade ortográfica.

O método de similaridade foi um dos métodos utilizados para estabelecer equivalências entre substantivos e adjetivos de espanhol e o português, nos experimentos que são apresentados na Seção 6.1.3.1. Além disso, o método foi testado na compilação do léxico bilíngue de cognatos e falsos cognatos entre as duas línguas, como será apresentado na Seção 6.2.1.

6.1.1.2 Método baseado em tradução automática

O segundo método testado procura as traduções em português das palavras ou expressões em espanhol. Para obter as traduções mais prováveis foi treinado um modelo de tradução por meio do *toolkit* Moses (Koehn et al., 2007). O modelo de tradução foi treinado, seguindo o *pipeline* oferecido pelo Moses e utilizando os pares de sentenças do *corpus* paralelo Fapesp (Seção 5.1.1). O *pipeline* do Moses foi executado sequencialmente, seguindo a configuração estabelecida como *baseline*¹ na competição de sistemas de tradução, apresentada na *EMNLP 2011 Sixth Workshop on Statistical Machine Translation*.

O *corpus* paralelo, alinhado sentencialmente, foi convertido em dois arquivos: o primeiro arquivo com as sentenças em espanhol e o segundo arquivo

¹<http://www.statmt.org/wmt11/baseline.html>

com as sentenças em português. Todas as sentenças de ambas as línguas foram tokenizadas e cada arquivo trazia uma sentença por linha, com as etiquetas de começo e fim da sentença. As sentenças paralelas foram alinhadas lexicalmente por meio do Giza++ (Och e Ney, 2000a). O Giza++ procura o melhor alinhamento lexical entre duas sentenças paralelas, empregando os modelos IBM (Brown et al., 1993) e HMM (Vogel et al., 1996). O Giza++ foi treinado com todo o corpus paralelo, considerando a configuração padrão na versão 2.0. Foram realizadas as iterações dos modelos IBM-1, IBM-3, IBM-4 e HMM para obter os melhores alinhamentos dos pares de sentenças. Os alinhamentos lexicais foram realizados nas duas direções espanhol-português e português-espanhol e depois os alinhamentos em ambas as direções foram combinados, utilizando o algoritmo apresentado por Och e Ney (2003) para obter a tabela de tradução lexical.

Os arquivos com as sentenças em espanhol e em português do corpus paralelo Fapesp foram utilizados para obter modelos de línguas de 5-gramas para as línguas. A ferramenta *SRLIM*, comentada na Seção 3.1.3 foi empregada com essa finalidade. O método *Kneser-Ney smoothing* foi utilizado para estimar a contagem dos n-gramas. O modelo de língua portuguesa treinado de tamanho 5-gramas é composto por:

- 100.459 1-gramas,
- 980.972 2-gramas,
- 2.458.822 3-gramas,
- 3.295.080 4-gramas e
- 3.504.354 5-gramas.

Já o modelo de língua espanhola treinado é composto por:

- 106.441 1-gramas,
- 949.965 2-gramas,
- 2.367.569 3-gramas,
- 3.346.543 4-gramas e
- 3.696.864 5-gramas.

No decorrer desta seção, o modelo de tradução obtido com o *toolkit* Moses será chamado de método estatístico.

Após a configuração do modelo estatístico de tradução utilizando os parâmetros especificados, este foi testado na tarefa de traduzir palavras e expressões para se obter os léxicos bilíngues propostos.

6.1.1.3 Método empírico

O método de tradução baseado em frases, disponibilizado pelo *toolkit* Moses, se mostrou como uma alternativa eficiente para estabelecer as equivalências entre o espanhol e o português. Contudo, na Seção 3.2 foram estudados métodos que, baseados em alinhamentos lexicais, conseguiam aprimorar a saída do *toolkit* Moses. Um dos métodos propostos chamou a atenção da pesquisadora pelo seu desempenho na tradução de multipalavras, um dos problemas que precisam ser tratados nesta pesquisa. Seguindo essa linha, o método empírico apresentado por Bai et al. (2009) foi testado na tarefa de se obter a tradução de expressões fixas e semifixas contidas nos léxicos propostos.

O método proposto por Bai et al. (2009) será chamado nesta pesquisa de método empírico. Ele foi detalhado na Seção 3.2 e a seguir serão comentadas algumas modificações adotadas para adaptar o método às necessidades da presente pesquisa. O método empírico também usa o alinhamento lexical do corpus paralelo Fapesp. O método recebe como entrada o corpus paralelo alinhado lexicalmente. O alinhamento é o mesmo utilizado no modelo treinado com a *toolkit* Moses, e desta forma a probabilidade de traduzir palavras independentes é a mesma. O algoritmo empírico recebe como entrada uma palavra ou multipalavra em espanhol e procura candidatos à tradução em português. Para isso, o método analisa o contexto das palavras candidatas à tradução na língua alvo e forma sequências candidatas à tradução. Após a extração das candidatas, o algoritmo foca na ordenação das traduções candidatas. O método empírico obtém as sequências de tradução aplicando diferentes heurísticas.

Nestes experimentos, todos os passos apresentados no Algoritmo 1 da Seção 3.2 foram implementados em *Python*. Para estimar as heurísticas dos passos três, quatro e cinco, foi preciso estimar a probabilidade de tradução de cada palavra dentro das expressões. Esse cálculo foi obtido considerando o alinhamento lexical do corpus paralelo Fapesp, obtido com a ferramenta Giza++. Após a implementação do algoritmo original de Bai et al. (2009), foram observadas algumas deficiências na hora de gerar as traduções candidatas (passo

seis e sete).

No algoritmo original (Algoritmo 1 Seção 3.2), as traduções candidatas (passo seis) se formavam combinando todas as palavras extraídas das sentenças. Após gerar as sequências candidatas, no passo sete do algoritmo eram extraídas subsequências, a partir das sequências candidatas. Essas subsequências também pertenciam à lista de traduções candidatas. Esse último passo era realizado com a finalidade de aumentar a cobertura do método, aumentando o número de traduções candidatas. Após implementar o algoritmo original de Bai et al. (2009) foi comprovado que no passo seis aceitava-se qualquer combinação de palavras candidatas, gerando muitas sequências que não existiam na língua portuguesa. Por exemplo, para obter a tradução mais provável do marcador discursivo em espanhol “*por supuesto*” foram obtidas as seguintes palavras candidatas: “claro”; “é”; “evidentemente”; “por”; “certamente”; “que”; “ou”; “seu”; “naturalmente” e “e”. A heurística do passo seis combinou todas as palavras candidatas, gerando diversas sequências, algumas como: “é claro”; “é claro que”; “e claro que”; “é claro por seu que e que”; “evidentemente que ou”. Observando as sequências geradas no passo seis pode-se concluir que muitas delas são sequências ruidosas e improváveis na língua portuguesa. Além disso, o número de sequências ruidosas, geradas no passo seis, pode ser ainda maior se diversas *stopwords* pertencem à lista de palavras candidatas.

No passo sete do algoritmo um fenômeno similar ao do passo seis aconteceu. Bai et al. (2009) propuseram gerar subsequências candidatas, a partir das sequências do passo seis. Embora os autores comprovassem que o passo sete do algoritmo é importante, pois aumenta a cobertura do método, nele a quantidade de sequências geradas é tão elevado que para determinadas expressões, deteriora o desempenho computacional e na maioria dos casos as expressões geradas não existem na língua portuguesa.

A partir dos problemas identificados no passo seis e sete do algoritmo, nesta pesquisa, as expressões candidatas foram geradas considerando as palavras candidatas, obtidas no passo cinco do algoritmo e o contexto em que elas aconteceram na sentença. Isto é, as traduções candidatas foram n-gramas de tamanho $n = 1, 2, 3, 4, 5$, em que sempre aparecia uma palavra candidata extraída no passo cinco. Esse critério foi válido, porque as entradas dos léxicos bilíngues eram expressões com palavras consecutivas. Os passos subsequentes ao passo sete foram mantidos iguais ao algoritmo original. O método empírico foi testado durante a geração de alguns dos léxicos propo-

tos nesta pesquisa.

6.1.1.4 Avaliação dos métodos de equivalências

Para avaliar intrinsecamente os três métodos propostos para gerar as equivalências entre o espanhol e o português foram criados manualmente dois *gold standards*: um composto por palavras em espanhol, com suas equivalências em português e o outro composto por multipalavras equivalentes nas duas línguas. A lista bilíngue de palavras equivalentes foi construída durante a compilação de um léxico de substantivos e adjetivos (Seção 6.1.3.1) e a lista bilíngue de expressões foi compilada durante a criação do léxico de expressões formulaicas (Seção 6.2.4). Os três algoritmos foram avaliados na procura de equivalências entre palavras, enquanto o método estatístico e o empírico foram avaliados na obtenção de traduções de multipalavras.

A lista com 970 substantivos e adjetivos em espanhol foi traduzida manualmente por uma linguista² falante nativa do português. Os métodos foram avaliados considerando as métricas de cobertura e precisão, apresentadas na Seção 3.1.3. Na Tabela 6.1, mostram-se os resultados do processo de avaliação e, além disso, mostra-se o desempenho dos três métodos em dois cenários: (i) considerando todas as palavras e (ii) considerando as palavras que aconteceram no corpus paralelo com uma frequência maior que quatro.

Tabela 6.1: Avaliação dos métodos de equivalências entre palavras

	Frequência > 0		Frequência > 4	
	precisão	cobertura	precisão	cobertura
Similaridade	0,54	0,51	0,54	0,51
Empírico	0,68	0,68	0,85	0,85
Estatístico	0,76	0,76	0,86	0,86

Na Tabela 6.1 é possível observar como os métodos empírico e estatístico são os que têm melhor precisão e cobertura. É importante apontar que ambos os métodos têm precisão e cobertura iguais, porque sempre que a palavra existe no corpus vão conseguir oferecer a tradução. Por outro lado, a precisão e cobertura dos dois métodos são ainda melhores se unicamente se consideram palavras com uma frequência maior que quatro no corpus (frequência > 4), alcançando uma precisão de 0,85 para o método empírico e 0,86 para o estatístico. Desta forma, pode-se concluir que ambos os métodos são fortemente influenciados pela frequência das palavras no corpus. Como consequência,

²Magali Sanches Duran

quanto menor for a frequência da palavra no corpus, maior será a probabilidade dos algoritmos não conseguirem identificar a tradução correta.

Foram combinados os testes estatísticos teste-F e teste T-student para avaliar o método de melhor desempenho. Os testes estatísticos foram realizados com um nível de confiança de 0,05. Quando comparados, os métodos estatístico e empírico com o método baseado em similaridade, a diferença entre eles foi significativa, considerando todas as palavras do conjunto e as palavras com frequência maior que quatro. Desta forma, como mostrado na Tabela 6.1 os dois primeiros métodos têm melhor precisão e cobertura que o método baseado em similaridade. Além disso, também foram realizados os testes estatísticos para comparar o método empírico e o estatístico. Os testes mostraram que há diferenças significativas no desempenho dos algoritmos, quando considerado todo o conjunto de palavras, sendo que os algoritmos têm precisão e cobertura similares, quando procurada a tradução de palavras com frequências maior que quatro. Embora, não haja diferença estatística da precisão e cobertura dos métodos empírico e estatístico, o resultado anterior indica que o método estatístico é menos susceptível à ocorrência de palavras de baixa frequência no corpus. Considerando esse quesito, o desempenho do método estatístico foi melhor.

Embora a precisão e a cobertura do método baseado em similaridade sejam diferentes dos valores obtidos com os outros dois métodos, 7% das vezes o algoritmo de similaridade conseguiu oferecer a equivalência correta, enquanto o método empírico falhou e 5% das vezes o método de similaridade superou ao método estatístico. O fato anterior acontece porque o método baseado em similaridade não é influenciado pela frequência das palavras no corpus. Seguindo essa linha, foi avaliado o desempenho do método empírico e estatístico combinado com o método baseado em similaridade. Na Tabela 6.2, mostram-se os resultados desse experimento.

Tabela 6.2: Avaliação da combinação dos métodos de equivalências entre palavras

	Frequência>0	Frequência>4
	precisão	precisão
Empírico_Similaridade	0,82	0,90
Estatístico_Similaridade	0,85	0,92

Os resultados apresentados na Tabela 6.2 mostram que houve uma melhoria no desempenho dos algoritmos empírico e estatístico quando combinados com o método baseado em similaridade. A precisão do método empírico me-

lhorou 14% com todas as palavra do experimento e 5% considerando somente as palavras com frequência maior que 4. Já, o método estatístico melhorou 9% com todas as palavras do conjunto e 6% para as palavras com frequência maior que quatro. O teste estatístico realizado para comparar se eram significativas as diferenças entre os algoritmos empírico e empírico combinado com similaridade e estatísticos e estatístico combinado com similaridade mostraram em ambos os casos que a melhoria na precisão dos algoritmos foi significativa. Os resultados anteriores mostram que o método baseado em similaridade pode auxiliar os métodos estatístico e empírico, principalmente quando palavras de baixa frequência são traduzidas.

Em um segundo experimento, os métodos estatísticos e empíricos foram avaliados na tarefa de traduzir para o português expressões formulaicas da escrita acadêmica em espanhol. Para isso, foi selecionada uma lista com 264 expressões com três e quatro palavras. As expressões formulaicas foram extraídas e traduzidas para o português, manualmente. Cada expressão recebeu uma ou várias traduções e em alguns casos símbolos de pontuação foram incluídos na expressão extraída, como consequência, eles apareceram, também, nas traduções. Por exemplo, a expressão em espanhol "*es decir*," foi extraída considerando o símbolo de pontuação e sua tradução candidata foi "é dizer,".

Na Tabela 6.3 aparecem os valores de precisão e cobertura obtidos para os métodos empírico e estatístico na tradução para o português de expressões formulaicas em espanhol. Segundo a avaliação apresentada na Tabela 6.3 , o método estatístico teve melhor precisão e cobertura que o método empírico na tradução de expressões com três ou quatro palavras. As diferenças de precisão e cobertura de ambos os métodos são estatisticamente significativas segundo a combinação dos testes estatísticos teste-F e teste T-student. Analisando os resultados, foi observado que o método empírico é muito influenciado pela ocorrência da expressão no corpus. Caso a expressão não apareça no corpus, o método não consegue oferecer a tradução, por isso há diferenças entre os valores de precisão e cobertura do método. Por outro lado, o método estatístico sempre retorna uma tradução para uma dada expressão, desta forma a precisão e a cobertura são iguais. O método estatístico também é influenciado pela ocorrência da palavra no corpus, porém quando a expressão não acontece no corpus o método retorna uma tradução literal da expressão. A alternativa anterior dá certo em muitos casos, produto da similaridade entre as duas línguas.

Ambos os métodos mantiveram valores equivalentes de precisão quando

Tabela 6.3: Avaliação dos métodos de equivalências entre expressões formulaicas

	Empírico		Estatístico	
	3-gram	4-gram	3-gram	4-gram
precisão	0.83	0.84	0.92	0.91
cobertura	0.77	0.72	0.92	0.91

traduzidas expressões de diferentes tamanhos. No entanto, a cobertura do método empírico foi diferente na obtenção de expressões com tamanhos diferentes. De qualquer forma, os testes estatístico teste T-student e teste-F mostraram que as diferenças não eram significativas.

Não foi possível comparar o desempenho do método empírico e estatístico na tradução de palavra e multipalavras, pois os cenários dos experimentos eram diferentes. No experimento para traduzir palavras foi considerado um maior número de palavras e, além disso, uma única tradução foi oferecida, manualmente, para cada palavra dada como entrada. Na anotação manual das expressões formulaicas para algumas expressões, mais de uma tradução foi oferecida manualmente, esse pode ser um dos motivos que influenciou em um melhor desempenho do método. No entanto, outro motivo que influenciou nos desempenhos diferentes na tradução de palavras e expressões está relacionado ao fato de que as palavras foram extraídas de dicionários e muitas não apareciam no *córpus* paralelo, acontecendo o contrário com as expressões: a maioria estava incluída no *córpus*. Esse fato mostra como a qualidade do *córpus* paralelo é importante para obter algoritmos com bons desempenhos.

Os métodos apresentados para obter as equivalências entre as palavras e/ou expressões em espanhol e português foram utilizados na compilação dos léxicos bilíngues propostos nesta pesquisa. Nas próximas seções deste capítulo, é mostrado como cada método foi utilizado. Os comparativos entre os métodos de tradução estatístico e empírico mostraram que o método estatístico, que utilizou o *toolkit* Moses, é menos suscetível a ruídos que podem aparecer durante a criação dos recursos, por exemplo, expressões ou palavras que não acontecem nos *córpus* de referência. Seguindo essa linha, ele foi o método selecionado na construção dos léxicos bilíngues, no caso de uso de método de tradução.

6.1.2 Escolhas Lexicais em Português

O segundo auxílio implementado nesta pesquisa concentra-se na apresentação de sinônimos de palavras e expressões em português. Como mostrado no Capítulo 5, muitas vezes o hispano falante utiliza uma palavra ou expressão que existe em português, mas que não é frequentemente utilizada ou que não é empregada em um dado contexto. Esse tipo de inadequações foram etiquetadas durante a anotação, apresentada na Seção 5.3.3 como erros de Convencionalidade. No Exemplo 1, mostra-se como o hispano falante utiliza a expressão "*prover serviços*" que não é comum em português, sendo mais frequente escrever "oferecer serviços".

Exemplo 1: "A zona se encontra em risco devido a projetos turísticos de grande envergadura que podem ocasionar graves deterioros dos ecossistemas e de suas capacidades de *prover serviços*/(oferecer serviços)."

No Exemplo 1, há vários erros, porém o foco é mostrar como o hispano falante selecionou o verbo "prover", que existe nas duas línguas, mas a escolha não foi adequada. A detecção e correção da inadequação em foco requer uma análise mais elaborada, a qual pode seguir os métodos propostos para detectar e corrigir erros de colocações, em que a frequência da expressão é utilizada como parâmetro para detectar o erro e medidas de associação indicam a melhor troca. Contudo, utilizou-se o exemplo para mostrar como um método, que procure sinônimos, pode ser utilizado, como alternativa para solucionar o problema.

O uso reiterativo de itens lexicais foi outro problema observado no Capítulo 5. Parece que o hispano falante ao aprender uma nova palavra e/ou expressão para determinado assunto sempre a usa, gerando um texto com pouca diversidade lexical. Para tratar esse problema, a apresentação de sinônimos também aparece como uma alternativa de solução válida.

Para o português do Brasil já existem recursos que apresentam os sinônimos de itens lexicais. Um dos recursos mais conhecidos é a *WordNet.Br* (Di Felippo e Dias da Silva, 2007). A *WordNet.Br* é um dicionário eletrônico de sinônimos e antônimos enriquecido com a especificação de outras relações de natureza lógico-conceitual. Além da *WordNet.Br*, outro recurso que também apresenta a relação de sinônimos e antônimos é o *TeP 2.0* (Maziero et al., 2008). O *TeP 2.0* é um *Thesaurus* Eletrônico para o português do Brasil e foi construído combinando informações da *WordNet.Br* e a base de dados lexical *Diadorim* (Gregghi et al., 2002) e apresenta sinônimos de itens lexicais das categorias substantivo, adjetivo, verbo e advérbio. Outro recurso que apresenta

sinônimos e antônimos de verbos é a *OpenWordNet-PT* (De Paiva et al., 2012), um recurso lexical recentemente construído que relaciona semanticamente itens lexicais de diversas categorias. Os recursos anteriores (*OnpenWordNet-Pt*, *WordNet.Br* e *TeP 2.0*) poderiam ter sido utilizados para obter os sinônimos dos itens lexicais do projeto HABLA, porém eles foram construídos para uso geral e não para apoiar auxílios à escrita científica. Como consequência, para alguns casos os recursos apresentam muitos sinônimos da língua geral que não deveriam ser utilizados na escrita acadêmica. Por exemplo, para o verbo “*elucidar*” a *WordNet.Br* recupera os seguintes sinônimos: “*aclarar*”, “*alumiar*”, “*desanuviar*”, “*esclarecer*”, “*iluminar*”, “*ilustrar*”, “*instruir*”, “*abrir*”, “*aclarar*”, “*dilucidar*”, “*desatar*”, “*desvelar*”. A apresentação dessa quantidade de sinônimos, longe de ajudar, atrapalha o auxílio à escrita, considerando que o próprio usuário precisa desambiguar e escolher o sentido adequado do verbo.

É importante lembrar que os recursos do HABLA foram produzidos para auxiliar a escrita acadêmica, por isso, deseja-se concisão na apresentação das sugestões. Outro problema observado com os recursos disponíveis é que eles não fornecem sinônimos para marcadores discursivos, que é um dos itens lexicais considerado nos auxílios do projeto HABLA.

Para evitar as deficiências anteriores e tentado obter sinônimos para os marcadores discursivos em português, foi proposto um método que extraia sinônimos do corpus paralelo Fapesp. Considerando os itens que aparecem no corpus, esperava-se que a lista de sinônimos apresentada pelo método não fosse muito extensa. Além disso, devido ao fato do corpus paralelo descrever pesquisas acadêmicas, esperava-se que os sinônimos extraídos fossem aceitos na produção escrita desse gênero. Finalmente, como as informações do método são extraídas de um corpus paralelo, é possível manter as equivalências entre o espanhol e um conjunto de equivalentes em português que são sinônimos.

O método para extrair sinônimos em português utiliza a língua espanhola como pivô. Usando as equivalências entre o espanhol e o português, o método obtém os sinônimos de itens lexicais em português. Seguindo essa linha, o método permite criar recursos que mantêm a ligação entre os itens em espanhol e diversas traduções em português. Para isso, é necessário criar listas bilíngues de itens de uma mesma categoria gramatical. Nesta pesquisa as listas bilíngues foram criadas por meio do método estatístico, apresentado na Seção 6.1.1.2.

Considerando uma lista de itens lexicais em espanhol e o conjunto de seus

respectivos equivalentes de tradução em português, apresenta-se um método que agrupa esses itens em português. O agrupamento assume que, se vários itens em português se alternam como equivalentes de um item lexical em espanhol, provavelmente eles têm a mesma função discursiva e são sinônimos. Da mesma forma, se um item em português ocorre como equivalente de vários itens em espanhol, os itens lexicais em espanhol provavelmente têm a mesma função discursiva e são sinônimos.

Na Figura 6.2, o item lexical em espanhol w_{e0} é traduzido para o português como os itens $w_{p0}; w_{p1}ew_{p2}$, os quais constituem o conjunto de equivalentes de w_{e0} . Por outro lado, o item w_{en} é traduzido para os itens $w_{p0}; w_{p2}ew_{p3}$, os quais constituem o conjunto de equivalentes de w_{en} . A hipótese para se obter os sinônimos em português é a de que itens lexicais como w_{p0} ; e w_{p2} que aparecem nos dois conjuntos de equivalentes, provavelmente são itens com a mesma função discursiva em português. Da mesma forma, como os itens em espanhol w_{e0} ; e w_{en} têm equivalentes em comum, também teriam a mesma função. Caso não exista alternância dos itens em português nas diferentes traduções, os itens que compõem o conjunto de equivalentes também são sinônimos, por exemplo: $w_{p0}; w_{p1}ew_{p2}$. Um método similar já foi utilizado por Wu e Zhou (2003b) para obter sinônimos de colocações entre as línguas chinês e inglês.

Figura 6.2: Estrutura da saída do procedimento para obter sinônimos entre itens lexicais.

O método para obter sinônimos com base nas equivalências de tradução foi proposto para categorizar os marcadores discursivos segundo sua função

discursiva. Para isso, o método foi combinado com medidas de associação e é apresentado na Seção 6.2.3. O método também foi utilizado para apresentar sinônimos de verbos em português, como será exemplificado na Seção 6.2.2.

6.1.3 Detecção e Correção de Erros

O último auxílio considerado nesta pesquisa de doutorado foi a detecção e correção de erros. A maioria dos estudos dedicados a detectar automaticamente erros é guiada por um corpus de aprendizes previamente anotado com as diferentes categorias de erros. A anotação de erros em um corpus de aprendizes, contudo, é uma tarefa demorada que requer anotadores especialistas. Além disso, para ser um recurso útil na identificação automática de erros, é necessário que o corpus anotado contenha um grande número de erros, o que requer a anotação de um grande volume de dados. A anotação apresentada na Seção 5.3.3 mostrou que o corpus de erros Espanhol-Acadêmico-Br não era uma recurso capaz de auxiliar um método que detecte e corrija erros automaticamente, via métodos de aprendizado de máquina. No entanto, a análise dos erros, apresentada na Seção 5.3.3, indicou a natureza de alguns dos erros. A partir da natureza do erro, surge nesta pesquisa a alternativa de gerar erros artificiais para detectar e corrigir erros.

O propósito de uso da abordagem de geração de erros é diferente dos objetivos seguidos nos estudos de Foster e Oistein (2009) e Dicknson (2010). Enquanto Foster e Oistein (2009) e Dicknson (2010) produziram erros artificiais para criar exemplos negativos que aprimorassem algoritmos de aprendizado de máquina, neste projeto de doutorado a geração de erros artificiais foi uma alternativa para produzir um léxico de erros que apoiasse o algoritmo para detectar e corrigir erros. Outra diferença está relacionada com o grupo de aprendizes considerados para gerar os erros. Na nossa abordagem, os erros reais específicos de nativos do espanhol, quando aprendem português, foram gerados, enquanto Foster e Oistein (2009) usam um corpus que contém erros produzidos por nativos de diferentes nacionalidades quando estudam o inglês e Dicknson (2010) não especifica o corpus de aprendiz empregado como inspiração.

No corpus Espanhol-Acadêmico-Br foi comprovado que, às vezes, o erro não era um erro ortográfico e sim um erro produzido após a tentativa de “adivinhar” uma palavra ou expressão equivalente em português. Esse tipo de problema também foi detectado no estudo de Duran (2008b) e a autora acredita que esses erros podem ser detectados pelos sistemas que identificam erros

de nativos, entretanto estes sistemas não conseguiriam corrigir o problema. Para tratar esses problemas, o sistema teria que reconhecer a interlíngua que o aprendiz produz. Seguindo esta linha, para gerar o léxico de erros artificiais usou-se o mesmo raciocínio observado no corpus Espanhol-Acadêmico-Br. Um exemplo do raciocínio lógico dos hispano falantes é apresentado no Exemplo 2.

Exemplo 2: O problema principal dos métodos de coocorrência é que *se bem* (**embora**) podem detectar que existe uma interação, não podem detectar o tipo de interação entre duas proteínas.

No Exemplo 2, o hispano falante produz a expressão em português “*se bem*”, a partir da tradução literal do marcador discursivo em espanhol “*si bien*”. Considerando unicamente a frequência da expressão “*se bem*”, sublinhada e em cursiva no exemplo, em um corpus produzido por nativos de português, poderíamos alertar o aprendiz que a troque por outra, pelo fato da frequência ser muito baixa. Porém, o principal problema do exemplo não é a detecção do erro e sim oferecer a melhor sugestão para corrigir esse erro. Observando a anotação realizada no exemplo, pode-se inferir que o aprendiz quis usar a equivalência em português do marcador discursivo em espanhol “*si bien*” e errou porque, nesse caso, a tradução correta seria “embora” e não a tradução literal “*se bem*”. Desta forma, o método de geração de erros tenta prever diferentes categorias de erros, sendo seu principal diferencial **manter o controle da forma correta da palavra ou expressão**. Assim, é possível oferecer a correção do erro ou pelo menos sugestões bem sucedidas.

O léxico de erros, chamado de *Spanish-native-like errors* é composto de três formas:

1. a forma em espanhol,
2. a interlíngua (forma errada) e
3. a forma em português.

A forma da interlíngua é usada para tentar identificar um possível erro cometido por um nativo de espanhol, a forma em espanhol (FS) é usada para produzir sugestões:

- “Você quis dizer FS em Português?”
- “A equivalência pode ser FP” (a forma em português).

Segundo o exemplo anterior, a sugestão seria:

- “Você quis dizer *si bien*?”
- “A equivalência pode ser embora”.

O ponto inicial da abordagem de geração de erros artificiais foi o corpus de aprendiz Espanhol-Acadêmico-Br (Seção 5.2) e a tipologia de erros criada após analisar este corpus (Seção 5.3.2). Nesta pesquisa foram gerados automaticamente os erros de marcadores discursivos, de expressões formulaicas e de transferência de sufixos. As heurísticas utilizadas para gerar erros artificiais diferem de um recurso lexical para outro. Por isso, a geração dos léxicos de erros será detalhada nas seções em cada um dos léxicos bilíngues for apresentado. No entanto, o método de geração de erros de sufixos é um método que combina heurísticas simples por meio das quais foi possível gerar erros artificiais. O sucesso desse método deu impulso (e garantias) para a geração de outras categorias de erros; por isso, ele será explicado a seguir.

6.1.3.1 Geração de Erros de Transferência de Sufixos

Um dos erros identificados no corpus Espanhol-Acadêmico-Br, causado pela interferência do espanhol, foi provocado pela substituição de sufixos em espanhol por sufixos em português. Esse erro é produzido porque o hispano falante conhece a transferência dos sufixos entre o espanhol e o português e para produzir a palavra em português, mantém o lexema da palavra em espanhol e coloca o sufixo em português. Por exemplo, a palavra *vecindad* em espanhol foi transferida para o português como "*vecindade*", mas na verdade ela deveria ter sido transferida como "*vizinhança*". No exemplo a seguir, observa-se como o hispano falante utiliza a palavra "*validez*" do espanhol e não faz a troca do sufixo para que a palavra fique na forma correta "*validade*".

Exemplo 3: Neste aspecto ainda não tem sido muito abordada a *validez*/ (**validade**) dessa *assunção* / (**afirmação**) ...

A constatação anterior incentivou a geração dos primeiros erros artificiais do projeto HABLA. Para gerar erros produzidos pela transferência de sufixos foi criada uma tabela com as equivalências dos sufixos nas línguas espanhol e português. Essa tabela foi construída por uma linguista³, que colaborou em diferentes tarefas deste projeto. A tabela completa com os sufixos equivalentes é apresentada no Apêndice D, sendo que alguns destes sufixos são idênticos, enquanto outros são diferentes.

³Magali Sanches Duran

Após a construção da lista paralela de sufixos em ambas as línguas, foram extraídas listas de palavras em espanhol que compartilhavam o mesmo sufixo. Para isso, foi utilizado o dicionário em espanhol⁴, da Universidade de Carlos III de Madrid. Esse dicionário é composto por 54.000 lemas e suas formas derivadas. Para construir o léxico com os erros de mudança de sufixo, o sufixo de cada palavra em espanhol da lista foi substituído pelo sufixo equivalente em português, segundo a tabela do Apêndice D.

O próximo passo da geração de erros de sufixos foi verificar se a palavra gerada pertencia ou não ao português. Para isso, foi utilizado o dicionário em português⁵ desenvolvido no NILC e que contém aproximadamente 880.000 palavras. As palavras geradas que estavam no dicionário em português eram cognatas nas duas línguas, enquanto as palavras geradas que não estavam no dicionário foram etiquetadas como “erros”. Nesse último caso foi necessário procurar a equivalência correta em português. Para isso, foi empregado o método estatístico combinado com a similaridade ortográfica, apresentado na Seção 6.1.1.2, que obteve uma precisão de 92%.

Como o método que gera os erros artificiais é combinado com o método que procura as equivalências mais prováveis em português, é possível criar um recurso que mantém o erro e a forma correta. Esse recurso pode ser útil em tarefas de detecção e correção de erros.

Nos experimentos foram gerados os erros para todos os sufixos do Apêndice D. Na Tabela 6.4 são mostrados alguns exemplos dos erros artificiais gerados para os pares de sufixo “*aje-agem*”. Na primeira coluna da Tabela 6.4 é mostrada a lista com as palavras em espanhol que terminam com o sufixo “-*aje*”. Como a lista de palavras foi extraída de um dicionário, e como o dicionário é um recurso linguístico limitado, podem existir outras palavras do espanhol com a terminação “-*aje*” que não foram incluídas no recurso de erros artificiais. Na coluna 2 da Tabela 6.4 são listadas as palavras geradas e na coluna 3 da mesma tabela mostra-se como as palavras geradas podem ser erros, porque não estão no dicionário em português. Evidentemente, como o dicionário em português é um recurso finito algumas palavras geradas podem existir em português, mas são classificadas como erro, pelo fato delas não estarem no dicionário. As palavras geradas e achadas no dicionário são similares ortograficamente e desta forma são cognatos das línguas português e espanhol. Para essas palavras não é necessário procurar sugestões, porém as palavras

⁴<http://www.datsi.fi.upm.es/coes/>

⁵<http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html>

geradas e classificadas como erros podem ser possíveis erros dos hispano falantes quando produzem textos em Português. As palavras classificadas como “Erros”, automaticamente, fazem parte do léxico de erros “*Spanish-native-like errors*”.

Tabela 6.4: Exemplo de erros artificiais da categoria Transferência de Sufixos

Espanhol	Português (Gerado)	Classe	Tradução Correta
<i>lenguaje</i>	<i>linguagem</i>	Erros	linguagem
<i>linaje</i>	<i>linagem</i>	Erros	linhagem
<i>pasaje</i>	<i>pasagem</i>	Erros	passagem
<i>peaje</i>	<i>peagem</i>	Erros	pedágio
<i>drenaje</i>	<i>drenagem</i>	Existe	
<i>embalaje</i>	<i>embalagem</i>	Existe	

Na Tabela 6.5 foi selecionada uma amostra dos sufixos considerados no experimento para mostrar o resultado obtido com a geração de erros. Na primeira coluna da Tabela 6.5, aparecem os pares de sufixos espanhol-português selecionados; na segunda coluna, é mostrada a quantidade de palavras do espanhol extraídas do dicionário para cada sufixo; a terceira coluna apresenta o total de palavras que foram geradas e que pertencem ao português; finalmente, na quarta coluna, está o total de palavras geradas que são candidatas a erros.

Tabela 6.5: Amostra da geração de erros de sufixos

Sufixos Paralelos		Geração de Erros		
espanhol/ português	Palavras Espanhol	Palavras Português	Erros Candidatos	
-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	

A similaridade entre o espanhol e o português foi uma das motivações para criar o método de geração de erros. Para medir essas similaridades durante o processo de geração de sufixos foram comparadas as palavras em espanhol, extraídas do dicionário (coluna dois da Tabela 6.5) com as palavras geradas que pertencem ao português (coluna três da Tabela 6.5) e as traduções manuais dos possíveis erros. Evidências dessas similaridades são observadas

na Figura 6.3, sendo que 91,90% das palavras equivalentes em ambas línguas têm lexemas idênticos ou diferentes em uma única letra. Além disso, na Figura 6.3 observa-se que 3,91% dos casos usam um sufixo em espanhol, enquanto empregam outro diferente em português. Finalmente, em 4,41% dos casos os lexemas das palavras equivalentes diferem em pelo menos duas letras.

Figura 6.3: Similaridade entre substantivos do espanhol e do português.

Como comentado, a mesma abordagem para gerar erros artificiais foi proposta nesta pesquisa para gerar outros tipos de erros. Para isso, aplicaram-se heurísticas diferentes gerando erros lexicais que foram testados na tarefa de detectar e corrigir erros, produzidos pela interferência da língua nativa, como será mostrado na Seção 6.2.3.

6.2 Geração de Léxicos Bilíngues para Apoiar a Escrita dos Hispano Falantes

Nesta Seção, são apresentados os procedimentos para compilar os léxicos bilíngues propostos neste projeto de pesquisa.

6.2.1 Cognatos e Falsos Cognatos

O léxico de cognatos e falsos cognatos tem a finalidade de alertar o hispano falante da existência desses itens lexicais que, quando usados incorretamente, provocam problemas de compreensão dos textos. Como comentado na Seção 3.3.1, cognatos entre duas línguas são pares de palavras que são traduções mútuas e, além disso, ortograficamente similares. Por outro lado, os falsos

cognatos são pares de palavras que compartilham a ortografia, mas tem sentidos diferentes.

Entre línguas próximas como o espanhol e o português são inúmeros os léxicos similares. Conhecer esses léxicos é uma vantagem, pois o hispano falante consegue aproveitar muito conhecimento do espanhol na escrita em português. Da mesma forma, a similaridade entre as línguas atrapalha o hispano falante, porque às vezes os léxicos não são cognatos e são falsos cognatos, que usados inadequadamente podem representar um erro. Seguindo essa linha, disponibilizar um recurso que alerte os hispano falantes caso um falso cognato for utilizado é muito útil. Esse recurso daria mais segurança nas tarefas da escrita, pois quando usado um falso cognato o aprendiz sabe que precisa verificar se o sentido do item é o indicado no contexto específico.

Para identificar automaticamente cognatos de falsos cognatos e, consequentemente, criar o recurso bilíngue, nesta pesquisa foram realizados vários experimentos. Nos experimentos, diversas configurações foram testadas para criar um recurso que auxiliasse os hispano falantes mediante alertas. O recurso compilado oferece o primeiro auxílio da Figura 6.1; ele consegue indicar que o hispano falante usou um falso cognato e mostrar sentenças em que o item foi utilizado. Da mesma forma, o léxico pode indicar se um determinado item em espanhol tem um cognato ou um falso cognato em português.

Segundo a revisão da literatura da Seção 3.3.1, os estudos dedicados a identificar cognatos e falsos cognatos reportam melhor desempenho quando combinam várias abordagens. Nesta pesquisa, os experimentos para separar cognatos de falsos cognatos combinam aspectos das abordagens propostas por Frunza e Inkpen (2009) e Mulloni et al. (2007). Medidas de similaridade ortográficas e fonéticas e probabilidades de tradução de palavras foram utilizadas para descrever os dados, que foram classificados em cognatos e falsos cognatos mediante algoritmos de classificação.

O método proposto para gerar o recurso de cognatos e falsos cognatos segue o procedimento apresentado na Figura 6.4. Como indicado na Figura, o método utiliza como base de conhecimento o corpus paralelo Fapesp e vários dicionários online das línguas espanhol e português. Os dicionários online são utilizados no segundo passo do algoritmo para extrair cognatos e falsos cognatos entre as línguas. As palavras extraídas são classificadas manualmente e depois algoritmos de classificação são aplicados para separar cognatos de falsos cognatos.

Passo 1: Pré-processamento

Figura 6.4: Procedimento para gerar o léxico bilingue de cognatos e falsos cognatos.

A preparação do corpus paralelo é o primeiro passo deste procedimento. Nestes experimentos, o corpus paralelo foi utilizado para gerar um dicionário bilingue entre as línguas espanhol e português, por meio da ferramenta *NA-Tools*⁶. Para isso, como mostrado na Figura 6.4, o corpus paralelo precisa ser tokenizado e alinhado sentencial e lexicalmente, segundo os procedimentos descritos na Seção 6.1.1.2. A saída deste primeiro passo é um léxico bilingue do português e espanhol, que contém todas as palavras que aparecem no corpus. Para cada palavra, em espanhol, incluída no léxico bilingue foram obtidas uma ou várias traduções. As traduções aparecem ordenadas segundo a probabilidade de que a palavra em português seja tradução da palavra em espanhol, de forma que, a primeira palavra é a melhor tradução e a última é a mesmo provável. Essa probabilidade de tradução será empregada no passo 3 do algoritmo.

Passo 2: Levantamento de Listas de Cognatos e Falsos Cognatos

No segundo passo do método proposto, 980 pares de palavras foram classificadas manualmente em Cognatos, Falsos cognatos e Palavras Não Relacionadas. Dois anotadores treinados, um nativo do português e outro nativo do espanhol realizaram a classificação manual. As palavras classificadas como Não Relacionadas eram sinônimos nas duas línguas que não eram similares ortográfica ou foneticamente. Estatísticas do conjunto de dados, classificado manualmente, são mostradas na Tabela 6.6. As palavras foram selecionadas dos seguintes recursos: dicionário online português do Brasil e espanhol⁷;

⁶NATools é um conjunto de ferramentas desenvolvidas para trabalhar com corpus paralelos, disponível livremente em <http://natura.di.uminho.pt/natura/natura/> sob as especificações da *GNU-General Public License*

⁷<http://www.myzips.com/software/ACCESS-WordlistSpanishPortuguese.phtml>

dicionário online português-espanhol⁸; lista das palavras mais frequentes em português e espanhol⁹ e a lista online de diferentes palavras em português e espanhol¹⁰. No conjunto ilustrado na Tabela 6.6 não há palavras compostas e os números que aparecem entre parênteses indicam que existem 122 palavras cognatas e 199 falsos cognatos com ortografia idêntica. Este último fato aumenta a complexidade dos métodos de classificação propostos no Passo 3 do algoritmo.

Tabela 6.6: Conjunto de treinamento para separar cognatos de falsos cognatos (o número entre parêntesis indica o total de palavras com ortografia idêntica)

Conjunto de dados de treinamento	
Cognatos	338 (122)
Não Relacionadas	238 (0)
Falsos Cognatos	372 (199)
Total	948

Passo 3: Classificação de Cognatos e Falsos Cognatos

O terceiro e último passo do método proposto concentrou-se em testar algoritmos de classificação para separar cognatos de falsos cognatos.

A primeira etapa de qualquer algoritmo de classificação é selecionar o conjunto de características (*features*) que descreverão os dados. Nestes experimentos, as instâncias que compõem os dados são pares de palavras em espanhol e em português. Para descrever os dados foram selecionadas as seguintes sete medidas de similaridade ortográficas: IDENT, PREFIX, DICE, TRIGRAM, LCSR, EDIT, NED. Todas as medidas utilizadas para caracterizar os dados já foram descritas na Seção 3.3.1. Uma medida de similaridade fonética foi incluída utilizando como base o algoritmo SOUNDEX. A implementação em *Python* do algoritmo SOUNDEX para a língua portuguesa foi feita nesta pesquisa, mediante a transformação de algumas regras e tomando-se como base a implementação do algoritmo para o espanhol¹¹. Para estimar a similaridade entre os códigos gerados pelo algoritmo SOUNDEX, a distância NED foi empregada. Além de considerar as medidas individuais como características de classificação foi considerada também a média destas medidas.

O conjunto de medidas de similaridade foi estimado para os pares de palavras português-espanhol e esses valores foram utilizados como entrada dos

⁸<http://en.bab.la/dictionary/spanish-portuguese/>

⁹<http://www.wordfrequency.info/spanish-portuguese.asp>

¹⁰<http://rudhar.com/lingtics/ptesdiff.htm>

¹¹Disponível em: <http://oraclenotepad.blogspot.com/2008/03/soundex-en-espaol.html>

algoritmos de classificação empregados na tarefa de classificar pares de palavras nas classes cognatos, falsos cognatos ou palavras não relacionadas. Desta forma, o conjunto de dados foi representado como uma matriz, em que cada linha tinha um par de palavras e as colunas eram as medidas de similaridade que caracterizavam cada par. Diversas modificações foram testadas nos experimentos para obter a melhor alternativa de construção do recurso.

Cada experimento procurou modificar o conjunto de características para obter a melhor classificação dos dados. Para isso, foram avaliados diferentes algoritmos de classificação, selecionados do pacote *Weka* (Witten e Frank, 2005). Além disso, avaliou-se a influência das medidas de similaridade na tarefa de classificação. Os três experimentos realizados são relatados abaixo.

No primeiro experimento (**Experimento 1**), o conjunto de dados foi dividido em duas classes: Cognatos/Falsos Cognatos e Não-Relacionados. Essa divisão foi possível porque as características ortográficas desses conjuntos eram similares. Após avaliar a classificação do primeiro experimento, em um segundo experimento (**Experimento 2**) foi avaliado o desempenho dos algoritmos na divisão do conjunto de dados nas classes Cognatos e Falsos Cognatos.

Antes de avaliar o desempenho dos algoritmos de classificação, empregou-se o algoritmo *InfoGainAttributeEval* do pacote *Weka* para estimar a importância de cada atributo na tarefa de classificação. Esse algoritmo é uma técnica de seleção de atributos que ordena os atributos que oferecem mais informação na classificação. Na Figura 6.5, pode-se observar o comportamento dos 8 atributos em cada experimento. A medida de similaridade LCSR foi a melhor ordenada nos dois experimentos. No experimento 1 foram eliminados, incrementalmente, os atributos com pior classificação até ficar somente com o atributo LCSR como única características de classificação. O algoritmo IBK, implementação do KNN em *Weka*, com $K=3$ manteve um desempenho de 95,56% empregando a medida LCSR como atributo de classificação. O desempenho do KNN mudava apenas nas casas decimais com a eliminação dos atributos com pior classificação. A medida EDIT e a média de todas as medidas de similaridade se mostram como atributos decisivos na classificação dos algoritmos. Como resultado esperado, observamos que a medida que menos informação ofereceu foi a IDEM. Segundo o algoritmo de seleção de atributos, as medidas de similaridade ortográfica não oferecem muita informação para o algoritmo de classificação no Experimento 2. Desta forma, era esperada uma queda no desempenho dos algoritmos nesse experimento.

Utilizando a técnica de validação cruzada 10 *folds* para amostrar os dados,

Figura 6.5: Resultados da técnica de seleção de atributos para cada experimento

o conjunto de treinamento em cada execução foi dividido em 10 partições, 9 para treinamento e 1 para teste. Usando o pacote Weka, foram testados os seguintes algoritmos de classificação: o método probabilísticos *Naïve Bayes*, IBK, *Multilayer Perceptron*, *Support Vector Machine* (SMO), o método baseado em regras (JRip) e o algoritmo de árvore de decisão J48. Diferentes modificações foram realizadas nos algoritmos de classificação para se obter seus melhores desempenhos no nosso conjunto de dados. Por exemplo, no caso do algoritmo IBK, observou-se variações no desempenho, variando o número de vizinhos K. No Experimento 1, o melhor resultado foi para K=3, enquanto que para o Experimento 2 foi para K=9. Os diferentes valores de K indicam que, para separar cognatos/falsos cognatos de palavras não relacionadas com a análise de 3 vizinhos é suficiente, já para separar cognatos de falsos cognatos é necessário investigar as características de um número maior de vizinhos.

Na Tabela 6.7, ilustram-se os resultados obtidos para cada experimento usando os algoritmos de classificação. O melhor desempenho foi obtido com o algoritmo IBK, enquanto o pior foi com *Naïve Bayes*. É observado nessa tabela que os algoritmos separam com uma alta taxa de acerto as classes Cognatos/Falsos Cognatos das Não Relacionadas (Experimento 1). Esse resultado está relacionado ao fato de que as palavras Não Relacionadas apresentam pouca similaridade ortográfica, enquanto os Cognatos e Falsos Cognatos são muito similares. No Experimento 2, pode-se observar uma queda brusca no desempenho dos algoritmos; como esperado, esse era o grande gargalo do nosso problema: conseguir identificar os Cognatos dos Falsos Cognatos. Na Figura 6.5, observa-se que as medidas de similaridade não oferecem muita informação para a classificação de cognatos e falsos cognatos. Uma análise

do número de pares de palavras classificadas incorretamente pelo algoritmo de melhor desempenho, IBK, mostrou que o número de falsos positivos no Experimento 2 era alto.

Tabela 6.7: Desempenho dos classificadores testados para os experimentos propostos

Classificadores	Experimento 1	Experimento 2
Instâncias classificadas corretamente		
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%

Um dos fatores que causaram os erros na separação de Cognatos e Falsos Cognatos está relacionado com o número elevado de palavras com ortografia idêntica no conjunto de treinamento. Após observarmos os resultados insatisfatórios do Experimento 2, propusemos um terceiro experimento, no qual foi adicionado um novo atributo baseado na probabilidade de traduzir palavras do espanhol para o português.

No **Experimento 3**, um atributo chamado tradução foi adicionado como uma nova característica dos dados. O atributo tradução foi obtido procurando os significados das palavras no dicionário bilíngue espanhol-português, gerado a partir do corpus paralelo no Passo 1 do algoritmo. Esperava-se que as palavras classificadas como Cognatos ou Não Relacionadas aparecessem como traduções mútuas no dicionário, enquanto os Falsos Cognatos deviam aparecer com significados diferentes dependendo do contexto em que foram usados. No dicionário construído via o corpus da Fapesp, as palavras podiam ter vários significados dependentes da probabilidade de ter acontecido no corpus. A probabilidade de ocorrência das palavras foi empregada como característica de classificação. Se a palavra aparecia no dicionário essa era sua probabilidade de ocorrência, caso contrário a probabilidade era zero. Uma das deficiências desse experimento é que muitas das palavras que pertenciam à lista original não apareciam no corpus. Na Tabela 6.8 ilustra-se o número de palavras que aparecem no dicionário. O número entre parênteses são as palavras do conjunto de dados que não aparecem no dicionário.

Novamente, foi aplicada a técnica de seleção de atributos, dividindo o conjunto de dados em dois experimentos: no Experimento 3.1 em 3 classes (Cog-

Tabela 6.8: Número de palavras no dicionário (número entre parênteses representa o número de palavras do conjunto de treinamento que não estavam no dicionário).

Cognatos	338 (116)
Não Relacionadas	238 (95)
Falsos Cognatos	372 (172)
Total	948

atos, Falsos Cognatos e Não Relacionadas), e no Experimento 3.2, em 2 classes Cognatos/Falsos Cognatos. Desta forma, é possível verificar na Figura 6.6 que no Experimento 3.1 o atributo pior ranqueado foi a tradução, enquanto que no Experimento 3.2 foi o melhor ranqueado. Esse resultado é interessante porque se confirmou que o atributo tradução contribui para a diferenciação das classes Cognatos/Falsos Cognatos.

Figura 6.6: Resultados da técnica de seleção de atributos para cada experimento

Na Tabela 6.9, apresenta-se o desempenho dos algoritmos de classificação para o Experimento 3.2. Pode-se observar que a adição do novo atributo de tradução aumentou o desempenho de todos os classificadores. Desta vez, o IBK continuou mostrando melhor desempenho, enquanto o *Naive Bayes* foi o pior. Acredita-se que se o número de palavras do conjunto de treinamento contidas no dicionário tivesse sido maior, os desempenhos dos algoritmos também teriam aumentado. No Experimento 3.2, o número de falsos positivos continua sendo alto, mais o desempenho melhorou significativamente se comparado com o obtido no Experimento 2. O Experimento 3 foi similar aos experimentos apresentados por Mulloni et al. (2007) nos quais foram combinadas características ortográficas e semânticas para identificar cognatos e falsos cognatos entre os pares de línguas inglês e espanhol e foi obtido um

desempenho de 91,85%. Entretanto, é importante observar que esses autores trataram o problema da ausência de palavras no dicionário, sendo essa uma das deficiências do nosso Experimento.

Tabela 6.9: Desempenho dos classificadores testados no Experimento 3.2

Classificadores	Experimento 3.2
	Instâncias classificadas corretamente
<i>Naive Bayes</i>	71,73%
IBK (K = 9)	76%
SVM	72%
<i>MultilayerPerceptron</i>	76%
JRIP	75%
J48	74%

O léxico bilíngue de cognatos e falsos cognatos foi criado considerando a configuração apresentada no experimento 3. Dessa forma, obteve-se uma lista de pares de palavras do espanhol e do português que estão categorizadas em cognatos e falsos cognatos. Como comentado, esse recurso pode ser muito útil alertando os hispano falantes caso um item identificado como falso cognato for utilizado. O ideal é apresentar junto com o alerta, sentenças em que o falso cognato é utilizado com sentidos variados, caso o item tiver mais de um sentido. Desta forma, o hispano falante além do alerta, consegue observar exemplos e detectar qual é o sentido que ele deseja utilizar. Porém, para isso é necessário ter um corpus com exemplos do falso cognato e seus sentidos. Além disso, para levantar sentenças com diferentes sentidos é necessário ter um desambiguador lexical, tarefa que embora tem sido muito estudada (Sobrevilla Cabezado, 2015; Nóbrega e Pardo, 2014), somente recentemente foram disponibilizados métodos para o português para desambiguação de verbos e substantivos; os trabalhos citados são contemporâneos a esta pesquisa de doutorado, inviabilizando a aplicação das ferramentas. Por outro lado, foi confirmado que a maioria dos falsos cognatos entre o espanhol e o português acontecem na língua falada e são poucos os exemplos na escrita, sendo menor ainda o número de exemplos presentes no corpus paralelo Fapesp, que é um corpus com textos que descrevem pesquisas acadêmicas. A partir destas considerações, na maioria das vezes, o léxico é incapaz de mostrar sentenças de exemplo dos falsos cognatos.

6.2.2 Verbos

O léxico bilíngue de verbos dá suporte aos dois primeiros auxílios propostos nesta pesquisa. Ele pode ser utilizado pelos hispano falantes para consultar qual é a melhor equivalência em português de um verbo em espanhol e para obter os melhores sinônimos de um verbo em português. É importante ressaltar que os sinônimos foram ordenados segundo seu uso na escrita acadêmica, guiando o aprendiz nas escolhas dos verbos mais frequentes nos textos acadêmicos.

Como mostrado na Seção 5.3.3 os hispano falantes têm dificuldades em selecionar/utilizar os verbos utilizados na escrita acadêmica em português. As inadequações são evidenciadas por meio do uso de verbos com ortografia equivalente ou muito similar nas duas línguas, mas que não são frequentes na escrita em português ou verbos que são utilizados em português, mas não na escrita acadêmica. Esse tipo de problema foi categorizado durante a anotação apresentada na Seção 5.3.3 dentro da categoria Convencionalidade. A maioria desses problemas representam inadequações e não erros e é complexa sua identificação automática, pois é subjetivo apontar que um item lexical é utilizado ou não em textos acadêmicos. Inclusive, um item pode ser adequado para textos acadêmicos de uma área específica, enquanto para textos de outras áreas não é. Os verbos sublinhados nos Exemplo 1 e 2 mostram essas inadequações que se distanciam do discurso dos nativos e marcam o discurso dos hispano falantes.

Exemplo 1: Os pesquisadores recopilaram/(**compilar**) informação sobre as fortalezas da aplicação e entre elas encontrou-se que os estudantes se sentem motivados ao realizarem as atividades com o porta-fólio, e, com isto, demonstram um bom desempenho.

Exemplo 2: Enquanto que as políticas de CTI estavam sendo adotadas e implementadas dentro de tais modelos, a taxa de câmbio tecnológico acelerou grandemente e ocorreram mudanças no processo de inovação, passando de um modelo linear a um não linear e evolucionando/(**evoluindo**) para um modelo "aberto".

Como mencionado na Seção 3.3.2, os recursos que aparecem na literatura para apontar os verbos em português mais frequentes na escrita acadêmica são insuficientes, pois eles são pequenos e muitos dos verbos mais frequentes não aparecem nas listas. Seguindo essa linha, nesta seção é apresentado o método de compilação de um recurso criado para guiar o hispano falante nas escolhas dos verbos mais utilizados na escrita acadêmica.

Na Figura 6.7, mostra-se a construção do léxico bilingue de verbos, realizada em quatro passos. O primeiro passo consiste em selecionar o corpus, que utilizado como base de conhecimento permite a extração do léxico bilingue de verbos. O segundo passo do algoritmo concentra-se na extração da lista de verbos das línguas português e espanhol. Para isso, foi necessário anotar morfossintaticamente os erros e filtrar alguns erros que aconteceram no processo de lematização e etiquetamento morfossintático. No terceiro passo, os verbos em espanhol são utilizados para extrair do corpus paralelo as respectivas equivalências em português, usando o algoritmo de tradução estatística. Por último, no quarto passo, os verbos em português são agrupados em sinônimos, que são ordenados segundo sua frequência no gênero acadêmico.

Figura 6.7: Procedimento para gerar o léxico bilingue de verbos.

Passo 1: Pré-processamento

Esta etapa consiste em preparar o corpus paralelo Fapesp para usá-lo como base de conhecimento para extrair os pares de verbos em espanhol e português. Para isso, o método de tradução estatística apresentado na Seção 6.1.1.2 foi utilizado e o corpus foi tokenizado, separado em sentenças e alinhado segundo o procedimento especificado na mesma seção.

Para compilar uma lista monolíngue de verbos em português foi utilizado o corpus acadêmico, apresentado na Seção 5.1.2, e um corpus da escrita geral, composto por textos do Projeto *Lácio-Web*¹², especificamente do gênero jornalístico e de assunto religioso.

¹²<http://www.nilc.icmc.usp.br/lacioweb/>

Os corpus acadêmico e geral foram lematizados e etiquetados morfossintaticamente. Para etiquetar morfossintaticamente os textos em espanhol se empregou o etiquetador *TreeTagger*¹³ (Schmid, 1995), já para anotar os textos em português utilizou-se a ferramenta *MXPOST* (Ratnaparkhi, 1996). Ambos os etiquetadores utilizam métodos estatísticos para, baseados em corpus, realizar a anotação das etiquetas morfossintáticas. A principal motivação para empregar o lematizador e anotador morfossintático *TreeTagger* para a língua espanhola foi o fato de ser uma ferramenta independente de língua e que, além disso, tem sido utilizada em diversos trabalhos que estudam o espanhol. O *Treetagger* foi treinado com textos da língua espanhola e esses modelos estão disponíveis livremente na Web¹⁴. Por outro lado, o *MXPOST* foi treinado com os textos em português do corpus *Mac-Morpho*, cujo modelo está disponível no Projeto Lácio-Web¹⁵.

Passo 2: Extrair e filtrar verbos

O segundo passo do algoritmo extrai automaticamente listas monolíngues de verbos do espanhol e do português. Neste passo usam-se as etiquetas morfossintáticas para extrair dos corpus as listas de verbos lematizados e suas frequências. A lista monolíngue de verbos em espanhol foi extraída dos textos em espanhol do corpus paralelo Fapesp. Por outro lado, a lista monolíngue de verbos em português foi extraída dos textos que integram o corpus acadêmico, dessa forma somente verbos que aparecem em textos acadêmicos foram incluídos. Inicialmente, foram extraídas 562.139 ocorrências de verbos em espanhol, que representaram 3.372 lemas de verbos. Para o português foram extraídas 267.743 ocorrências de verbos que representavam 6.927 lemas de verbos.

Uma análise superficial das listas monolíngues geradas mostrou que muitos dos verbos que apareciam com baixa frequência no corpus, eram itens errados. Nesses casos estavam incluídos verbos com erros de grafia, símbolos, palavras não reconhecidas pelos etiquetadores, palavras em outras línguas e participípios passados usados como substantivo ou adjetivos.

Espúrios da lista de verbos foram eliminados por meio de um filtro que retirava da lista todas as palavras que não terminavam em *-ar*”, *-er*” e *-ir*”, lembrando que os verbos da lista estavam lematizados. Após aplicar o filtro a lista de verbos em espanhol ficou com 487.615 ocorrências que representam 2.778 lemas, enquanto a lista de verbos em português contém 255.614

¹³<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

¹⁴<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

¹⁵<http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html>

ocorrências que representam 3.099 lemas. A maioria das palavras etiquetadas erradamente pelo *TreeTagger* incluíam verbos seguidos do pronome "se": (*comunicarse, volverse, definirse, aplicarse*). Os erros cometidos pelo *MXPOST* incluíam: palavras do inglês (*economic, development, large, analysis* e passados participípios fazendo função de adjetivos (direcionada, atendida)).

Passo 3: Extrair equivalentes dos verbos em ES para o PT

Para extrair do corpus Fapesp os equivalentes de tradução em português para os verbos em espanhol, foi utilizado o modelo de tradução criado com o *toolkit* Moses e apresentado na Seção 6.1.1.2. O procedimento procura traduções candidatas em português para cada verbo em espanhol que aparece na lista levantada no Passo 2. A lista de equivalências é filtrada mantendo unicamente os verbos incluídos na lista de verbos em português, também construída no Passo 2 do algoritmo. Dessa forma, há garantias que as equivalências dos verbos em espanhol são utilizadas em textos acadêmicos. Por exemplo, para o verbo "*fijar*" foram obtidas as seguintes traduções: "fixar"; "estabelecer"; "fazer"; "ver"; "selecionar". A ordem em que aparecem as equivalências é a ordenação realizada pelo método de tradução que indica que a melhor equivalência do verbo "*fijar*" é "fixar".

Como observado nos exemplos, o auxílio para oferecer as equivalências dos verbos apresenta várias alternativas para que o usuário selecione a melhor. Quanto mais sentidos tiver o verbo maior a quantidade de equivalências. No entanto, algumas equivalências são menos adequadas do que as outras e cabe ao usuário decidir qual é a melhor escolha. Para ajudar o usuário a decidir foi proposto um método que procura sinônimos dos verbos em português. Desta forma, mais exemplos dos apresentados pelo método de equivalências são mostrados ao usuário. Os sinônimos são ordenados em função de sua adequação ao gênero acadêmico como será explicado no passo quatro do algoritmo.

Passo 4: Obter e ordenar sinônimos em PT

Neste passo é utilizado o algoritmo de sinônimos baseado em tradução, apresentado na Seção 6.1.2, para obter sinônimos para os verbos em português. Esse passo da compilação do léxico bilíngue de verbos é responsável por sustentar o segundo auxílio proposto nesta pesquisa. Este auxílio é preparado para hispano falantes com níveis mais avançado e que procuram o emprego mais diversificado dos verbos nos textos. Para cada verbo em português o recurso disponibiliza um ou vários sinônimos.

A partir do procedimento proposto para obter sinônimos, os sinônimos do

verbo "fixar" foram os seguintes: "realizar"; "colocar"; "fazer"; "ver"; "prever"; "dar"; "estabelecer"; "implantar"; "reforçar"; "aprofundar"; "assentar"; "aderir"; "combinar"; "selecionar"; "colar"; "grudar"; "estipular". Na lista de sinônimos aparecem muitos verbos que representam diversos sentidos do verbo "fixar". No entanto, como o método para obter os sinônimos é baseado na saída do método de tradução estatística e como o método de tradução privilegia os itens lexicais mais frequentes no corpus, nem sempre os sinônimos apresentados são sinônimos verdadeiros. Por exemplo, sabe-se que os verbos "fazer" e "dar" são verbos usados com muita frequência nos textos e, além disso, eles são verbos polissêmicos, que quando combinados com outros verbos podem adquirir os sentidos do outro verbo, mas individualmente eles têm um sentido próprio. No caso do verbo "fixar" tanto "fazer" quanto "dar" não são sinônimos.

Identificada a deficiência do método para obter sinônimos foi testado o recurso *WordNet.Br* para procurar sinônimos de verbos em português selecionados aleatoriamente. Uma característica padrão da saída da *WordNet.Br* foi apresentar listas com muitos sinônimos, sendo uma saída que não é amigável para um usuário aprendiz da língua. Por exemplo, para o verbo "fixar" a *WordNet.Br* apresentou os seguintes sinônimos: "afincar", "afirmar", "afixar", "arbitrar", "arquivar", "arraigar", "assentar", "assinalar", "bater", "consignar", "consolidar", "cravar", "decorar", "decretar", "designar", "determinar", "dispor", "ditar", "embeber", "encravar", "enfincar", "enraizar", "espetar", "estabelecer", "estatuir", "estipular", "estudar", "fincar", "firmar", "imobilizar", "impor", "indicar", "mandar", "memorizar", "meter", "ordenar", "preceituar", "pegar", "prender", "prescrever", "segurar".

Os sinônimos apresentados pela *WordNet.Br* para o verbo "fixar" somavam 41, enquanto o método baseado em tradução apresentou 17 sinônimos. Apesar do método baseado em tradução apresentar menos sinônimos que a *WordNet.Br* muitos dos sentidos que apareceram na saída da *WordNet.Br* também apareceram na saída do método de tradução. Em alguns casos o mesmo sentido aparece com diferentes palavras. Por exemplo, a *WordNet.Br* indica que "cravar" é um sinônimo de "fixar", enquanto o método de sinônimos por tradução indica "aprofundar"; "cravar" e "aprofundar" podem ser considerados sinônimos em alguns contextos. O verbo "cravar" é menos utilizado na escrita acadêmica do que o verbo "aprofundar", com isso mostra-se que a saída do método baseado em tradução indica palavras mais adequadas na escrita de textos acadêmicos. Esse fato acontece porque o método extrai os sinônimos de um corpus paralelo que usa léxicos comuns da escrita acadêmica e mantém

somente os verbos equivalentes que aconteceram em textos do corpus acadêmico. Lembrando que o objetivo perseguido por meio do léxico dos verbos é guiar o hispano falante na escolha de verbos comuns da escrita acadêmica, os verbos extraídos pelo método de tradução podem ser mais adequados que os apresentados pela *WordNet.Br* que é um recurso da língua geral.

Combinando os motivos descritos, acredita-se que o método baseado em tradução pode ser mais adequado para auxiliar aos hispano falantes. Contudo, o método ainda apresenta sinônimos que não são sinônimos, como os verbos "fazer" e "dar", se no léxico monolíngue tivessem sido incluídas expressões verbais, provavelmente, o método conseguiria recuperar essas expressões como possíveis sinônimos. Como expressões verbais não foram consideradas, os verbos "fazer" e "dar" deveriam sair da lista de sinônimos do verbo "fixar" ou pelo menos aparecerem nas últimas posições da lista.

Com essa finalidade, a lista de sinônimos foi ordenada em função da proximidade na sinonímia dos verbos e de sua ocorrência nos textos acadêmicos. Após a ordenação é possível fixar um número determinado de equivalentes do verbo.

Para ordenar os sinônimos considerando a proximidade com o verbo em questão a saída da *WordNet.Br* é utilizada. Embora algumas deficiências desse recurso fossem detectadas ele é um recurso útil para qualquer aplicação de PLN. Analisando os sinônimos de ambos os recursos para o verbo "fixar" observa-se que os seguintes três verbos coincidem: "estabelecer", "estipular" e "assentar". O fato dos três verbos ocorrerem em ambas as saídas é um indicativo de que eles são realmente sinônimos. Desta forma, para ponderar a lista de sinônimos, aqueles sinônimos que aconteceram em ambas as saídas receberam um peso fixo de 1,5. Esse peso forma parte da combinação linear que além de considerar a ocorrência do verbo nas duas listas de sinônimos, considera a ocorrência do verbo no corpus acadêmico.

Por outro lado, para determinar se um verbo é adequado na escrita acadêmica foi calculada a *Tf-idf*¹⁶(Luhn, 1957) do verbo nos corpus geral e acadêmico. *Tf-idf* é uma estatística numérica que reflete como um termo é importante em um documento ou coleção de documentos. Ele é frequentemente usado como um fator de ponderação na recuperação de informação e de mineração de texto. O valor *Tf-idf* aumenta proporcionalmente ao número de vezes que uma palavra aparece no documento, mas é compensado pela frequência da palavra no corpus, o que ajuda a controlar o fato de que algumas palavras

¹⁶do inglês *Term Frequency-Inverse Document Frequency*

são geralmente mais comuns do que outras. Uma análise similar foi desenvolvida por Simpson-Vlach e Ellis (2010) para levantar expressões formulaicas da língua geral e da escrita acadêmica, porém Simpson-Vlach e Ellis (2010) fizeram o levantamento estimando o logaritmo da probabilidade das expressões.

Matematicamente, a *Tf-idf* é calculada segundo a Equação 6.1, que multiplica duas estatísticas: a frequência do termo e a frequência inversa do termo no documento. A frequência do termo é calculada segundo a seguinte operação logarítmica: $1 + \log(f_{i,j})$, em que $f_{i,j}$ é a frequência do termo i no documento j . A frequência inversa do termo é calculada segundo a fração logarítmica $\log(\frac{N}{n_j})$, que divide o total de documentos no corpus N , pelo total de documentos n_j em que o termo aconteceu.

$$w_{i,j} = \begin{cases} (1 + \log(f_{i,j})) \times \log(\frac{N}{n_j}) & \text{se } f_{i,j} \geq 0 \\ 0 & \text{se caso contrário} \end{cases} \quad (6.1)$$

Termos mais frequentes dentro de um documento e termos mais raros possuem um peso *Tf-idf* maior. Seguindo essa linha, o cálculo da *Tf-idf* pode ser utilizado para determinar a importância de um termo na escrita acadêmica e na geral. Enquanto maior for o valor de *Tf-idf*, mais significativo vai ser o termo. Por exemplo, o *Tf-idf* do verbo “ser” é pequeno porque ele é um verbo usado na maioria dos textos, isto é, não é relevante para nenhum dos textos. Como consequência, esse valor é equivalente tanto no corpus acadêmico quanto no corpus geral, o que mostra que o termo é usado da mesma forma nos dois corpus. A partir dessa análise, supõe-se que quando o *Tf-idf* do termo é maior no corpus acadêmico do que no geral é um indicio de que o termo é adequado para a escrita acadêmica. Caso contrário, não é possível afirmar nada, mas se existir um sinônimo adequado na escrita acadêmica e outro que segundo a heurística não é, pode-se ponderar os termos de forma a privilegiar o primeiro caso.

Seguindo essa análise, o valor da subtração do *Tf-idf* dos verbos no corpus acadêmico e geral foi incluído na combinação linear para ordenar os sinônimos. Na Figura 6.8 mostra-se um gráfico com o valor da subtração do *Tf-idf* dos verbos em português, incluídos no recurso. Analisando o gráfico, pode-se inferir que os verbos para os quais o valor da subtração é positivo são verbos que são utilizados na escrita acadêmica. Alguns dos picos na região positiva do gráfico são os valores da subtração dos verbos “sumarizar”, “desambiguar” e “instanciar”, que são verbos frequentes nos textos acadêmicos de PLN. Por

outro lado, os valores perto de zero na Figura 6.8 representam os seguintes verbos: "estar", "dever", "ficar", "considerar", "ocorrer", "existir". Finalmente, dentre os verbos que tiveram um *Tf-idf* negativo encontram-se:

- palavras de outra língua como: "over", "mother";
- verbos escritos erradamente: "assitir", "sequir";
- verbos que não acontecem nos textos acadêmicos: "descabelar", "remendar", "tamponar", "provisionar", "outorgar", "instigar".

Os exemplos mostraram que o cálculo do *Tf-idf* é útil para penalizar os verbos que não acontecem com frequência nos textos acadêmicos e ordenar os sinônimos de cada verbo segundo esse critério.

Figura 6.8: Subtração da *Tf-idf* nos corpúsculo acadêmico e geral para os verbos em português.

Observando os exemplos dos verbos, para os quais a subtração do *Tf-idf* do corpúsculo acadêmico e geral foi negativa, foram levantados alguns erros do processo de filtragem, Passo 2 do algoritmo proposto. Uma alternativa para filtrar esses erros é incluir na análise um léxico do português de forma que os verbos que não aparecem no léxico sejam excluídos. Para essa finalidade poderia ser utilizado o léxico do NILC¹⁷ (Nunes et al., 1996).

A combinação linear para ordenar os sinônimos dos verbos foi calculada como a seguir:

¹⁷Disponível em: <http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html>

$$ordena_{o} = 1,5 * Acontece_{NaWordNet.Br} + (Tf - idf_{Acadêmico} - TF - idf_{Geral}) \quad (6.2)$$

$Acontece_{NaWordNet.Br}$ é uma variável binária que toma os valores zero se o sinônimo não está incluído na lista da *WordNet.Br* e um caso contrário. Por outro lado, o algoritmo que estima o valor de *Tf-idf* considerando a frequência de cada verbo nos corpúscos acadêmico e geral foi implementado em *Python* segundo a fórmula 6.1.

Após estimar a combinação linear para ordenar os verbos, para o verbo "fixar" foram obtidos os sinônimos ordenados como aparecem na Tabela 6.10. Na segunda coluna da tabela aparecem os sinônimos, na terceira aparece o peso dado, caso o sinônimo também tenha sido apontado pela *WordNet.Br*, na quarta coluna aparece o resultado da subtração da *Tf-idf* nos corpúscos acadêmico e geral e na última coluna aparece o resultado da combinação linear, usado para ordenar os sinônimos.

Tabela 6.10: Ordem dos sinônimos do verbo "fixar"

Top	Sinônimos	WordNet.Br	Subtração <i>Tf-idf</i>	Ordenação
5	assentar	1,5	0,28	1,78
	estabelecer	1,5	0,06	1,56
	fixar	1,5	-0,33	1,17
	grudar	0	0,70	0,70
	estipular	1,5	-0,80	0,70
10	colar	0	0,48	0,48
	aderir	0	0,40	0,40
	ver	0	0,36	0,36
	reforçar	0	0,33	0,33
	fazer	0	0,25	0,25
20	aprofundar	0	0,22	0,22
	dar	0	0,17	0,17
	colocar	0	0,16	0,16
	prever	0	-0,007	-0,007
	realizar	0	-0,04	-0,04
	selecionar	0	-0,07	-0,07
	combinar	0	-0,20	-0,20
	implantar	0	-0,21	-0,21

Os primeiros cinco verbos (Top 5) que aparecem na Tabela 6.10 são sinônimos do verbo "fixar". Quatro destes primeiros cinco também foram apontados pela *WordNet.Br* e todos menos os verbo "fixar" e "estipular" foram os menos indicados para usá-los na escrita acadêmica, porém os valores de *Tf-idf* são

próximos de zero. Todos os verbos, incluídos no segundo grupos (Top 10) da Tabela 6.10 também são sinônimos do verbo fixar, exceto o verbo "fazer", mas ele aparece na última posição do Top 10. Todos os verbos incluídos no grupo do Top 10, segundo o calculo do *Tf-idf*, podem ser utilizados na escrita acadêmica. O último grupo da Tabela 6.10 (Top 20) está composto por verbos que não são sinônimos do verbo "fixar". Para o verbo "fixar" o método foi eficiente ordenando os sinônimos, limitar a quantidade de sinônimos na saída representaria uma melhora do método.

O método para obter sinônimos utilizando a saída da tradução ainda não foi validado, uma alternativa para isso, seria realizar uma validação manual do recurso por lexicógrafos ou linguistas. Outra possibilidade é inserir o recurso em uma ferramenta de auxílio à escrita e fazer uma avaliação extrínseca da validade do recurso no auxílio à escrita dos hispano falantes. No Apêndice E apresentam-se alguns verbos em português com seus sinônimos obtidos e ordenados segundo o método proposto. Nesse apêndice é possível observar os verbos obtidos como sinônimos e quais deles aparecem também na *Word-Net.Br*. A saída apresentada no Apêndice E é outra evidência do sucesso do método proposto para obter sinônimos.

O léxico bilíngue de verbos foi implementado de forma que é possível relacionar o verbo em espanhol com os sinônimos de seus melhores equivalentes em português. Por exemplo, considerando a ordenação do algoritmo de tradução, a melhor equivalência para o verbo "*fijar*" em espanhol foi "fixar", Passo 3 do algoritmo. No Passo 4 do algoritmo foi possível obter os sinônimos do verbo "fixar". Segundo essa linha, as saídas dos Passo 3 e 4 são combinadas e é possível mostrar para o aprendiz que o verbo "*fijar*" em espanhol pode ser equivalente aos seguintes verbos em Português: "assentar", "estabelecer", "fixar", "grudar" e "estipular". Além de apresentar os equivalentes, sentenças em que eles foram utilizados também podem ser extraídas dos corpus, auxiliando o aprendiz a escolher a melhor equivalência.

O léxico bilíngue de verbos contém 1.999 verbos em espanhol e 1.897 verbos em português. Da lista monolíngue de verbos em português, somente foram incluídos no léxico bilíngue aqueles verbos que apareceram como equivalências de verbos em espanhol. Além disso, da lista inicial em espanhol foram eliminados 779 verbos, 546 deles não obtiveram equivalências em português. Desse total, para 194 verbos que ocorrem no corpus com uma frequência menor que quatro não foram obtidas traduções e para 12 verbos também não foram obtidas tradução, apesar de ocorrerem no corpus em mais de quatro

ocasiões. Uma alternativa para obter as equivalências desses verbos que não apareceram no *cópus* pode ser empregar o método de similaridade ortográfica, apresentado na Seção 6.1.1.1, porém preferiu-se aplicar esse algoritmo em trabalhos futuros. Para o resto dos verbos que saíram da análise, os equivalentes obtidos não apareceram na lista de verbos em português, criada no Passo 2. Neste último grupo estão incluídos exemplos que representam erros do etiquetador morfossintático.

6.2.3 Marcadores Discursivos

O *léxico* bilíngue de marcadores discursivos oferece todos os auxílios apresentados na Figura 6.1. Nesta seção, o procedimento é apresentado em duas partes. Em uma primeira parte é detalhado o procedimento para se obter o *léxico* bilíngue. O recurso criado nessa primeira parte será capaz de oferecer os dois primeiros auxílios. Na segunda parte desta seção, são detalhados os experimentos propostos para detectar e corrigir erros de marcadores discursivos, oferecendo o último auxílio proposto nesta pesquisa.

Como mostrado na Figura 6.9, a construção do *léxico* bilíngue é separada em quatro passos. O primeiro passo consiste em selecionar um *cópus* paralelo que, utilizado como base de conhecimento, permita a extração do *léxico* bilíngue de marcadores. O segundo passo é levantar na literatura uma lista inicial monolíngue de marcadores discursivos em português e em espanhol, ou seja, listas independentes, sem relação de equivalência. No terceiro passo, os marcadores em espanhol são utilizados para extrair do *cópus* as respectivas equivalências, usando um algoritmo de tradução estatístico. No quarto passo, as traduções obtidas são filtradas pela lista de marcadores em português. No quinto passo, as traduções filtradas são utilizadas para criar agrupamentos entre os marcadores de ambas as línguas (um mesmo marcador pode ter várias equivalências e isso indica que os equivalentes pertencem a uma mesma categoria).

Passo 1: Pré-processamento

Esta etapa consiste em preparar o *cópus* Fapesp para usá-lo como base de conhecimento para extrair os pares de marcadores em espanhol e português. Para isso, o método de tradução estatística apresentado na Seção 6.1.1.2 foi utilizado e o *cópus* foi tokenizado, sentenciado e alinhado segundo o procedimento especificado nessa seção.

Passo 2: Levantamento de listas monolíngues de marcadores discursivos

Figura 6.9: Procedimento para gerar o léxico bilíngue de marcadores discursivos.

O segundo passo do método proposto é dedicado a levantar na literatura listas monolíngues de marcadores nas línguas espanhol e português. Essas listas são necessárias, pois não há nenhum *tagger* ou *parser* de português que rotule automaticamente todos os marcadores discursivos.

A lista de marcadores em espanhol foi construída com dados de gramáticas e sites da web. A lista em espanhol foi utilizada para levantar no *corpus* Fapesp, sentenças em espanhol em que ocorriam marcadores discursivos. Como há itens lexicais que podem ser usados tanto como marcadores discursivos quanto em outras funções, as sentenças foram analisadas manualmente para validar a lista de marcadores no *corpus* acadêmico, resultando em um conjunto de 97 marcadores discursivos. A fim de aumentar a abrangência da lista criada, combinamos essa lista com a lista de 119 marcadores identificados por Fernández (2005) em *corpus* jornalístico. Eliminadas as duplicidades, a lista ficou com 160 marcadores do espanhol.

A lista de marcadores em português partiu da lista do projeto Dizer (Pardo e Nunes, 2006) e foi acrescida dos marcadores identificados por Fernández (2005) em *corpus* jornalístico. O Dizer tem 150 marcadores, enquanto Fernández (2005) tem 122, ambas as listas têm 54 marcadores em comum. Eliminadas as duplicidades, a lista ficou com 218 marcadores.

Não poderíamos ter utilizado apenas as listas levantadas por Fernández (2005), pois elas são oriundas de um *corpus* jornalístico e, como pudemos verificar pela interpolação com outras listas, o léxico de marcadores discursivos e a frequência com que ocorrem, é dependente de gênero. Após o levantamento realizado, confirmou-se a conclusão de Fernández (2005) sobre o fato de que

o número de marcadores em português é maior que o número de marcadores em espanhol.

As lista em espanhol foi utilizada no Passo 3 como entrada do algoritmo para extrair do *córpus* paralelo seus equivalentes em português. Já a lista em português foi utilizada no Passo 4 para filtrar os equivalentes extraídos do *córpus*, pois muitas das equivalências podem não ser marcadores discursivos.

Passo 3: Extração dos equivalentes do espanhol em português

Para extrair do *córpus* paralelo os equivalentes de tradução em português para os marcadores discursivos em espanhol, foi utilizado o modelo de tradução criado com *toolkit* Moses (Seção 6.1.1.2). O procedimento procura traduções candidatas em português para cada marcador em espanhol que aparece na lista levantada no Passo 2. Por exemplo, para o marcador "*con todo*" foram obtidas as seguintes traduções: "com tudo"; "mas"; "contudo"; "com todo"; "porém"; "no entanto"; "com toda"; "a todo"; "em tudo"; "a tudo".

Passo 4: Filtragem dos equivalentes que são marcadores discursivos

Como o recurso lexical foi desenvolvido para apoiar auxílios à escrita, é importante mostrar para o aprendiz unicamente as traduções em que a expressão em espanhol tenha a função de marcador discursivo. Para isso, os equivalentes de tradução produzidos no Passo 3 foram filtrados, usando como referência a lista de marcadores em português previamente compilada. Das traduções do marcador "*con todo*" mostradas como exemplo no Passo 3, somente quatro são marcadores discursivos em português que constam da lista levantada no Passo 2: "mas", "contudo", "porém" e "no entanto". Consideram-se válidos apenas esses quatro equivalentes, porque temos certeza de que eles são marcadores discursivos. Os demais foram desprezados e poderiam até ser marcadores, mas não se tem meios automáticos de verificar isso. No exemplo citado, nenhum dos equivalentes desprezados é marcador discursivo.

O filtro retorna alguma tradução sempre que entre os equivalentes de tradução candidatas houver um marcador discursivo que conste da lista inicial em português, caso contrário, nenhum item lexical em português é apresentado. Achou-se preferível não apresentar uma sugestão a apresentar uma sugestão errada. Por exemplo, as traduções obtidas para o marcador "*de todas formas*" foram: "de todas formas"; "em todas formas", "dos todas formas"; "da todas formas". Embora "de todas formas" tenha equivalentes em português ("de toda forma" e "de qualquer forma"), eles não apareceram entre os equivalentes de tradução extraídos do *córpus*.

Passo 5: Agrupamento dos marcadores discursivos

Os marcadores discursivos foram agrupados por função discursiva, considerando a lista de marcadores em espanhol e o conjunto de seus respectivos equivalentes de tradução em português, seguindo o método para obter sinônimos (Seção 6.1.2). O agrupamento assume que se vários marcadores se alternam como equivalentes de um marcador em espanhol, provavelmente eles têm a mesma função discursiva, ou seja, são sinônimos. Da mesma forma, se um marcador em português ocorre como equivalente de vários marcadores em espanhol, esses marcadores em espanhol provavelmente têm a mesma função discursiva. A partir das equivalências de tradução entre o espanhol e o português, uma representação matricial foi obtida estimando-se a coocorrência dos marcadores discursivos em português nos diversos conjuntos de equivalentes.

Para avaliar a coocorrência dos marcadores, medidas de associação foram propostas. O problema a ser tratado e as características de cada medida de associação são elementos importantes que influenciam na escolha da medida de associação. A associação entre os marcadores discursivos em português foi calculada utilizando as seguintes medidas de associação: coeficiente Dice, *Pointwise mutual information* e *log-likelihood ratio*. As três medidas citadas mostraram resultados quase idênticos, porém o coeficiente Dice já tinha sido empregado nesta pesquisa para avaliar associação entre contextos de palavras na implementação do algoritmo apresentado por Bai et al. (2009).

Para categorizar os marcadores discursivos, uma matriz Y de sinônimos foi criada a partir do cálculo do coeficiente Dice. Na matriz quadrada $n \times n$ as filas e as colunas são os marcadores discursivos em português, sendo n o total de marcadores em português. A diagonal principal da matriz Y é 1, porque indica a associação de um marcador em português com ele mesmo. Para estimar o coeficiente Dice entre os marcadores discursivos, a probabilidade de cada marcador $p(w_p)$ foi calculada, considerando-se a quantidade de vezes que o marcador w_p acontece nas traduções. Por outro lado, a probabilidade conjunta $p(w_{p0}, w_{p1})$ é estimada considerando-se a quantidade de vezes que w_{p0} e w_{p1} acontecem nos mesmos conjuntos de equivalentes de tradução. A equação utilizada para calcular o coeficiente Dice é a mesma apresentada no algoritmo 1.

Para cada marcador em português, há uma lista de marcadores que pertencem aos mesmos conjuntos de equivalentes aos quais o marcador pertence. Às vezes, esta lista é vazia porque o marcador acontece como única tradução para um marcador do espanhol. Nesse último caso, a associação do marcador com algum outro será zero. A matriz Y é a informação que o algoritmo

de agrupamento *AffinityPropagation* (AP) (Frey e Dueck, 2007) emprega para criar o agrupamento dos marcadores discursivos. Foi selecionado o algoritmo *AffinityPropagation* porque ele é indicado quando os dados são esparsos, de forma que a matriz de entrada contenha muitos zeros, como acontece nos experimentos aqui reportados. O algoritmo AP já foi utilizado em aplicações para diversos fins. Por exemplo, Al-Rfou (2012) empregou o algoritmo para agrupar famílias de línguas utilizando a matriz de confusão como uma matriz de similaridade entre as línguas.

AP é um algoritmo que identifica exemplos dentre as instâncias dos dados e forma grupos ao redor dessas instâncias. O algoritmo considera que todas as instâncias são exemplos potenciais e por meio da troca de mensagens entre os dados seleciona as instâncias mais características para realizar o agrupamento. O algoritmo requer unicamente uma matriz de similaridade como entrada, por exemplo, uma instância 1×1 da matriz Y , sendo Y_{ij} o valor da medida que indica quão similares são os exemplos i e j . Os valores da matriz de similaridade podem ser tanto positivos quanto negativos; quanto maior é o valor, mais similares são as instâncias e, caso existam valores $-\infty$, as instâncias envolvidas são consideradas como absolutamente dissimilares. A matriz não precisa ser simétrica (Frey e Dueck, 2007).

Adicionalmente à matriz de similaridade, outro parâmetro importante do algoritmo AP é chamado de *preference*, que é interpretado como a tendência que as instâncias têm de se converterem em exemplos. O parâmetro *preference* determina quão granulado será o agrupamento e, conseqüentemente, quantos grupos serão criados. Desta forma, uma das vantagens do algoritmo AP é que não é necessário estabelecer um número de grupos, como acontece em diversos algoritmos de agrupamentos. Segundo Frey e Dueck (2007), um valor adequado para inicializar o parâmetro poderia ser a mediana dos valores de similaridades dentro da matriz, o que resultaria em um número moderado de grupos, mas também poderia ser o valor mínimo de similaridade, resultando no número mínimo de grupos. Em nossos experimentos foi empregada a implementação do algoritmo AP disponível em *Python* e foi mantido o valor predefinido da *preference*, que neste caso é a mediana dos valores da matriz de similaridade.

Após detalhar o modo de compilar o léxico de marcadores discursivos, a seguir são apresentados alguns dos resultados obtidos com o método proposto.

Atualmente, o léxico bilíngue de marcadores discursivos é composto por 133 marcadores em espanhol e 130 marcadores em português. O total de

marcadores selecionados na segunda etapa de geração do recurso diminuiu porque alguns dos marcadores em português não apareceram como equivalentes de nenhum marcador em espanhol ou não aconteceram no corpus Fapesp.

Como resultado do processo de tradução, obtém-se, para cada marcador em espanhol, um marcador ou uma lista de marcadores em português. A seguir apresenta-se a saída do procedimento para os marcadores discursivos em espanhol “*así que*”, “*además*”, “*de este modo*”.

“*así que*” → “assim que” “*además*” → “além disso”; “aliás”; “ainda” “*de este modo*” → “desse modo”; “assim”; “dessa forma”; “dessa maneira”

Os exemplos anteriores mostram o sucesso do método proposto para construir o léxico bilíngue. Além disso, por meio dos exemplos são analisadas algumas das estratégias válidas que podem ser adotadas para estabelecer equivalências entre os marcadores discursivos das duas línguas. Estas informações são mostradas no recurso e podem ser utilizadas como estratégias para detectar e/ou corrigir erros. No primeiro exemplo, com uma tradução literal do espanhol para o português, foi gerado um marcador correto em português. Já no segundo exemplo, a tradução literal geraria o marcador “ademais” em português. Embora o marcador “ademais” não represente um erro do português, ele quase não é usado pelos nativos, pois caiu em desuso. Por último, no terceiro exemplo, para se obter as equivalências corretas foi necessário empregar adequadamente as contrações do português (preposição e pronome demonstrativo neste caso).

A partir destes primeiros resultados, o recurso está pronto para oferecer o auxílio de equivalências entre o espanhol e o português. É importante ressaltar que o léxico só inclui marcadores discursivos nas duas línguas, então não há perigo de apresentar uma equivalência que não seja marcador discursivo, coisa que poderia acontecer caso um método de tradução fosse utilizado pelos hispano falantes.

Utilizando-se as equivalências dos marcadores em espanhol e português, os marcadores das duas línguas foram categorizados. Essa categorização é apresentada no Apêndice F e representa uma alternativa de apresentar sinônimos a os hispano falantes, contribuindo para melhorar a diversidade lexical nos textos. Para categorizar os 130 marcadores discursivos em português que aconteceram no corpus Fapesp foi calculada uma matriz de similaridade de 130×130 , que contém a estimativa do coeficiente Dice entre os marcadores.

Inicialmente, os marcadores discursivos em português foram agrupados em

29 grupos. Desses 29 grupos, foram eliminados 3 grupos. O primeiro grupo eliminado era composto por 31 marcadores discursivos. Nele foram incluídos marcadores que aconteceram como uma única tradução e, como consequência, não estavam relacionados com nenhum outro marcador discursivo. O segundo grupo excluído era formado pelos marcadores “aí” e “por favor”, que não são empregados na escrita acadêmica e por isso saíram da análise. Por último, o grupo formado pelos marcadores “ou” e “e”, também foi eliminado, pois estes são itens muito ambíguos; não sendo considerados pelos autores como marcadores discursivos. Desta forma, dos 29 grupos criados automaticamente, 26 grupos compõem o recurso atualmente. Com relação ao agrupamento, é importante apontar que nos 29 agrupamentos criados não houve ambiguidades, isto é, um mesmo marcador não aparece em mais de um grupo ao mesmo tempo. Depois de se criar os grupos em português, os grupos em espanhol foram obtidos por casamento com os grupos em português.

Na Seção 6.2.3 são apresentados experimentos desenvolvidos para detectar e corrigir erros de marcadores discursivos.

Detecção e Correção de Erros de Marcadores Discursivos

Como comentado na Seção 5.3.3, os hispano falantes cometem desvios da língua que muitas vezes são produzidos por interferências de sua língua nativa, o espanhol. Em alguns casos, usam marcadores discursivos do espanhol que são itens lexicais que existem em português com outra função sintática, os quais podem ser considerados falsos cognatos entre as duas línguas. Em outros casos, em uma tentativa de gerar marcadores discursivos em português, os aprendizes produzem expressões que não existem em nenhuma das duas línguas. Em qualquer dos casos, a detecção automática de erros de marcadores discursivos não é uma tarefa trivial, porque além de identificar o erro é necessário reconhecer que o erro é de marcador discursivo.

Em uma análise manual do corpus Fapesp foram observados padrões no uso dos marcadores discursivos que nos mostram que grande parte dos marcadores discursivos pode ocorrer: a) no início de uma das orações relacionadas ao marcador, seguido de vírgula; b) entre as duas orações que relaciona, precedido de vírgula ou c) no meio de uma das orações que relaciona, entre vírgulas.

- (i) Ele estudou, no entanto não foi aprovado no exame.
- (ii) Ele estudou. No entanto, não foi aprovado no exame.

- (iii) Ele estudou. Não foi, no entanto, aprovado no exame.

Uma exceção são os marcadores compostos por conjunções e locuções concessivas, nas orações subordinadas adverbiais concessivas:

- (iv) Embora ele tenha estudado, não foi aprovado no exame.
- (v) Ele não foi aprovado no exame, embora tenha estudado.

Foram usados os padrões detectados para identificar automaticamente os marcadores discursivos nos textos. Para simplificar a tarefa, decidiu-se usar na identificação apenas as características comuns aos dois tipos de marcadores: por exemplo, para ser considerado marcador, o item lexical tem que estar no início de sentença ou ser precedido de vírgula. Essas duas características foram inseridas em um processo de geração de erros artificiais que será explicado a seguir. Em textos de aprendizes, contudo, essas características podem não ser, eventualmente, suficientes para se identificar erros caso eles não façam uso adequado da vírgula. Aliás, como observamos que em textos bem escritos existem regras para o uso da vírgula junto a marcadores discursivos, pretendemos em trabalhos futuros dar apoio ao aprendiz também nesse aspecto.

1. Geração Artificial de Erros

Como comentado na Seção 5.3.3, a proximidade entre o espanhol e o português torna o problema de interferência da língua materna ainda maior, permitindo que os hispano falantes tentem “adivinhar” palavras ou expressões equivalentes na segunda língua. O exemplo a seguir mostra como o aprendiz se apoiou na tradução literal do marcador discursivo em espanhol “*por una parte*” para gerar, de forma errada, a expressão “*por uma parte*”, que no Exemplo 3 está fazendo função de marcador discursivo.

Exemplo 3: A aplicação deste modelo permitirá, *por uma parte*/(**por um lado**), efetuar o reconhecimento, descrição e medição das variáveis qualitativas ...

Para detectar e corrigir automaticamente problemas causados pela interferência da língua nativa é necessário reconhecer a interlíngua que o aprendiz produz e sua natureza. Muitas vezes, os erros como os do exemplo anterior não são detectados pelos sistemas dedicados a identificar erros (corretores gramaticais) e outras vezes o sistema detecta o erro, mas não consegue oferecer a correção adequada.

Seguindo esta linha, para gerar erros artificiais de marcadores discursivos o procedimento apresentado na Seção 6.1.3 foi adotado, gerando o erro e mantendo o controle da palavra ou expressão correta em português. Assim, é possível oferecer a correção do erro ou pelo menos sugestões bem sucedidas. O léxico de erros de marcadores discursivos também foi integrado no léxico "Spanish-native-like errors", sendo composto das seguintes três formas:

- a forma em espanhol,
- a interlíngua (forma errada) e
- a forma em português.

Para o Exemplo 3, a sugestão seria: “Você quis dizer "por una parte"?” “A equivalência pode ser “por um lado”. A geração do léxico de erros artificiais usa o mesmo raciocínio dos aprendizes, observado no corpus Espanhol-Acadêmico-Br.

O procedimento proposto para gerar erros artificiais é apresentado na Figura 6.10. A geração de erros artificiais aparece como alternativa para detectar erros quando a anotação de erros no corpus não é suficientemente grande como também para guiar um processo de detecção automática.

Figura 6.10: Procedimento para gerar erros artificiais de marcadores discursivos.

Como mostrado na Figura 6.10, o método para gerar erros artificiais recebe como entrada o léxico bilíngue de marcadores discursivos. Diversas heurísticas (descritas a seguir) são aplicadas aos marcadores discursivos em espanhol, incluídos no léxico bilíngue, com a finalidade de gerar candidatos a erros

em português. Como a geração artificial de erros foi acoplada ao método apresentado para compilar o léxico bilíngue, é possível gerar o erro e obter a sua correção. Os erros artificiais são gerados segundo as seguintes heurísticas:

1. Língua fonte: Palavras ou expressões em espanhol são os primeiros candidatos a erros. Quando não sabem uma palavra em português, os aprendizes “arriscam” usar palavras do espanhol porque essa estratégia tem grande probabilidade de acerto devido à similaridade entre as línguas. **Exemplo:** “Se observou também um incremento no número de artigos escritos por autores que trabalham em América do norte como por autores que trabalham em outras regiões; *sin embargo*(**mas; contudo; porém; não obstante; no entanto**), os autores norte americanos publicam a maioria dos trabalhos.”

2. Língua fonte + Padrão vírgula: Às vezes o item lexical que constitui o marcador em espanhol existe em português, mas não como marcador discursivo. Em função disso, acrescentamos a vírgula no erro gerado a fim de desfazer a ambiguidade. **Exemplo:** “Têm a vantagem de por em evidência sintomas prematuros do dano causado pelo contaminante e, *por tanto*,/(**portanto**) podem ser utilizados como sinais de alarme ante a presença de contaminantes químicos.”

3.Tradução literal: Cada palavra que compõe a expressão em espanhol é traduzida literalmente por meio do modelo de tradução utilizado para obter as equivalências entre o espanhol e o português. **Exemplo:** “A aplicação deste modelo permitirá, *por uma parte*/(**por um lado**), efetuar o reconhecimento, descrição e medição das variáveis qualitativas ...”

4. Tradução literal + Padrão vírgula: Uma expressão traduzida literalmente pode ser aceita em português mas não como marcador discursivo, por isso, a vírgula é acrescentada ao erro para desambiguar a opção do aprendiz de usar a expressão como marcador discursivo. **Exemplo:** “*em concreto*/(**Na verdade**), aplicamos categorias de análises coerentes com tal perspectiva ao estudo dos processos de construção de conteúdos

Devido às similaridades entre o português e o espanhol, é possível gerar erros tendo como entrada o espanhol. Porém, existe uma alta probabilidade de que o candidato a erro exista em português. Para verificar se os erros gerados são realmente erros em português, utiliza-se o modelo de língua, apresentado na Seção 5.1.2, que valida a existência ou não das expressões geradas. Quando a palavra ou frase não pertence ao modelo de língua, ela é inserida no léxico de erros automaticamente. O léxico de erros é a saída principal do procedimento apresentado na Figura 6.10.

Embora os erros gerados de forma artificial reflitam os erros identificados no *córpus* de aprendizes, não é possível prever todos os erros que um aprendiz pode produzir. O fato anterior deve determinar uma baixa cobertura do recurso criado, pois só é possível identificar os erros dos marcadores incluídos no recurso. Outra deficiência do método, que também deve afetar sua cobertura, está relacionada ao fato de o *córpus* paralelo ser um recurso limitado e, portanto, quando um marcador específico não estiver incluído no *córpus*, o método não vai conseguir obter a equivalência em português e, caso detecte um erro, não poderá mostrar sua correção. Uma alternativa para solucionar o problema anterior é aumentar o *córpus* paralelo, porém são poucos os *córpus* paralelos disponíveis, sendo menores ainda os do gênero acadêmico para as línguas português e espanhol. Para obter, automaticamente, textos paralelos podem ser utilizados os métodos propostos por Sridhar et al. (2011); Barbosa et al. (2012). Com a finalidade de aumentar a cobertura da tarefa de detectar erros foi implementado um algoritmo que, utilizando um modelo de língua, identifica erros que são corrigidos por meio de similaridade ortográfica. O método proposto é explicado a seguir.

2. Modelo de Língua e Similaridade Ortográfica para Detectar e Corrigir Erros

O método apresentado na Figura 6.11 tem a mesma finalidade do método da Figura 6.10: detectar e corrigir erros de marcadores discursivos. Os erros detectados pelo método da Figura 6.10, porém, não são gerados de forma artificial, ao contrário, são extraídos automaticamente dos textos dos aprendizes. Como apresentado na Figura 6.11, em uma primeira etapa o método combina um conjunto de heurísticas e um modelo de língua para detectar expressões não adequadas. Em um segundo momento, o método procura correções para as expressões levantadas usando um algoritmo de similaridade ortográfica. Finalmente, o método é avaliado na tarefa de detectar e corrigir erros.

Os textos dos aprendizes são analisados com a finalidade de se extrair *n*-gramas que representem erros de marcadores discursivos. Tipicamente, os *n*-gramas são sequências contínuas de palavras que têm tamanhos diversos. Os *n*-gramas não reconhecidos pelo modelo de língua e apresentado na Seção 5.1.2 foram levantados como candidatos a erros. A maioria dos *n*-gramas candidatos a erros não se constitui erro de marcador discursivo, foco destes experimentos. Além disso, como o modelo de língua ainda é um recurso pequeno, que não contém todas as variações permitidas pela língua portuguesa, alguns dos *n*-gramas selecionados como candidatos a erros podem represen-

Figura 6.11: Procedimento para detectar e corrigir erros de marcadores discursivos por similaridade entre expressões.

tar falsos positivos. Em outras palavras, eles existem em português, mas não aparecem no modelo de língua.

Como alternativa para diminuir o número de falsos positivos na detecção de erros, incluímos na análise o léxico do NILC¹⁸ (Nunes et al., 1996), que contém palavras e expressões. O léxico do NILC foi utilizado para excluir da análise todos os n-grams que não aparecem no modelo de língua e que existem na língua portuguesa (pelo menos existem no léxico). Após aplicar o filtro anterior, um segundo filtro é considerado, desta vez com a finalidade de manter unicamente os n-gramas com alta probabilidade de conter erros de marcadores discursivos. Para isso, foram considerados os n-gramas que aconteceram no começo ou no meio das sentenças e que estejam anteceditos ou precedidos de vírgula ou entre vírgulas. Como comentado anteriormente, há fortes indícios de que esses n-gramas contenham erros relacionados a marcadores discursivos.

Após selecionar os n-gramas com erros no texto dos aprendizes, utilizou-se um algoritmo baseado na similaridade das expressões para corrigir esses erros. Por exemplo, nos textos dos aprendizes aparece a expressão “*a pesar disto*”, que é um erro de junção/separação de palavras, e sua correção seria “*apesar disso*”, que é uma expressão similar ao erro. Para isso, estima-se a distância de edição d com o algoritmo *Levenshtein* (Levenshtein, 1966; Damerau, 1964a) entre o n-grama candidato a erro n e S , em que S é um conjunto

¹⁸Disponível em: <http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html>

de palavras e expressões extraídas do léxico do NILC. O conjunto S foi criado considerando todas as entradas do léxico do NILC com a etiqueta morfosintática “LOC”. A etiqueta “LOC” é utilizada no léxico do NILC para denotar marcadores discursivos, conectivos, expressões fixas e advérbios. A seleção do conjunto S é realizada com a finalidade de diminuir o espaço de busca das correções dos erros e, conseqüentemente, aumentar as chances de detectar a sugestão correta. As sugestões obtidas com o algoritmo de similaridade são ordenadas segundo a distância de edição d e foram consideradas correções candidatas aquelas expressões com menos de quatro caracteres de diferença.

Embora tenha havido uma preocupação em se detectar unicamente erros de marcadores discursivos na implementação deste método, é natural que alguns outros tipos de erros sejam levantados e até corrigidos incidentalmente, devido ao fato de que outras expressões, além dos marcadores discursivos, foram mantidas no subconjunto S do léxico do NILC. Por isso, na avaliação do método, consideramos sua capacidade de detectar e corrigir erros de marcadores discursivos e outros tipos de erros. Portanto, outra diferença entre este método e o método baseado em erros artificiais é que enquanto o último método identifica e corrige exclusivamente erros de marcadores discursivos, este pode identificar e corrigir outros erros de expressões.

3. Resultados e Avaliação da Detecção e Correção de Erros

A avaliação dos experimentos propostos para detectar e corrigir erros de marcadores discursivos foi realizada por meio de uma verificação manual dos erros levantados e suas correções. Essa estratégia foi necessária devido à ausência de um recurso com erros previamente anotados e corrigidos que pudesse ser utilizado como referência para avaliar o desempenho dos métodos. A abordagem anterior foi utilizada por Gamon et al. (2008), que também não tinha um corpus com erros anotados, e que após a verificação da saída do sistema disponibilizou um grande conjunto de exemplos.

As Figuras 6.12a e 6.12b mostram o total de erros detectados e corrigidos em cada um dos métodos apresentados para detectar e corrigir erros de marcadores discursivos. Como observado, o número de erros identificados pelo método que combina o modelo de língua com a similaridade ortográfica é maior que o número de erros detectados pelo recurso de erros artificiais. O método que utiliza similaridade sempre oferece uma sugestão para o erro detectado, já o léxico com erros artificiais às vezes não oferece nenhuma correção. O fato anterior acontece porque para o marcador em espanhol foi gerado

um erro artificial, mas não sua tradução correta, provavelmente porque o marcador discursivo não aconteceu no corpus paralelo ou porque nas equivalências obtidas não apareceu nenhum marcador discursivo da lista levantada no Passo 2 do algoritmo para compilar o léxico bilíngue de marcadores discursivos. Contudo, quando o léxico de erros artificiais retorna uma correção, há garantias de que ela seja correta, pois o método mantém a forma errada da palavra e sua correção.

Figura 6.12: Total de erros detectados e corrigidos pelos métodos propostos para detectar e corrigir erros de marcadores discursivos.

Como comentado, o método que combina o modelo de língua com a similaridade ortográfica, além de detectar erros de marcadores discursivos, levantou outros tipos de erros, por exemplo, erros de expressões fixas como “*de acordo a*” quando deveria ter sido utilizado “*de acordo com*”. Além disso, detectou com bastante frequência erros de advérbios terminados em “-mente” que apareceram com acento. Este último tipo de erro representa uma interferência do espanhol, pois em espanhol os advérbios terminados em “-mente” levam acento sempre que a palavra leve acento, já em português não acontece a mesma coisa. O erro dos advérbios é detectado e corrigido pelos corretores gramaticais mais utilizados na língua portuguesa, no entanto o erro na expressão “*de acordo a*” não é identificado.

O procedimento proposto para verificar a saída do sistema e, consequentemente, para avaliar os métodos que detectam e corrigem erros, é apresentado na Figura 6.13. A avaliação dos algoritmos foi realizada nos diferentes níveis que aparecem na Figura 6.13. Sabe-se que o léxico de erros artificiais só identifica e corrige erros de marcadores discursivos. Desta forma, nas sentenças levantadas com este método, verifica-se o desempenho do método na detecção correta de erros ou verdadeiros positivos (nível 1 da Figura 6.13).

Adicionalmente, o método é avaliado na tarefa de corrigir erros de marcadores discursivos sempre que o erro tenha sido identificado corretamente; caso contrário, avalia-se a capacidade do método de falhar na correção/detecção, após detectar um falso positivo (nível 3 da Figura 6.13). Os resultados da avaliação do léxico de erros artificiais é apresentado na Tabela 6.11.

Figura 6.13: Procedimento para verificar manualmente a saída do sistema.

Por outro lado, o método baseado no modelo de língua e na similaridade ortográfica detecta outros tipos de erros, além de erros de marcadores discursivos, por isso, este método foi avaliado considerando todos os níveis da Figura 6.13. Primeiramente, todas as sentenças levantadas pelo método são utilizadas para avaliar a capacidade do método em identificar corretamente erros (nível 1 da Figura 6.13). Em uma segunda etapa, aquelas sentenças que realmente continham um erro foram utilizadas para avaliar a capacidade do método em identificar erros de marcadores discursivos e outros tipos de erros (nível 2 da Figura 6.13). Finalmente, o método é avaliado na tarefa de oferecer sugestões de erros de marcadores discursivos e de outros tipos de erros, assim como sua capacidade de não estragar as sentenças com erros detectados incorretamente ou falsos positivos (nível 3 da Figura 6.13). Os resultados da avaliação deste último método são mostrados na Tabela 6.12.

A detecção e correção de erros foi avaliada por meio da precisão dos métodos. A precisão foi calculada segundo a Equação 6.3:

$$\text{precisão} = \frac{\text{verdadeiro positivo}}{\text{verdadeiro positivo} + \text{falso positivo}} \quad (6.3)$$

Sendo que os verdadeiros positivos são os erros identificados corretamente como erros e os falsos positivos são exemplos corretos que foram identificados como erros. A mesma definição é empregada para a correção dos erros, sendo

verdadeiros positivos as correções avaliadas como corretas por humanos e falsos positivos, as correções avaliadas como incorretas.

O cálculo da precisão depende da verificação do sistema, que foi realizada por dois humanos treinados. Os humanos avaliaram as saídas dos métodos e a concordância entre os anotadores foi obtida por meio do cálculo do *Kappa*. Valores de concordância entre os anotadores indicam a dificuldade da tarefa para os humanos e é um parâmetro adicional para avaliar a saída dos métodos. O *Kappa* obtido na tarefa de identificar erros (nível 1 da Figura 6.13)) foi de 0,75, o que indica que os anotadores concordaram na anotação de 88% dos casos. O *Kappa* na tarefa de identificar se o erro é ou não de marcador discursivo foi calculado só para sentenças levantadas pelo método que utiliza o modelo de língua. Neste caso, o valor do *Kappa* foi de 0,67, o que indica que os anotadores concordaram em 86% dos casos. Embora o valor do *Kappa* indique uma alta concordância entre os anotadores, as discordâncias mostram que, mesmo para humanos, é difícil reconhecer erros de marcadores discursivos, talvez porque eles tenham dificuldade em reconhecer os itens que compõem essa categoria. Por exemplo, durante a anotação foi observado que para um anotador a expressão “de acordo com” é marcador discursivo, mas para o outro anotador não é. O segundo anotador está correto se se considera o léxico de marcadores discursivos levantado na literatura e esse exemplo de discordância aconteceu várias vezes na avaliação, impactando o *Kappa*.

Usando como referência a verificação manual, foi calculada a precisão dos métodos na detecção e correção de erros. Apesar de ser alta a concordância entre os anotadores, a precisão só foi calculada para os exemplos em que houve máxima concordância.

Na Tabela 6.11 aparecem os resultados da avaliação do léxico de erros artificiais na tarefa de detectar e corrigir erros. O método apresentou uma precisão de 83% na detecção de erros de marcadores discursivos. Os falsos positivos, que representam 17% dos erros detectados, foram motivados pelo fato de que o erro artificial gerado existe em português, mas não apareceu no corpus utilizado para construir o modelo de língua que verificava os erros gerados artificialmente para detectar erros de aprendizes. Por outro lado, o método teve uma precisão de 86% na tarefa de oferecer sugestões e, além disso, nenhuma das sugestões oferecidas piorou o texto quando um falso positivo foi detectado.

Os resultados do método baseado no modelo de língua e na similaridade ortográfica mostram que houve bastantes falsos positivos na detecção de er-

Tabela 6.11: Precisão na detecção e correção de erros com o léxico de erros artificiais.

Léxico de Erros Artificiais		
Nível 1	(1) Detecta erros de marcadores discursivos	83%
Nível 3	(2) Corrige erros de marcadores discursivos	86%
	(3) Sugestão piora o texto após detectar falsos positivos	0%

ros, como mostrado na Tabela 6.12. Dos erros detectados corretamente 39% deles eram erros de marcadores discursivos, enquanto 16% erram outros tipos de erros. Porém, o método teve uma precisão de 89% na correção de erros de marcadores discursivos, mostrando que muitos dos erros ocorrem porque os hispano falantes tentam gerar expressões a partir da similaridade entre as duas línguas. Além disso, o método teve uma precisão de 25% na correção de outros tipos de erros, diferentes dos marcadores discursivos. Embora a precisão anterior seja baixa, ela mostra que o método consegue tratar outros tipos de erros, tais como erros de ortografia e de preposições. A maior deficiência do método está na detecção de erros, e como consequência, na capacidade do método de piorar as sentenças em que erros foram detectados de forma incorreta. Neste último caso, o método consegue uma precisão de 83%. No entanto, a maioria dos falsos positivos contém nomes próprios que não aparecem no modelo de língua nem no léxico do NILC; para reduzir o número de falsos positivos uma alternativa seria incluir uma ferramenta que automaticamente detectasse entidades nomeadas¹⁹ (Cardoso, 2012). Essa é uma abordagem que será incluída em trabalhos futuros.

Tabela 6.12: Precisão na detecção e correção de erros com o método baseado no modelo de língua e na similaridade ortográfica.

Método Baseado em Modelo de Língua e Similaridade Ortográfica		
Nível 1	(1) Detecta Erros	55%
Nível 2	(2) Detecta erros de marcadores discursivos	39%
	(3) Detecta outros tipos de erros	16%
Nível 3	(4) Corrige erros de marcadores discursivos	89%
	(5) Corrige outros tipos de erros	25%
	(6) Sugestão piora o texto após detectar falsos positivos	83%

O recurso de erros artificiais teve maior precisão na detecção de erros se comparado com o método que identificou os erros usando um modelo de língua. Entretanto, o léxico de erros artificiais só detecta os erros que aparecem

¹⁹Uma possível ferramenta seria o REMBRANDT, disponível em <http://xldb.fc.ul.pt/wiki/Rembrandt>

no recurso, enquanto que o método baseado no modelo de língua tem maior abrangência na diversidade de erros detectados. Para melhorar a precisão na detecção e correção de erros do léxico de erros artificiais, uma alternativa seria aumentar o tamanho do cópulo paralelo, o que aumentaria a probabilidade de que os marcadores discursivos em espanhol e em português aparecessem no cópulo. Por outro lado, ambos os métodos são bons na correção do erro detectado: o léxico de erros artificiais só não ofereceu sugestão se o marcador não aparecia no cópulo utilizado para obter as equivalências entre o espanhol e o português e o outro método só errou na sugestão quando não era possível identificar o erro por meio de similaridade ortográfica.

O método baseado em modelo de língua e na similaridade ortográfica é mais factível que o método baseado em léxico de erros artificiais, pois o primeiro detecta erros que realmente ocorrem no cópulo de aprendizes e o outro prevê possíveis erros que talvez nunca aconteçam. Além disso, para melhorar o léxico de erros artificiais é necessário adicionar mais textos ao cópulo paralelo, tarefa que sempre é complexa, pois textos paralelos espanhol-português do gênero acadêmico não são encontrados facilmente em grande quantidade. Para aprimorar o método baseado no modelo de língua, uma alternativa seria construir um modelo de língua maior, sendo mais viável obter textos em uma língua só. Contudo, o léxico de erros artificiais foi compilado focando em erros de hispano falantes aprendizes de português e, desta forma, ele é um recurso mais controlado que mostra as principais dificuldades desses aprendizes na produção de marcadores discursivos em português. Adicionalmente, as sugestões oferecidas pelo recurso de erros artificiais também são mais controladas.

6.2.4 Expressões Formulaicas

O léxico bilíngue de expressões formulaicas foi construído para apoiar os primeiro e o terceiro auxílios à escrita apresentados no começo deste Capítulo. Dessa forma, o léxico mostra as equivalências em português de expressões formulaicas em espanhol. Além disso, utilizando as entradas em espanhol do léxico foi aplicado o procedimento de geração de erros artificiais para auxiliar à tarefa de detectar e corrigir erros de expressões formulaicas, produzidos pela interferência do espanhol.

Como comentado na Seção 3.3.4, as expressões formulaicas são sequências de palavras, contínuas ou descontínuas, que são ou parecem ser frases pré-fabricadas. Em geral, elas expressam um conceito e contribuem com a

fluidez do texto. Devido à diversidade destas expressões, a extração automática delas não é tarefa trivial, pois diversos padrões morfosintáticos as descrevem e, além disso, não há regularidade na frequência com que elas acontecem nos textos. Algumas das medidas que melhor as descrevem são medidas de associação que avaliam a coocorrência das palavras dentro das expressões, porém, isso também não é uma regra, pois algumas expressões formulaicas são fixas, mas outras não. Adicionalmente, algumas expressões formulaicas são dependentes do gênero, algumas que são empregadas com frequência em uma área específica de pesquisa não são utilizadas em outras.

Na Figura 6.14, mostra-se que a construção do léxico bilíngue de expressões formulaicas é separada em seis passos. O primeiro passo consiste em selecionar *córpus* que, utilizado como base de conhecimento, permite extrair o léxico bilíngue de expressões. O segundo passo do algoritmo concentra-se na extração automática da lista de expressões em espanhol. Devido à dificuldade em determinar se uma expressão é formulaica ou não, no terceiro passo do método propusemos uma verificação manual da sua saída. No quarto passo, as equivalentes em português das expressões em espanhol são extraídas do *córpus* paralelo, por meio do algoritmo de tradução estatístico. No quinto e sexto passo do algoritmo o léxico bilíngue compilado é utilizado para gerar erros artificiais.

Passo 1: Pré-processamento

Esta etapa consiste em preparar o *córpus* paralelo Fapesp para usá-lo como base de conhecimento para extrair os pares de expressões formulaicas em espanhol e português. Para isso, o método de tradução estatística apresentado na Seção 6.1.1.2 foi utilizado e o *córpus* foi tokenizado, sentenciado e alinhado segundo o procedimento especificado nessa seção.

Para compilar uma lista em espanhol de expressões formulaicas foram utilizados os textos em espanhol de dois *córpus*: o *córpus* paralelo Fapesp e o *córpus* acadêmico em espanhol, apresentado na Seção 5.1.2. Ambos os *córpus* foram utilizados porque espera-se extrair diferentes expressões de cada um.

Passo 2: Extrair Expressões Formulaicas

Como comentado na literatura, a extração automática de expressões formulaicas não é uma tarefa trivial, devido à diversidade de formas e tipos destas expressões. Diversos estudos têm sido propostos para analisar as expressões formulaicas, sendo diversos os métodos automáticos utilizados. A maioria dos trabalhos consultados extrai as expressões dos *córpus* combinando

Figura 6.14: Procedimento para gerar o léxico bilíngue de expressões formulaicas.

análises de frequências com medidas de associação mas, após a extração, comumente verificam a validade das expressões extraídas para determinar se realmente são expressões formulaicas.

Para extrair automaticamente as expressões formulaicas em espanhol foram analisadas várias alternativas. Primeiramente, foi estudado o emprego da ferramenta *JBow* (Seção 3.3.4), pois é uma ferramenta baseada em um modelo estatístico e, como consequência, é independente de língua, por isso podia ser utilizada em um corpus de língua espanhola. Outra motivação para o emprego da ferramenta *JBow* na tarefa de extração de expressões foi a possibilidade de se obter as expressões em espanhol categorizadas segundo as categorias retórica do resumo. Porém, para isso seria necessário anotar manualmente um conjunto de treinamento, em que as sentenças estivessem classificadas, segundo sua função retórica. Essa anotação manual é uma tarefa complexa que fogia dos interesses desta pesquisa. A partir disso, pensou-se em levantar uma lista de expressões formulaicas sem estar categorizadas nas funções retóricas, no entanto *JBow* gerava uma lista com todos os n-gramas dentro do corpus sendo necessário analisar manualmente toda a lista gerada. O fato da

ferramenta gerar uma extensa lista foi o principal motivo para desistir dessa alternativa para se extrair as expressões formulaicas em espanhol.

Outra ferramenta que se mostrou atrativa para a tarefa de extração das expressões formulaicas em espanhol foi o *MWEToolkit*. Para usá-la, é necessário criar padrões que caracteriam as expressões formulaicas de diferentes tipos. Como comentado na revisão da literatura, obter padrões para estas expressões não é uma tarefa simples. Da mesma forma que acontecia com a ferramenta *JBow*, a saída de *MWEToolkit* era uma lista com expressões e várias medidas de associação calculadas para cada expressão. Considerando esse fato, pensou-se em aplicar algoritmos de classificação ou agrupamentos para tentar selecionar um conjunto em que maior número de expressões formulaicas reais estivesse contido. Porém, neste último caso seria necessário avaliar a correção da lista obtida.

A partir das análises anteriores e considerando-se a revisão da literatura, decidiu-se adotar uma abordagem similar às propostas por Simpson-Vlach e Ellis (2010) e Prentice (2012) para extrair expressões formulaicas em espanhol. Da mesma forma que em Prentice (2012), foram extraídos n-gramas de tamanho três e quatro que aparecem no começo das sentenças. Como em Simpson-Vlach e Ellis (2010) a frequência das n-gramas foi combinada com a informação mútua para ordenar as expressões com maior probabilidade de serem expressões fixas. Usar o começo da sentença como referência para procurar expressões formulaicas é um critério válido que diminui o espaço de busca e aumenta as probabilidades de se obter um n-grama que represente uma expressão formulaica. Porém, podem ser extraídas 3-gramas que se sobrepõem com a lista de 4-gramas.

Do corpus Fapesp foi extraída uma lista de 100.054 3-Gramas e 138.391 4-Gramas, enquanto do corpus acadêmico em espanhol foram obtidas 5.340 3-Gramas e 6.437 4-Gramas. Como as listas de n-gramas foram ordenadas pela frequência e a informação mútua, de cada lista, selecionou-se as 150 expressões melhores ordenadas considerando a informação mútua. No total, foram extraídas 600 expressões como candidatas, 300 de cada corpus, em que 150 eram 3-gramas e 150 eram 4-gramas. Algumas das expressões extraídas do corpus Fapesp foram "*al mismo tiempo*", "*por tal motivo*", "*según el investigador*". Também foram obtidos marcadores discursivos, alguns deles não estavam incluídos no léxico bilíngue de marcadores discursivos, porque não apareciam na lista compilada previamente. Já para o corpus de resumos acadêmicos foram obtidas expressões, como: "*en este trabajo*", "*los resultados*

indican”, que não aconteceram no corpus Fapesp.

Das 600 expressões candidatas, somente, 30 3-gramas e 60 4-gramas aconteceram em ambas as listas. Isso comprova a importância de se extrair as expressões de corpus diferentes. Além disso, 191 3-gramas eram sub-sequências das 4-gramas. A grande maioria das sobreposições aconteceu pela inserção de símbolos de pontuação, por exemplo, a vírgula. Porém, foram observadas expressões que eram partes de expressões maiores. Por exemplo, foi selecionado o 3-grama "*al fin y*" e, além disso, foi selecionado o 4-grama "*al fin y al*". Ambas as sequências formam parte da expressão fixa "*al fin y al cabo*", que significa "afinal" em português. "*Al fin y al cabo*" é uma expressão fixa recorrente em espanhol, mas que não foi considerada nesta análise, pois tem tamanho 5. As expressões de tamanho 5 não foram consideradas na análise para evitar maior sobreposição entre as sub-sequências, lembrando que foram extraídas expressões que aconteceram no começo das sentenças. Eliminando-se as duplicações e as sobreposições, a lista de expressões formulaicas candidatas ficou com 319 expressões.

No conjunto de 319 expressões foram observadas expressões muito frequentes, como "*y eso es*", que não é uma expressão fixa. Também detectaram-se expressões menos frequentes, tais como "*en estados unidos*" que são privilegiadas pelo cálculo da informação mútua, por conterem entidades nomeadas que são raras em textos acadêmicos. Como esses fenômenos são comuns na extração de expressões fixas e com a finalidade de se obter uma lista de expressões formulaicas que auxilie a escrita em português dos hispano falantes, foi proposta uma verificação manual da lista de expressões candidatas, como mostrado no Passo 3 do algoritmo.

Passo 3: Julgamento Humano

As 319 expressões incluídas no Passo 2 do algoritmo foram revisadas por uma linguista treinada, com a finalidade de indicar qual delas eram realmente expressões formulaicas da língua espanhola. Na verificação manual foram apontadas 264 expressões formulaicas; destas 146 eram 3-gramas e 118 4-gramas. Desta forma, 83% das expressões candidatas eram realmente expressões formulaicas. Além disso, o anotador humano ofereceu diferentes traduções para cada uma das expressões candidatas. Essa tradução manual foi utilizada para avaliar os algoritmos de equivalências, apresentados na Seção 6.1.1. Algumas das expressões que se mantiveram na lista foram: "*en los últimos años*", "*uno de los objetivos*", "*de acuerdo con*", "*a partir de entonces*", etc..

Depois de criada a lista de expressões formulaicas em espanhol, para cada elemento da lista foi obtido seu equivalente em português, como mostrado no Passo 4.

Passo 4: Extrair Equivalente do ES em PT

Para se extrair do corpus paralelo Fapesp os equivalentes de tradução em português para as expressões formulaicas em espanhol, foi utilizado o modelo de tradução criado com o *toolkit* Moses e apresentado na Seção 6.1.1.2. O procedimento procura traduções candidatas em português para cada expressão em espanhol que aparece na lista do Passo 3. Por exemplo, a seguir aparecem algumas expressões formulaicas em espanhol e sua tradução em português.

- *al fin de cuentas*: afinal de contas
- *al mismo tiempo*: ao mesmo tempo
- *los investigadores creen que*: os pesquisadores acreditam que

Os exemplos mostram o sucesso do método de equivalência proposto e, além disso, por meio das equivalências é possível mostrar para o aprendiz que nem sempre a tradução literal entre o espanhol e o português está correta. Por exemplo, o hispano falante poderia ter escrito "*a fim de contas*" e ter gerado um erro. Por outro lado, a tradução literal era possível no segundo exemplo. A última equivalência mostra como há uma preferência por empregar "pesquisadores" e "acreditam", embora "investigadores" e "crêem" existam.

Um problema que afetou o desempenho do algoritmo de tradução estatística foi o fato de que algumas das expressões formulaicas, incluídas na lista em espanhol, não aconteciam no corpus paralelo, porque foram extraídas do corpus acadêmico em espanhol. Esse fato já foi comentado na Seção 6.1.1.4, quando apresentados os algoritmos de equivalências e o processo de avaliação mostrou que o método estatístico era menos substituível à ausência de palavras e expressões no corpus, porque nesses casos o algoritmo realizava uma tradução literal, que às vezes é bem sucedida devido à similaridade entre o espanhol e o português.

Passo 5 e 6: Gerar Erros Artificiais e Verificar Erros

O procedimento proposto para gerar erros artificiais (Seção 6.1.3) foi utilizado para gerar erros de expressões formulaicas. Esse procedimento foi utilizado e avaliado na geração de erros artificiais de marcadores discursivos (Seção 6.2.3). Utiliza-se como entrada do método o léxico bilíngue de expressões formulaicas para gerar os erros. Na geração de erros de marcadores

discursivos, diversas heurísticas foram propostas, como aparecem na Figura 6.10. Para se gerar erros de expressões formulaicas, somente foi proposta a geração de erros segundo a tradução literal, pois foram os padrões de erros associados a estes itens lexicais que foram observados com mais recorrência no corpus Espanhol-Acadêmico-Br. A saída do procedimento que gera erros de expressões formulaicas está composta de:

- a forma em espanhol,
- a interlíngua (forma errada) e
- a forma em português.

O mesmo modelo de língua utilizado para verificar a corretude do erro artificial na geração de erros de marcadores discursivos foi utilizado na verificação dos erros de expressões formulaicas. O modelo de língua empregado foi o apresentado na Seção 5.1.2 e a probabilidade de cada n-grama no modelo de língua serviu como referência para incluir ou não um item no léxico de erros.

A tradução literal consiste em traduzir cada palavra que compõe a expressão em espanhol, de forma individual, por meio do modelo de tradução utilizado para obter as equivalências entre o espanhol e o português. No Exemplo 4, observa-se como o hispano falante utilizou a expressão "*como consecuencia disto*". Essa expressão, embora não represente um erro, não é utilizada com frequência na escrita em português e acredita-se que foi produzida após uma tradução literal da expressão fixa em espanhol "*como consecuencia de ello*". A equivalência correta dessa expressão em espanhol é "*como consecuencia*" e foi obtida utilizando-se o algoritmo de tradução estatístico.

Exemplo 4: "*Como consecuencia disto* / (**como consecuencia**), é necessário seguir trabalhando na sua concretização"

Na primeira coluna da Tabela 6.13, são apresentadas algumas das expressões formulaicas em espanhol incluídas no léxico bilíngue. Na segunda coluna aparecem os erros gerados a partir da tradução literal e na última coluna aparece o equivalente correto em português das expressões em espanhol. Tanto as traduções literais, classificadas como erros, quanto as equivalências em português foram obtidas de forma automática com o método proposto.

Analisando-se os exemplos que aparecem na Tabela 6.13, observa-se que várias categorias de erros podem aparecer dentro das expressões geradas artificialmente. Os dois primeiros exemplos apresentados na Tabela 6.13 mostram erros lexicais produzidos pela interferência de expressões em espanhol.

Tabela 6.13: Exemplo de entradas incluídas no léxico de expressões formulaicas

Expressão em espanhol	Tradução Literal (Erro)	Expressão em português
<i>al margen de a decir verdad</i>	<i>ao margem de a dizer verdade</i>	além de na verdade, para dizer a verdade
<i>la evaluación se realizó</i>	<i>a avaliação se fez</i>	a avaliação foi feita
<i>el diseño experimental foi</i>	<i>o desenho experimental foi</i>	o design experimental foi
<i>a pesar de ello en este trabajo</i>	<i>a pesar disto em esse trabalho</i>	apesar disso nesse trabalho

O terceiro erro parece uma inadequação no uso de expressões em português, na verdade, uma preferência no estilo de escrita. Por outro lado, o quarto exemplo mostra como a tradução literal é válida para se obter uma expressão correta em português, porém a palavra "*diseño*" devia ter sido traduzida para "design" e não para "desenho". Neste último exemplo o erro foi produzido pelo uso incorreto de um falso cognato nas duas línguas. O quinto exemplo foi um erro muito recorrente entre os hispano falantes. O erro gerado nesse caso parece um erro produzido por interferência ortográfica combinado com escolhas lexicais, pois devia ter sido utilizado "disso" no lugar de "disto". O último erro criado mostra um erro de contração, pois o método de tradução literal não considera as contrações na geração dos erros. Por meio dos exemplos apresentados na Tabela 6.13, destaca-se que o método de geração de erros artificiais é capaz de gerar erros de naturezas diferentes, sendo capaz de corrigir esses erros, corrigindo a expressão formulaica como um todo.

O léxico bilíngue de expressões formulaicas contém 264 expressões paralelas. Destaca-se que às vezes o equivalente em português de uma expressão formulaica em espanhol não é uma expressão formulaica. Após a tradução literal das expressões em espanhol foram geradas 160 expressões que não aconteciam no modelo de língua e, como consequência, foram incluídas no léxico de erros.

Uma das deficiências do léxico proposto é que ele não foi validado na tarefa de detectar e corrigir erros, porém a validação proposta para verificar a saída do procedimento que detecta e corrige erros de marcadores discursivos pode ser aplicada para validar a precisão do léxico bilíngue na detecção e correção de erros.

6.3 Ferramenta de Auxílio à Escrita HABLA

Após apresentar os léxicos criados, na Figura 6.15 mostramos uma proposta de ferramenta que incorpora os léxicos bilíngues para apoiar os três auxílios propostos:

- (a) Equivalências entre o Espanhol e o Português;
- (b) Escolhas lexicais em português;
- (c) Detecção e Correção de erros.

A ferramenta foi planejada para auxiliar o hispano falante em tempo real, de forma que o usuário consegue consultar o sistema para obter informações das quatro bases lexicais disponíveis:

1. cognatos e falsos cognatos das línguas português e espanhol;
2. pares de marcadores discursivos;
3. expressões formulaicas que aparecem nos textos científicos e
4. verbos relacionados com pesquisa científica em português.

Ao mesmo tempo, a ferramenta também incorpora léxicos criados para identificar alguns tipos de erros, dessa forma também é um sistema de pós-edição dos textos em português.

A Figura 6.15 mostra o fluxo de uma ferramenta interativa que auxilia na composição dos textos em português, assim como na revisão de alguns erros lexicais produzidos pelas interferências da língua nativa. A ferramenta proposta apresenta uma tela de busca em que os usuários podem consultar os léxicos bilíngues para obter os melhores equivalentes em português. É importante apontar que os auxílios para obter os equivalentes em português trazem mais informações que um tradutor, embora técnicas de tradução tenham sido utilizadas para compilá-los. Isso acontece devido as heurísticas que foram combinadas com os algoritmos de tradução para se obter os equivalentes mais adequados.

Durante a escrita do texto, o usuário pode requisitar à ferramenta os equivalentes de um termo em espanhol. Junto com os equivalentes são apresentadas sentenças em português em que eles são utilizados. Na Figura 6.16, o aprendiz requisita ao sistema os equivalentes do marcador discursivo "por

Figura 6.15: Visão geral dos léxicos bilíngues incluídos em uma ferramenta de auxílio à escrita.

último”. Veja que vários sinônimos do marcador foram retornados para o usuário, sendo responsabilidade dele a escolha da melhor opção. Esse auxílio está disponível para todos os léxicos bilíngues propostos nesta pesquisa.

Figura 6.16: Tela com equivalências em português do marcador discursivo em espanhol "por último".

O segundo auxílio, apresentado na Figura 6.15, também é um auxílio de consulta em que o usuário requisita sinônimos em português. Como mostrado na Figura 6.15, esse auxílio está disponível para os verbos e marcadores dis-

cursivos. Durante a escrita do texto, o usuário pode requisitar por sinônimos de verbos ou marcadores discursivos. Na Figura 6.17, aparece um exemplo em que o usuário perguntou ao sistema o sinônimo do verbo "outorgar".

Figura 6.17: Tela com sinônimos do verbo "outorgar".

Os dois primeiros auxílios são de consulta e cabe ao usuário acioná-los. Esses auxílios estão direcionados para usuários mais avançados que têm interesse em diversificar os itens lexicais utilizados. No entanto, também são úteis para usuários em níveis iniciais de aprendizado que precisam de auxílio para transferir itens em espanhol para o português. Até este momento, os auxílios oferecidos pela ferramenta são similares aos auxílios apresentados pelos sistemas PENS (Huang et al., 2000) e WINGS (Dai et al., 2014), mas a abordagem proposta para construir os auxílios é similar à abordagem proposta por Huang et al. (2000) que se basearam em métodos de tradução automática.

O último auxílio, apresentado na Figura 6.15, recebe um texto produzido pelo usuário e detecta e corrige erros neste texto. Os erros identificados são erros de marcadores discursivos, de expressões formulaicas e de mudanças incorretas de sufixos. Esses são os erros disponibilizados até o momento, contudo na medida em que novos auxílios forem criados, eles podem ser inseridos na ferramenta. Como mostrado na Figura 6.18, o sistema detectou, automaticamente, o uso incorreto do marcador discursivo "em concreto", desta forma, utilizando os procedimentos propostos apresenta os marcadores corretos que corrigem o problema.

Uma vez detectado o erro, o usuário decide se deve ou não corrigi-lo. Para isso, é possível consultar o sistema e verificar exemplos que mostrem sentenças em que marcadores como "de fato" e "na verdade" foram utilizados.

O recurso que inclui cognatos e falsos cognatos também é acionado durante a revisão do texto, pois caso um falso cognato for detectado, o sistema mostra uma alerta indicando a presença do falso cognato entre as línguas espanhol e português. Além disso, mostra qual é o significado desse falso cog-

Figura 6.18: Tela com a detecção e correção do erro em que o item "em concreto" foi utilizado como marcador discursivo.

nato em espanhol, o que representa uma informação que pode ser utilizada pelo hispano falante para reconhecer qual é o sentido que ele deseja utilizar na sentença. Na Figura 6.19, mostra-se uma alerta de falso cognato, em que a palavra "patrão" foi utilizada incorretamente. Coincidentemente, o exemplo mostra o alerta junto com a ocorrência do erro, porém o sistema não consegue detectar erros de falsos cognatos, pois uma análise de desambiguação de sentidos seria necessária. No exemplo, quando as possíveis trocas para português são apresentadas, aparece a palavra para corrigir o erro; no exemplo seria a palavra "padrão". A desambiguação lexical do exemplo que aparece na Figura 6.19 é difícil até para humanos, pois é necessário ter conhecimento de domínio, e reconhecer que no exemplo se está falando de técnicas de reconhecimento de padrões.

Figura 6.19: Tela com uma alerta que mostra a ocorrência de uma falso cognato entre o espanhol e o português no texto.

Acredita-se que o alerta apresentado pelo sistema, na Figura 6.19, permita

uma interação entre o usuário e a ferramenta que mostra que esta forma de apresentar as sugestões é muito útil para os aprendizes da língua, pois o usuário consegue analisar as saídas do sistema e, como consequência, contribuir na aprendizagem da língua. No entanto, essa hipótese precisa ser verificada.

Adicionalmente, a existência da ferramenta é uma alternativa para compilar textos produzidos por hispano falantes. Todos os usuários que utilizem a ferramenta poderiam se cadastrar e os textos ficariam armazenados, o que representaria um aumento do número de textos no *córpus* de aprendizes Espanhol-Acadêmico-Br. Além disso, com a disponibilização de vários textos de um mesmo usuário seria possível realizar análises longitudinais para avaliar a evolução dos hispano falantes na produção escrita em português.

6.4 Considerações Finais

Os experimentos apresentados neste capítulo mostraram como as técnicas de tradução automática podem ser úteis na criação de recursos que apoiem a escrita em português de hispano falantes. Por meio de tradução foram criados os seguintes léxicos bilíngues: (i) listas de cognatos e falsos cognatos das línguas português e espanhol; (ii) listas de pares de marcadores discursivos; (iii) listas das expressões formulaicas que aparecem nos textos científicos, (iv) listas de verbos relacionados com pesquisa científica em português, recorrentemente utilizados em textos acadêmicos e (v) listas de substantivos e adjetivos que compartilham sufixos paralelos. Para compilar os recursos, diversas técnicas computacionais foram combinadas com as técnicas de tradução automática. Esse fato, junto com a dificuldade na tarefa de avaliar todos os recursos criados, representou um dos maiores desafios desta pesquisa. Todos os léxicos foram gerados considerando-se os erros ou inadequações identificados na anotação do *córpus* Espanhol-Acadêmico-Br.

As técnicas de tradução automática estiveram envolvidas na geração dos três auxílios propostos: (i) Equivalências entre o espanhol e português; (ii) Escolhas lexicais em português e (iii) Detecção e correção de erros. Três métodos foram testados para se obter as equivalências lexicais entre o espanhol e o português. Os métodos para se obter as equivalências foram avaliados intrinsecamente na tarefa de traduzir palavras e expressões. Acredita-se que os resultados obtidos na avaliação sejam equivalentes na compilação de todos os léxicos bilíngues. Os resultados da avaliação mostraram que o método baseado em similaridade auxilia os métodos estatístico e empírico, principal-

mente quando palavras de baixa frequência são traduzidas. Esse método foi utilizado na geração do léxico de palavras com sufixos paralelos, porém em trabalhos futuros, espera-se usá-lo para obter outras equivalências, inclusive entre expressões, sempre que sejam identificadas expressões similares entre as duas línguas.

A saída do método de tradução automática também foi utilizada para obter sinônimos entre itens lexicais em português. O método para obter sinônimos foi criado com a finalidade de categorizar os marcadores discursivos segundo sua função discursiva, porém o método também foi utilizado para obter sinônimos de verbos em português. Este método não foi avaliado; sugere-se que ele seja validado por um lexicógrafo ou linguista. Inclusive, pode ser avaliado por um professor de língua estrangeira com a finalidade de avaliar a utilidade do recurso no processo de ensino/aprendizado. Atualmente, a única possibilidade de avaliar os recursos de sinônimos é comparando-o com outro recurso. No caso dos marcadores discursivos isso é difícil, pois há muita discussão sobre os itens que são ou não marcadores discursivos e como eles são categorizados. Poderia ter sido comparado com a pesquisa de Fernández (2005), mas essa comparação não vai trazer muita informação, pois foi comprovado que a lista de Fernández (2005) não inclui marcadores que foram incluídos em outras pesquisas. A mesma análise é aplicada na comparação dos sinônimos obtidos com o método de tradução e os sinônimos da *WordNet.Br*. Caso existam divergências na comparação não é possível afirmar nada, pois a *WordNet.Br* não foi validada.

Além disso, o método de tradução automática foi utilizado para gerar erros artificiais. A similaridade entre o espanhol e o português foi uma das motivações para se criar o método de geração de erros. A outra motivação foi a ausência de um grande corpus com uma vasta análise linguística que guiara um processo automático de detecção e correção de erros. Embora, a geração de erros artificiais seja uma previsão dos erros dos aprendizes, sendo impossível prever todos os erros que podem ser cometidos, o método mostrou-se eficiente na tarefa de detecção e correção de erros. A baixa precisão do método na detecção/correção de erros está relacionada com o fato do item lexical não acontecer no modelo de língua. Esse resultado indica que um aumento no tamanho do corpus representará uma melhoria no desempenho do método.

Adicionalmente, o método de geração artificial de erros é um método simples que pode ser inserido em etapas que antecedam a anotação manual de erros, desta forma, poderiam ser reconhecidos alguns problemas antes da

anotação, que sempre é uma tarefa muito demorada que requer tempo. O método também é útil para gerar um corpus com erros artificiais, que depois pode apoiar tarefas de detecção e correção automáticas de erros. A geração artificial de erros reflete alguns dos mecanismos adotados pelos aprendizes que desencadeiam em um erro, a partir disso, análises linguísticas nos recursos criados podem aprimorar o tratamento de erros produzidos pela interferência da língua nativa.

O método proposto para avaliar a tarefa de detecção de erros se mostrou eficiente, pois sem previa anotação conseguiu-se levantar vários exemplos que mostraram a ocorrência dos erros artificiais. Segundo a literatura é comum que nos textos dos aprendizes aconteçam diversos tipos de erros em um mesmo contexto, esse fato foi comprovado com o método que detectou os erros gerados artificialmente. Nas sentenças em que um erro artificial aconteceu, outros erros foram detectados. Isso foi uma vantagem do método proposto, pois erros de outras categorias foram levantados no contexto em que aconteceram os erros artificiais.

A avaliação dos acertos na tarefa de correção dos erros identificados automaticamente foi uma avaliação extrínseca que mostrou que é possível aprimorar a escrita dos aprendizes por meio dos recursos lexicais compilados. Seguindo essa linha, esses experimentos validam a primeira hipótese desta pesquisa, que apontava que a utilização de métodos da área de tradução automática permitiria criar recursos lexicais para auxiliar a escrita em português. No entanto, para comprovar a veracidade da segunda hipótese seria necessário realizar uma avaliação extrínseca de todos os recursos criados na tarefa de aprimorar a escrita dos hispano falantes. Uma alternativa para realizar essa avaliação é que a ferramenta HABLA²⁰, com os recursos disponíveis seja validada pelos usuários. Essa validação representa um dos trabalhos futuros relatados nesta pesquisa.

²⁰<http://nilc.icmc.usp.br/habla>

Conclusões

Neste projeto de doutorado, denominado HABLA (Hispano Falantes Adquirindo uma Base Linguística Acadêmica), foi proposta uma aproximação entre as áreas de pesquisas: tradução automática estatística e auxílio à escrita entre línguas muito próximas, considerando a análise de erros extraídos de corpus de aprendizes. Seguindo esta linha, o projeto aliou os seguintes objetivos:

Objetivo I: Analisar os problemas que nativos do espanhol possuem ao escrever teses e dissertações em português, considerando as principais interferências do espanhol que esses alunos pesquisadores mantêm na escrita do português;

Objetivo II: Avaliar os resultados de métodos e recursos léxico-sintático-semânticos baseados em corpus paralelos alinhados, que são utilizados em TA, na área de Sistemas de Auxílio à Escrita, para auxiliar na pós-edição e composição de um texto científico escrito em português por um aluno pesquisador cuja língua materna é o espanhol.

A primeira hipótese levantada neste doutorado é a de que a utilização de métodos da área de tradução automática permite a criação de recursos lexicais para auxiliar a escrita em português. Acredita-se que esses recursos possam sanar alguns dos problemas, principalmente as interferências do espanhol, que ainda não foram resolvidos nem com o emprego de ferramentas que ensinam, principalmente, a estruturação retórica dos textos, como é o caso do sistema SciPo, nem com corretores ortográficos para o português, que trabalham com nativos do português como usuários alvos. Uma segunda hi-

pótese é a de que o uso de recursos linguístico-computacionais, desenvolvidos para apoiar a escrita no nível lexical, pode melhorar a qualidade linguística dos textos em português produzidos pelos nativos do espanhol.

Neste capítulo são apresentados os produtos obtidos no projeto HABLA que são divididos em: Contribuições (Seção 7.1), Limitações (Seção 7.2) e Trabalhos futuros (Seção 7.3), apresentados a seguir.

7.1 Contribuições

Por meio da metodologia apresentada no Capítulo 5 foram analisados os problemas que nativos do espanhol possuem ao escrever teses e dissertações em português. As análises focaram na identificação das principais interferências do espanhol que esses alunos mantêm na escrita do português. Como resultados destas análises foram realizados:

1. o levantamento e quantificação dos principais problemas causados pelas marcas do espanhol, deixadas nos textos acadêmicos em português escritos por nativos do espanhol;
2. a disponibilização do corpus Espanhol-Acadêmico-Br, que contém textos acadêmicos em português produzidos por hispano falantes;
3. a construção de uma tipologia dos principais erros dos hispano falantes;
4. a anotação manual de erros, gerando um corpus de treinamento;
5. a compilação do corpus paralelo com erros em português;
6. a implementação de métodos para gerar auxílios de apoio à escrita.

É importante apontar que o fato dos erros de aprendizes terem sido anotados por dois anotadores é um destaque desta pesquisa. Um dos trabalhos que se conhece, em que todos os erros dos aprendizes foram anotados por mais de um humano foi apresentado por Dahlmeier et al. (2013). Neste projeto, estimou-se o *Kappa* para avaliar a concordância entre os humanos na tarefa de detectar e corrigir erros. Os valores obtidos na detecção, classificação e correção de erros estão de acordo com os valores relatados por Dahlmeier et al. (2013). O corpus Espanhol-Acadêmico-Br está disponível publicamente¹ e pode ser utilizado em outras pesquisas. Além disso, a anotação dos erros

¹<http://nilc.icmc.usp.br/habla>

mostrou o conjunto de erros que difere dos erros que os nativos da língua cometem, assim como os erros que ainda não são tratados pelos corretores gramaticais mais utilizados na língua portuguesa.

A partir das dificuldades identificadas no *cópus* de aprendizes foram propostos métodos para gerar auxílios à escrita em forma de recursos lexicais. Esses métodos foram detalhados no Capítulo 6 e mostraram como os métodos baseados em *cópus* paralelos alinhados e em tradução automática podem auxiliar a criação de léxicos bilíngues que apoiam a composição ou pós-edição de um texto científico em português por um aluno pesquisador cuja língua materna é o espanhol. A seguir, são listados os auxílios oferecidos pelos léxicos bilíngues propostos neste projeto:

1. Equivalências entre o espanhol e português;
2. Escolhas lexicais em português;
3. Detecção e correção de erros.

Os métodos de alinhamento de textos paralelos e tradução automática combinaram-se com heurísticas para, de forma semi-automática, gerar léxicos bilíngues que podem ser inseridos em ferramentas de auxílio à escrita ou em processos de ensino/aprendizado. Os léxicos bilíngues gerados nesta pesquisa são compostos por palavras e multpalavras e para extrair esses itens dos *cópus* paralelos foram utilizadas as seguintes informações: (i) léxicos monolíngues; (ii) anotação morfossintática e (iii) frequência e medidas de associação. A extração dos itens utilizando léxicos monolíngues surgiu como alternativa para extrair marcadores discursivos dos textos, pois não há nenhum *tagger* ou *parser* de português que rotule automaticamente esses itens. Caso essa ferramenta seja disponibilizada em um futuro próximo, o processo de extração pode ser facilmente substituído.

Para estabelecer equivalências entre os léxicos monolíngues, foram propostos três métodos:

1. baseado em similaridade ortográfica;
2. baseado em tradução automática estatística e
3. modificação do método empírico apresentado por Bai et al. (2009).

Os métodos estatístico e empírico tiveram melhor desempenho que o método baseado em similaridade na obtenção de equivalências entre palavras, no

entanto, comprovou-se que o método baseado em similaridade auxilia os métodos estatístico e empírico, aprimorando os valores de precisão e cobertura. Como os métodos estatístico e empírico são influenciados pela frequência dos itens lexicais no corpus, diferentemente do método baseado em similaridade, esse último pode auxiliar na obtenção de equivalentes de palavras de baixa frequência. Por meio dos métodos propostos foram criados os seguintes recursos lexicais:

1. léxicos bilíngues de cognatos e falsos cognatos do português e espanhol;
2. pares de marcadores discursivos;
3. expressões formulaicas que aparecem nos textos científicos;
4. verbos relacionados com pesquisa científicas em português, recorrentemente utilizados em textos;
5. substantivos e adjetivos com sufixos equivalentes em espanhol e em português.

Além disso, nesta pesquisa foi proposto um método para obter sinônimos baseado nas equivalências por tradução. O método foi utilizado para apresentar sinônimos de verbos em português e para categorizar marcadores discursivos segundo sua função discursiva. Para categorizar os marcadores discursivos, o método de equivalências por tradução foi combinado com medidas de associação e um algoritmo de agrupamento.

Para finalizar, o léxico bilíngue de marcadores discursivos, expressões formulaicas, substantivos e adjetivos foram estendidos, por meio da geração artificial de erros. Seguindo essa linha, foi proposto um léxico de erros artificiais, cuja distinção é que ele mantém um controle do erro e da forma correta em português, detectando o erro e sugerindo a correção, simultaneamente. Para a geração artificial dos erros de cada léxico foram aplicadas diversas heurísticas.

O léxico de erros de marcadores discursivos foi testado na detecção e correção automática de erros de marcadores discursivos, produzidos pela interferência do espanhol. Além disso, essa tarefa também foi abordada com um método que combina um modelo de língua com similaridade ortográfica entre expressões. Os resultados dos métodos para detectar erros mostraram que ambas as abordagens são viáveis para tratar erros de marcadores discursivos produzidos pela interferência da língua nativa. As abordagens propostas se mostraram mais eficientes ainda na tarefa de corrigir os erros, que é um

dos principais desafios em sistemas que detectam e corrigem erros produzidos por aprendizes de uma língua estrangeira. O sucesso dos métodos representa um incentivo para estendê-los com a finalidade de aprimorar os resultados obtidos.

A verificação manual da saída dos métodos de detecção e correção de erros permitiu validar o desempenho dos métodos propostos para tratar erros de marcadores discursivos. Adicionalmente, foram levantadas sentenças em que outros erros, além dos erros de marcadores discursivos, aconteceram. Por exemplo, foram identificados erros de preposições, de artigos e erros de ortografia. O desempenho positivo do método de detecção e correção de erros validou a primeira hipótese desta pesquisa.

O método proposto para construir os léxicos bilíngues pode ser customizado e adaptado para diferentes áreas de pesquisas ou diferentes comunidades discursivas. Tendo como entrada um corpus paralelo e listas monolíngues de itens lexicais, é possível obter um léxico bilíngue entre duas línguas, que mostre as equivalências entre as línguas e, conseqüentemente, algumas das estratégias que podem ser utilizadas pelos aprendizes para gerar palavras e/ou expressões na segunda língua. Esses mecanismos de transferência são informações que, se utilizadas, podem enriquecer o processo de ensino/aprendizado dos hispano falantes. É importante apontar que os léxicos gerados segundo os métodos propostos devem ser refinados por lexicógrafos ou linguistas, com a finalidade de aprimorar seus efeitos positivos nos auxílios à escrita.

Um destaque da abordagem proposta para detectar e corrigir erros é que os métodos e os recursos criados não dependem da anotação de um grande corpus com erros para identificar os problemas. Desta forma, os recursos poderiam ser utilizados em etapas anteriores à anotação de erros em um corpus de aprendizes, para verificar a presença dos erros artificiais incluídos no recurso. Como em textos de aprendizes é comum a ocorrência de mais de um erro por sentença, espera-se que além dos erros gerados artificialmente, outros erros possam ser levantados. Uma análise manual do subconjunto de sentenças selecionado poderia resultar na identificação de alguns problemas dos aprendizes sob estudo. Esses problemas poderiam referenciar uma tipologia de erros, que considere erros que são recorrentes. Uma tipologia feita às cegas muitas vezes considera tipos de erros que não são representativos do grupo de aprendizes sob estudo. Além disso, os recursos poderiam indicar em uma etapa de pré-processamento os textos com maior número de erros. Esses

textos poderiam ser os selecionados para começar uma anotação manual.

Com a metodologia proposta foi possível responder as questões de pesquisa apontadas no início deste projeto, assim como validar a primeira hipótese proposta.

7.2 Limitações

A anotação de erros de aprendizes é uma das principais contribuições desta pesquisa, pois caracterizou a produção escrita dos hispano falantes. Contudo, o número de erros anotados não foi suficiente para utilizar o *córpus* em abordagens de aprendizado de máquina para a construção de um sistema computacional. Os erros foram de categorias diversas e aconteceram de forma esparsa nos textos selecionados.

Por outro lado, um dos maiores desafios enfrentados nesta pesquisa foi a compilação dos léxicos bilíngues, devido à diversidade de itens lexicais. Por esse motivo, várias abordagens computacionais e recursos linguísticos foram testados. A metodologia proposta para compilar os léxicos bilíngues possui várias etapas, nas quais foram combinadas diversas análises. Como consequência, foi necessária a avaliação da compilação dos léxicos nas diferentes etapas propostas. Um fato que foi constatado nesta pesquisa foi a necessidade de análises linguísticas para disponibilização de *gold standards*, os quais facilitariam a avaliação dos métodos propostos. Dada a inexistência desses *gold standards* foi necessário o uso do conhecimento (ou avaliação) humano em algumas etapas do processo. Dado o problema anterior, foi unicamente avaliado o léxico de marcadores discursivos na tarefa de auxílio à escrita dos hispano falantes. No entanto, a avaliação proposta para verificar o desempenho do léxico de marcadores discursivos pode ser replicada na avaliação dos outros léxicos.

Outra deficiência relacionada à metodologia de avaliação dos léxicos gerados nesta pesquisa é que eles poderiam ter sido avaliados extrinsecamente, depois de inseridos na ferramenta HABLA, considerando os auxílios apresentados aos hispano falantes. No entanto, a avaliação anterior requer a participação de hispano falantes, linguistas e/ou professores de ensino de língua, sendo uma tarefa demorada; a tarefa foi colocada como trabalho futuro. Por esse motivo, a segunda hipótese desta pesquisa não foi validada.

Outra limitação dos métodos propostos é que eles dependem de *córpus* paralelos e léxicos monolíngues que são recursos escassos. Nesta pesquisa, fo-

ram selecionados os *córpus* mais adequados, disponíveis na literatura, porém o tamanho desses *córpus* impactou diretamente no desempenho dos algoritmos. O aumento do *córpus* paralelo representaria um aumento na precisão e cobertura dos métodos utilizados para obter as equivalências entre o espanhol e o português. Como consequência, também melhoraria a precisão do método que corrige os erros artificiais. Por outro lado, o tamanho do modelo de língua também impactou na tarefa de detectar e corrigir erros. Compilar um modelo de língua maior pode ser mais simples que compilar um *córpus* paralelo, porém os *léxicos* bilíngues foram desenvolvidos para apoiar a escrita acadêmica e, por isso, recomenda-se que o modelo de língua mostre as principais características desse gênero.

7.3 Trabalhos Futuros

Esta pesquisa de doutorado trouxe várias contribuições para o PLN, especificamente na área de auxílio à escrita em português como segunda língua; assim diversos trabalhos futuros são elencados a partir dos experimentos desenvolvidos.

1. Como o *córpus* Espanhol-Acadêmico-Br está disponível para outras pesquisas, acredita-se que análises para identificar outros erros lexicais, diferentes dos que foram tratados nesta pesquisa seriam interessantes. Essas análises deveriam focar na identificação da natureza dos erros lexicais, como acontece no estudo apresentado por (Carrió-Pastor e Mestre-Mestre, 2013). Entender a natureza dos erros lexicais auxiliaria métodos computacionais a detectar e corrigir erros, os quais podem ser utilizados para aprimorar os corretores ortográficos em português para tratarem erros produzidos por hispano falantes.
2. Aprimorar o classificador de textos nos níveis de proficiência. Novos atributos podem ser explorados, tais como: categorias de erros.
3. Combinar os métodos de tradução estatística e empírica com o método baseado em similaridade para auxiliar na tradução de palavras que acontecem com baixa frequência no *córpus*. O mesmo método poderia ser testado na tarefa para traduzir expressões com baixa frequência.
4. Considerando mudanças recorrentes nas palavras em espanhol e em português é possível treinar um algoritmo de classificação que aprimore o

algoritmo de similaridade. Por exemplo: as sequências "cua" e "cue" em espanhol mudam com frequência para as sequências "qua" e "que", como nos pares de equivalentes "*cuanto* - quanto" e "*frecuencia* - frequência" (o algoritmo de similaridade não conseguiu identificar as correspondências entre estes itens). É importante reforçar que estas mudanças não são consideradas pelos corretores ortográficos que auxiliam os nativos da língua. A seguir outras mudanças recorrentes que foram identificadas:

- (a) Mudança de "s" pela "ss": *escasez* - escassez; *resonancia* - ressonância, *esencia* - essência;
- (b) Mudança de "MN" por "M" ou "N": *omnipotencia* - onipotência; *inmunodeficiencia* - imunodeficiência; *inminencia* - iminência, *somnolencia* - sonolência;
- (c) Mudança de (B por V) ou (V por B): *absorbencia* - absorvência; *inmovilidad* - imobilidade; *aprobación* - aprovação;
- (d) Mudança de CIA por ÇA: *sentencia* - sentença; *diferencia* - diferença, *herencia* - herança, *licencia* - licença;
- (e) Mudança de CT por T: *reluctancia* - relutância; *actualidad* - atualidade; *electricidad* - eletricidade;
- (f) Mudança de DH por D: *adherencia* - aderência;
- (g) Mudança de UCIÓN por UIÇÃO: *contribución* - contribuição; *institución* - instituição; *distribución* - distribuição;
- (h) Eliminação de H: *inhibición* - inibição; *rehabilitación* - reabilitação; *deshidratación* - desidratação.

5. Uma alternativa para aumentar a cobertura do léxico de cognatos e falsos cognatos é extrair as palavras do corpus paralelo Fapesp. Os pares seriam extraídos dos alinhamentos lexicais, obtidos com o Giza++. A lista de pares de palavras criada após o alinhamento pode ser separada em palavras cognatas e palavras não relacionadas utilizando um limiar da medida de similaridade LCSR, que mostrou bom desempenho na tarefa. A lista de cognatos poderia ser utilizada para classificar cognatos e falsos cognatos utilizando a configuração do experimento que teve melhor desempenho nesta pesquisa (Seção 6.2.1). Desta forma, a nova lista seria empregada como um conjunto de teste.

6. Outra alternativa para aprimorar a compilação do léxico de cognatos e falsos cognatos é explorar medidas de similaridade semânticas. Essas

medidas poderiam ser incluídas como novos atributos na configuração do experimento com melhor resultado na Seção 6.2.1. Uma estrutura taxonômica que poderia ser utilizada para estimar a similaridade semântica é a openWordnet-PT (De Paiva et al., 2012). Com a mesma finalidade, também pode ser avaliada a *Spanish and Catalan WordNet*² que segue o *framework* da *EuroWordNet* que é um recurso disponível com a mesma estrutura que a *WordNet* de *Princeton*.

7. Para detectar automaticamente erros de falsos cognatos nos textos dos aprendizes, uma alternativa seria empregar algoritmos de classificação que analisem o contexto em que o falso cognato foi utilizado. Para treinar o algoritmo de classificação, informações do contexto em que o falso cognato deve ser utilizado seriam necessárias. Nesse cenário, sugere-se estudar as pesquisas apresentadas por Sobrevilla Cabezudo (2015) e Nóbrega e Pardo (2014), as quais focaram em métodos de desambiguação lexical.
8. Para ordenar melhor os verbos a partir da estimativa da Tf-idf, sugere-se incluir na análise um corpus da língua falada, desta forma o cálculo da Tf-idf vai conseguir diferenciar melhor os termos dos textos acadêmicos. Os experimentos poderiam ser testados na detecção automática de erros relacionados com verbos que não são convencionais no gênero acadêmico. Além disso, sugere-se testar alguma ferramenta disponível para o português para desambiguar o sentido do verbo. Atualmente, foi finalizado um mestrado que foca na desambiguação lexical de verbos (Sobrevilla Cabezudo, 2015).
9. A geração artificial de erros de marcadores discursivos foi obtida a partir da tradução literal dos marcadores em espanhol e dos padrões da vírgula. Entretanto, às vezes os marcadores discursivos são compostos por artigos e preposições que, quando traduzidos para o português resultam em uma contração. Na geração de erros artificiais por tradução literal não é possível obter sequências em que as contrações do português sejam empregadas. Por exemplo, para o marcador em espanhol “*por lo visto*” é obtida a tradução literal “por o visto” e somente será obtida a sequência “pelo visto” combinando a tradução literal com a contração de “por + o”. Para refinar o processo de geração artificial de erros sugere-se incluir novas heurísticas, por exemplo: a geração de contrações.

²<http://nlp.lsi.upc.edu/web/index.php?option=comcontenttask=viewid=31Itemid=57>

10. No método de detecção de erros de marcadores discursivos utilizando um modelo de língua e similaridade entre expressões, a maioria dos n-gramas detectados como falsos positivos continha entidades nomeadas. Para diminuir o número de falsos positivos sugere-se inserir nos experimentos um detector de entidades nomeadas. Com esta finalidade, uma ferramenta que deve ser testada é o *Rembrandt* (Cardoso, 2012), que detecta entidades nomeadas da língua portuguesa.

Referências Bibliográficas

- Abdalla, R. M. e Teufel, S. (2006). A bootstrapping approach to unsupervised detection of cue phrase variants. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, páginas 921–928, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 89 e 94.
- Aidinlou, N. e Mehr, H. (2012). The effect of discourse markers instruction on efl learners' writing. *World Journal of Education*, páginas 10–16. Citado nas páginas 8, 10, 45, 77 e 86.
- Al-Rfou, R. (2012). Detecting english writing styles for non-native speakers. *CoRR*, abs/1211.0498:1–9. Citado na página 216.
- Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the aless learner corpus. *Komaba Journal of English Education*, 1:105–127. Citado nas páginas 20 e 22.
- Alonso, R., Wanner, L., Vincze, O., Casamayor del Bosque, G., Vázquez, V., S.E., M., e Prieto, G. (2010). Towards a motivated annotation schema of collocation errors in learner corpora. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., e Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). Citado nas páginas 8, 9 e 45.
- Aluísio, S. (1995). *Ferramentas para auxiliar a escrita de artigos científicos em inglês como língua estrangeira*. Tese de Doutorado, Instituto de Física de São Carlos, Universidade de São Paulo. Citado na página 98.

- Aluísio, S., Schster, E., Feltrim, V., Pessoa, A., e Oliveira, O. (2005). Evaluating scientific abstracts with a genre-specific rubric. In Looi, C.-K., McCalla, G. I., Bredeweg, B., e Breuker, J., editors, *AIED*, volume 125 of *Frontiers in Artificial Intelligence and Applications*, páginas 738–740. IOS Press. Citado na página 135.
- Aluísio, S., Specia, L., Gasperin, C., e Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '10, páginas 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 137.
- Amaral, A. e Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. In *RECALL 23(1)*, páginas 1–22. Citado na página 6.
- Andersen, O. (2007). Grammatical error detection using corpora and supervised learning. In *Proceedings of the Twelfth ESSLLI Student Session*, páginas 1–10. Citado nas páginas 8, 105, 108, 109 e 110.
- Araújo, S., Almeida, J., Simoes, A., e Dias, I. (2010). Apresentação do projecto Per-Fide: Paralelizando o português com seis outras línguas. *Linguamatica*, 2(2):71–74. Citado na página 96.
- Attali, Y. e Busrstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal Of Technology Learning And Assessment*, 4(3):1–30. Citado nas páginas 6 e 99.
- Atwell, E. (1997). How to detect grammatical errors in a text without parsing it. In *Proceedings of the Third Conference of the European Association for Computational Linguistics (EACL)*, páginas 38–45, Copenhagen. Citado nas páginas 8 e 110.
- Aziz, W., Pardo, T., e Paraboni, I. (2008a). Building a spanish-portuguese parallel corpus for statistical machine translation. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, WebMedia '08, páginas 369–371, New York, NY, USA. ACM. Citado nas páginas 51 e 52.
- Aziz, W., Pardo, T., e Paraboni, I. (2008b). An experiment in spanish-portuguese statistical machine translation. In *Proceedings of the 19th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence*,

- SBIA '08, páginas 248–257, Berlin, Heidelberg. Springer-Verlag. Citado na página 51.
- Aziz, W. e Specia, L. (2011). Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *STIL 2011*, Cuiabá, MT. Citado nas páginas 47, 125 e 128.
- Bai, M., You, J., Chen, K., e Chang, J. S. (2009). Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2009)*, páginas 478–486. Citado nas páginas 73, 74, 75, 76, 176, 179, 180, 215 e 247.
- Baptista, J., Costa, N., Guerra, J., Zampieri, M., Cabral, M., e Mamede, N. (2010). P-awl: Academic word list for portuguese. In *Computational Processing of the Portuguese Language*, volume 6001 of *Lecture Notes in Computer Science*, páginas 120–123. Springer Berlin Heidelberg. Citado nas páginas 85 e 86.
- Barbosa, L., Kumar, V. Sridhar, R., M., Y., e Bangalore, S. (2012). Harvesting parallel text in multiple languages with limited supervision. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, páginas 201–214. Citado nas páginas 47 e 222.
- Barlow, M. (2004). Collocate. Relatório técnico, Athelstan Publications. Citado na página 93.
- Begoña, V. e Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*, páginas 33–40, Trento, Italy. Citado nas páginas 47 e 52.
- Berber Sardinha, T. (2001). O corpus de aprendiz br-icle. Relatório técnico, São Paulo: Puc-Lael. Citado nas páginas 18, 20 e 21.
- Bonfante, A. e Nunes, M. (2000). Métodos empíricos para correção gramatical de línguas naturais. Relatório técnico, Relatórios Técnicos do ICMC-USP, 108 (NILC-TR-00-4). Citado nas páginas 5, 115 e 155.
- Briscoe, T., Carroll, J., e Watson, R. (2006). The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation ses-*

- sions, COLING-ACL '06, páginas 77–80, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 108 e 110.
- Brown, P., Cocke, J., Pietra, S., Pietra, V., Jelinek, F., Mercer, R., e Roossin, P. (1988). A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics - Volume 1*, COLING '88, páginas 71–76, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 52, 53, 54 e 55.
- Brown, P., Lai, J., e Mercer, R. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, páginas 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 49 e 50.
- Brown, P., Pietra, V., Pietra, S., e Mercer, R. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:263–311. Citado nas páginas 52, 53, 54, 55, 68, 74 e 178.
- Budanitsky, A. e Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47. Citado na página 107.
- Burstein, J., Chodorow, M., e Leacock, C. (2003). Criterion online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*. Citado nas páginas 6 e 99.
- Burstein, J., Chodorow, M., e Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3):27–36. Citado na página 99.
- Cardoso, N. (2012). Rembrandt-a named-entity recognition framework. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, páginas 1240–1243. Citado nas páginas 228 e 254.
- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254. Citado na página 23.
- Carlson, A. J., Rosen, J., e Roth, D. (2001). Scaling up context-sensitive text correction. In *Proceedings of the Thirteenth Conference on Innovative Applications of Artificial Intelligence Conference*, páginas 45–50. AAAI Press. Citado na página 107.

- Carpuat, M. e Diab, M. (2010). Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, páginas 242–245, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 73.
- Carrió-Pastor, M. e Mestre-Mestre, E. (2013). Lexical errors in second language scientific writing: Some conceptual implications. *International Journal of English Studies*, 14(1):97–108. Citado nas páginas 6, 8, 9, 158 e 251.
- Carvalho, A., Freire, J., e Da Silva, A. (2010). Teaching portuguese to spanish speakers. a case for trilingualism. In *Hispania*, páginas 70–75. Citado nas páginas 1, 2 e 31.
- Caseli, H. (2003a). Alinhamento sentencial de textos paralelos português-inglês. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Citado nas páginas 10, 46, 47, 49, 50 e 67.
- Caseli, H. (2007). *Indução de léxicos bilíngües e regras para a tradução automática*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Citado nas páginas 10, 45, 46, 51 e 75.
- Caseli, H. e Nunes, M. (2003). Sentence alignment of brazilian portuguese and english parallel texts. In *Proceedings of the Argentine Symposium on Artificial Intelligence (ASAI)*, páginas 1–11. Citado na página 51.
- Caseli, H., Paz Silva, A., e Nunes, M. (2004). Evaluation of methods for sentence and lexical alignment of brazilian portuguese and english parallel texts. In *Proceedings of the XVII Brazilian Symposium on Artificial Intelligence (SBIA) - Lecture Notes on Artificial Intelligence*, páginas 184–193, DBLP:conf/sbia/2004. Citado na página 46.
- Caseli, H., Ramisch, C., Nunes, M., e Villavicencio, A. (2009). Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 1:1–20. Citado nas páginas 51, 57, 67, 68, 71, 72, 75 e 76.
- Caseli, H. M. (2003b). Corpus paralelo e corpus paralelo alinhado: Propriedades e aplicações. In *Caderno de Resumos do 51º Semiário do GEL - Comunicação Coordenada: Lingüística de Corpus - construção de corpora para*

análise lingüística e treinamento de ferramentas de processamento de língua natural, páginas 209–209. Citado nas páginas 10, 45 e 47.

Castele, A. e Collewaert, K. (2013). The use of discourse markers in spanish language learners' written compositions. *Procedia - Social and Behavioral Sciences*, 95:550–556. Citado nas páginas 8, 86 e 87.

Cer, D., Galley, M., Jurafsky, D., e Manning, C. D. (2010). Phrasal: A toolkit for statistical machine translation with facilities for extraction and incorporation of arbitrary model features. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, HLT-DEMO '10, páginas 9–12, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 61.

Chang, Y., Chang, J., Chen, H., e Liou, H. (2008). An automatic collocation writing assistant for taiwanese efl learners: A case of corpus-based nlp technology. *Computer Assisted Language Learning*, 21(3):283–299. Citado nas páginas 9 e 112.

Chen, M., Huang, S., Hsieh, H., Kao, T., e Chang, J. S. (2012). Flow: A first-language-oriented writing assistant system. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, páginas 157–162, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 102.

Chodorow, M., Dickinson, M., Israel, R., e Tetreault, J. (2012). Problems in evaluating grammatical error detection systems. In *COLING'12*, páginas 611–628. Citado na página 118.

Chodorow, M., Gamon, M., e Tetreault, J. (2010). The utility of article and preposition error correction systems for english language learners: Feedback and assessment. *Language Testing*, páginas 335–353. Citado na página 99.

Chodorow, M. e Leacock, C. (2000). An unsupervised method for detecting grammatical errors. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, páginas 140–147, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Citado nas páginas 105, 107, 108 e 122.

Chodorow, M., Tetreault, J., e Han, N. (2007). Detection of grammatical errors involving prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, SigSem '07, páginas 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 105, 108, 109 e 122.

- Clark, S. e Curran, J. (2007). Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, 33(4):493–552. Citado na página 108.
- Coxhead, A. (2000). A new academic vocabulary list. *TESOL, Quarterly*, 34(2):213–238. Citado na página 85.
- 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. Citado nas páginas 2, 18, 20, 24, 25, 28, 29 e 31.
- Dagneaux, E., Denness, S., e Granger, S. (1998). Computer-aided error analysis. *System*, 26(2):163 – 174. Citado nas páginas 17 e 26.
- Dahlmeier, D., Ng, H. T., e Wu, S. M. (2013). Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, páginas 22–31, Atlanta, Georgia. Association for Computational Linguistics. Citado nas páginas 23, 120, 154, 169 e 246.
- Dahlmeier, D. e 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*, páginas 915–923, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 8, 20, 22, 105, 108 e 109.
- Dai, X., Liu, Y., Wang, X., e Liu, B. (2014). Wings: Writing with intelligent guidance and suggestions. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, páginas 25–30. Association for Computational Linguistics. Citado nas páginas 47, 100, 101 e 239.
- Daille, B., Dufour-Kowalski, S., e Morin, E. (2004). French-english multi-word term alignment based on lexical context analysis. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, páginas 919–922. Citado nas páginas 71, 72 e 75.
- Dale, R., Anisimoff, I., e Narroway, G. (2012). Hoo 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the*

- Seventh Workshop on Building Educational Applications Using NLP*, páginas 54–62, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 106 e 107.
- Damerau, F. J. (1964a). A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176. Citado nas páginas 62, 106, 107 e 223.
- Damerau, F. J. (1964b). A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176. Citado na página 106.
- Dayrell, C. (2008). Sense-related items and lexical patterning in english and portuguese scientific abstracts. In *International Symposium on Using Corpora in Contrastive and Translation Studies*, páginas 25–27, UCCTS 2008, Hangzhou. Citado nas páginas 20, 23 e 28.
- Dayrell, C. e Aluísio, S. (2008). Using a comparable corpus to investigate lexical patterning in english abstracts written by non-native speakers. In *Proceedings of the 6th international conference on Language Resources and Evaluation*, páginas 61–66, LREC 2008 Marrakech (Morocco). Citado nas páginas 8 e 20.
- Dayrell, C., Candido, J., Curi, M., Tagnin, S., e Aluísio, S. (2012). *Livro do IX encontro de linguística de corpus*, chapter Padrões Léxico-Gramaticais na Especificação de Propósito e Resultado em Abstracts de Artigos Científicos: Aplicações no Ensino de EAP e na Construção de Ferramentas de Suporte à Escrita Científica. Editora Mercado de Letras, Porto Alegre. Citado na página 5.
- De Felice, R. (2008). *Automatic error detection in non-native English*. University of Oxford. Citado nas páginas 8, 105, 108, 109, 110 e 121.
- De Melo, G. e Weikum, G. (2009). Extracting sense-disambiguated example sentences from parallel corpora. In *Proceedings of the 1st Workshop on Definition Extraction*, WDE '09, páginas 40–46, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 47, 51, 67 e 68.
- De Paiva, V., Rademaker, A., e de Melo, G. (2012). Openwordnet-pt: An open brazilian wordnet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics*, páginas 353–359. Citado nas páginas 186 e 253.

- Deng, Y. e Byrne, W. (2005). Hmm word and phrase alignment for statistical machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, páginas 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 51 e 56.
- Di Felippo, A. e Dias da Silva, B. C. (2007). Towards an automatic strategy for acquiring the wordnet.br hierarchical relations. *Proceedings of the 5th Workshop in Information and Human Language Technology*, 12:1717–1720. Citado na página 185.
- Dicknson, M. (2010). Generating learner-like morphological errors in russian. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-10)*, páginas 259–267, Beijing. Citado nas páginas 8, 111 e 188.
- Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment*, 5(1):5–36. Citado na página 99.
- Dos Santos, V. (2006). Vocabulário e leitura: A elaboração de uma lista de palavras de uso acadêmico em português do brasil. Dissertação de Mestrado, Uiversidade Estadual de Campinas, Instituto de Estudos da Linguagem. Citado nas páginas 85 e 86.
- Dts-Software-Ltda (1995). *Manual do revisor DTS Windows*. Rio de Janeiro:DTS. Citado na página 114.
- Duran, M. (2008a). Customização de corretores ortográficos para aprendizes de línguas estrangeiras. In *Anais do VII Encontro de Linguística de Corpus*, São José do Rio Preto, São Paulo-SP. Citado nas páginas 5, 9, 10, 21 e 45.
- Duran, M. (2008b). *Parâmetros para a elaboração de dicionários bilingues de apoio à codificação escrita em línguas estrangeiras*. Tese de Doutorado, Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista, Campus de São José do Rio Preto. Citado na página 188.
- Duran, M. S., Scarton, C., Aluísio, S. M., e Ramisch, C. (2013). Identifying pronominal verbs: Towards automatic disambiguation of the clitic 'se' in portuguese. In *Proceedings of the 9th Workshop on Multiword Expressions*, páginas 93–100, Atlanta, Georgia, USA. Association for Computational Linguistics. Citado na página 71.

- Durão, A. (1999). *Análisis de errores e interlengua de brasileños aprendices de español y españoles aprendices de portugués*. Editora UEL. Universidade Estadual de Londrina. Citado nas páginas xi, 2, 4, 18, 20, 26, 27, 29, 30, 31, 33, 36, 37, 38, 39, 40, 41, 42, 43, 132, 144, 169 e 170.
- Durrani, D., Fraser, A., Schmid, H., e Hoang, H. (2013). Can markov models over minimal translation units help phrase-based smt. In *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Citado na página 73.
- Ellis, N. C., Simpson Vlach, R., e Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and tesol. *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect*, 42(3):375–396. Citado nas páginas 90 e 92.
- Esteban, J., Lorenzo, J., Valderrábanos, A. S., e L., G. (2004). Transtype2 - an innovative computer-assisted translation system. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, páginas 94–97, Barcelona, Spain. Association for Computational Linguistics, Association for Computational Linguistics. TT2. Citado na página 103.
- Evers, A., Finatto, M., e Pasqualini, B. (2011). Corpus de aprendizes de português como língua adicional (pla): Compilação inicial e primeiros resultados. In *18 InPLA - Intercâmbio de Pesquisas em Linguística Aplicada*, 18 InPLA, LAEL, PUCS, São Paulo, SP. Citado nas páginas 2, 25, 28 e 29.
- Fattah, M., Bracewell, D., Ren, F., e Kuroiwa, S. (2007). Sentence alignment using p-nnt and gmm. *Computer Speech Language*, 21(4):594–608. Citado nas páginas 46, 49, 50, 77 e 78.
- FELLBAUM, C., editor (1998). *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA. Citado na página 108.
- Feltrim, V. (2004). *Uma abordagem baseada em córpus e em sistemas de crítica para a construção de ambientes Web de auxílio à escrita acadêmica em português*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Citado nas páginas 5 e 90.
- Feltrim, V., Aluísio, S., e Nunes, M. (2003). Analysis of the rhetorical structure of computer science abstracts in portuguese. In *In: Corpus Linguistics 2003*,

2003, Lancaster, UK. *Proceedings of Corpus Linguistics 2003 (Also as UCREL Technical Report)*, páginas 212–218. Citado na página 87.

Feltrim, V., Pelizzoni, J., Teufel, S., Nunes, M., e ALUÍSIO, S. (2004). Applying argumentative zoning in an automatic critiquer of academic writing. In Bazzan, A. L. C. e Labidi, S., editors, *SBIA*, volume 3171 of *Lecture Notes in Computer Science*, páginas 214–223, conf/sbia/2004. Springer. Citado nas páginas 133, 134 e 135.

Fernandez, P. (2008). Translating formulaic expressions in instruction manuals: A corpus study. *Newcastle Working Papers in Linguistics*, 14:51–60. Citado nas páginas 94 e 95.

Fernández, S. (2005). *Los Marcadores Discursivos en la Argumentación Escrita: Estudio Comparado en el Español de España y en el Portugués de Brasil*. Tese de Doutorado, Universidade de Salamanca. Departamento de Lengua Española. Citado nas páginas 8, 86, 87, 88, 89, 213 e 242.

Ferraro, G., Nazar, R., M., A. R., e Wanner, L. (2014). Towards advanced collocation error correction in spanish learner corpora. *Language Resources and Evaluation*, 48(1):45–64. Citado nas páginas 111 e 112.

Fossati, D. e Di Eugenio, B. (2007). A mixed trigrams approach for context sensitive spell checking. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 4394 of *Lecture Notes in Computer Science*, páginas 623–633. Springer. Citado nas páginas 106 e 159.

Fossati, D. e Di Eugenio, B. (2008). I saw tree trees in the park: How to correct real-word spelling mistakes. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC 2008*, páginas 897–901. Citado nas páginas 106 e 159.

Foster, J. e Oistein, A. (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*, páginas 82–90, Boulder, Colorado, USA. Citado nas páginas 8, 111 e 188.

Foster, J. e Vogel, C. (2004). Parsing ill-formed text using an error grammar. *Artificial Intelligence Review*, 21:269–291. Citado nas páginas 105, 108 e 109.

Frey, B. e Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814):972–976. Citado na página 216.

- Frunza, O. e Inkpen, D. (2009). Identification and disambiguation of cognates, false friends, and partial cognates using machine learning techniques. *International Journal of Linguistics*, 1(1):1–37. Citado nas páginas 77, 79, 81, 176 e 194.
- Fung, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, páginas 236–243, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 68 e 76.
- Fung, P. (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Parallel Text Processing*, páginas 1–17. Springer. Citado nas páginas 46, 51 e 68.
- Furtado, M. M. (2012). A afinidade das línguas portuguesa e espanhola: estratégias de ensino/ aprendizagem dos falsos cognatos. Dissertação de Mestrado, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa. Citado na página 78.
- 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. Citado nas páginas 2, 3 e 31.
- Gagner, J., Foster, J., e Van Genabith, J. (2007). A comparative evaluation on deep and shallow approaches to the automatic detection of common grammatical errors. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'2007)*, páginas 112–121. Citado nas páginas 105, 108, 109 e 110.
- Gale, W. e Church, K. (1991). A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, páginas 177–184. Citado nas páginas 49 e 50.
- Gale, W. e Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, páginas 77–102. Citado nas páginas 49 e 50.
- Gale, W. A., Church, K. W., e Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439. Citado na página 106.

- Gamallo, G. e García, M. (2012). Extraction of bilingual cognates from wikipedia. In *International Conference on Computational Processing of the Portuguese Language PROPOR 2012*, páginas 63–72. Citado na página 84.
- Gamon, M., Gao, J., Brockett, C., Klementiev, A., B. Dolan, W., Belenko, D., e Vanderwende, L. (2008). Using contextual speller techniques and language modeling for esl error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*. Citado nas páginas xi, 8, 99, 120 e 224.
- Gamon, M., Leacock, C., Brockett, C., Dolan, W., Gao, J., Belenko, D., e Klementiev, A. (2009). Using statistical techniques and web search to correct esl errors. *Calico Journal*, Vol 26, No. 3, páginas 491–511. Citado nas páginas 63, 99 e 110.
- Gao, Q. e Vogel, S. (2008). Parallel implementations of word alignment tool. In *Proc. of the ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*, páginas 49–57. Citado na página 57.
- Gardner, D. e Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, páginas 1–30. Citado na página 85.
- Gaussier, E., Renders, J., Matveeva, I., Goutte, C., e Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of ACL-04*, páginas 527–534. Citado na página 68.
- Genoves, L., Lizotte, R., Schuster, E., Dayrell, C., e 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, 2007, Borovets*, páginas 225–239, Proceedings of the RANLP-2007. Citado nas páginas 8, 20, 24, 28 e 105.
- Golding, A. R. e Roth, D. (1999). A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3):107–130. Citado na página 107.
- 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. Citado nas páginas 2, 3 e 31.
- Gomes, L. (2009). Parallel texts alignment. Dissertação de Mestrado, Universidade Nova de Lisboa, Lisboa. Citado na página 78.

- Gonçalves, D. (2013). A interferência linguística no uso de heterossemânticos por brasileiros aprendizes de espanhol. *Revista de Letras, Goiânia*, 2(2):50–70. Citado na página 78.
- Granger, S. (1998). The computer learner corpus: A versatile new source of data for sla research. Teubert, W., *Corpus Linguistics: Critical Concepts in Linguistics, Routledge*, páginas 166 – 182. Citado nas páginas 26, 28 e 29.
- Granger, S. (2004). Computer learner corpus research: Current status and future prospects. *Language and Computers*, 52:123–145. Citado nas páginas 18, 19, 20 e 26.
- Granger, S., Kraif, O., Ponton, C., Antoniadis, G., e Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, 19(3):252–268. Citado na página 19.
- 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*, páginas 1–31, PUC-Rio, Rio de Janeiro, Brasil. Citado nas páginas xi, 2, 3, 4, 18, 20, 26, 28, 29, 30, 31, 33, 34, 35, 36, 42, 132, 144, 169 e 170.
- Gregghi, J., Martins, R., e Nunes, M. (2002). Diadorim - a lexical database for brazilian portuguese. In *LREC*, páginas 1346–1350. European Language Resources Association. Citado na página 185.
- Han, N., Chodorow, M., e Leacock, C. (2006). Detecting errors in english article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129. Citado na página 109.
- Hasegawa, R., Martins, R., e Nunes, M. (2002). Regra 2002: Características e desempenho. Relatório técnico, Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil. Citado nas páginas 5, 114, 116, 117 e 155.
- Hayashibe, Y., Hagiwara, M., e Sekine, S. (2012). phloat: Integrated writing environment for esl learners. In *Proceedings of Second Workshop on Advances in Text Input Methods (WTIM 2)*, páginas 25–30. Citado nas páginas 47, 97, 98, 100 e 101.

- 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. Citado nas páginas 2, 3, 4, 31 e 32.
- Hermet, M. e Désilets, A. (2009). Using first and second language models to correct preposition errors in second language authoring. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, páginas 64–72, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 8, 62 e 110.
- Hoflan, K. (1996). A program for aligning english and norwegian sentences. In HOCKEY, S.; IDE, N.; PERISSINOTTO, G. (eds). *Research in Humanities Computing*, páginas 165–178, Oxford: Oxford University. Citado na página 51.
- Hu, Y., Lu, R., Li, X., Chen, Y., e Duan, J. (2007). A language modeling approach to sentiment analysis. In Shi, Y., van Albada, G., Dongarra, J., e Sloot, P., editors, *Computational Science - ICCS 2007*, volume 4488 of *Lecture Notes in Computer Science*, páginas 1186–1193. Springer Berlin Heidelberg. Citado na página 63.
- Hua chen, M., Huang, C.-C., Huang, S.-T., e Chang, J. S. (2010). Grasp: Grammar- and syntax-based pattern-finder for collocation and phrase learning. In Ootoguro, R., Ishikawa, K., Umemoto, H., Yoshimoto, K., e Harada, Y., editors, *PACLIC*, páginas 357–364, conf/paclic/2010. Institute for Digital Enhancement of Cognitive Development, Waseda University. Citado nas páginas 105 e 112.
- Huang, C., Gao, J., L, T., e Ming, Z. (2000). Pens: A machine-aided english writing system for chinese users. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Citado nas páginas 47, 100, 101 e 239.
- Huang, C., Yang, P., Hua Chen, M., Ting Hsieh, H., Hui Kao, T., e Chang, J. S. (2012). Transahead: A writing assistant for cat and call. In Daelemans, W., Lapata, M., e Màrquez, L., editors, *EACL*, páginas 16–19. The Association for Computer Linguistics. Citado na página 103.
- Indurkha, N. e Damerau, F. (2010). *Handbook of natural language processing*. Chapman & Hall/CRC, 2nd^a edição. Citado nas páginas 46, 47, 49 e 51.
- Informática, L. (1999). *Gramática eletrônica*. Lexicon Informática Ltda. Citado na página 114.

- Inkpen, D., Frunza, O., e Kondrak, G. (2005). Automatic identification of cognates and false friends in french and english. In *Proceedings of the international conference on recent advances in natural language processing*, (RANLP-05), páginas 251–257. Citado na página 81.
- Jalilifar, A. (2008). Discourse markers in composition writings: The case of iranian learners of english as a foreign language. *Journal of CCSE, English Language Teaching*, 1(2):196–202. Citado na página 86.
- James, C. (1998). *Errors in language learning and use: Exploring error analysis*. London: Longman. Krishnamurthy, R. & Kosem, I. Citado na página 9.
- Johnson, M., Griffiths, T. L., e Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19*. Citado na página 63.
- Jurafsky, D. e Martin, J. H. (2009). *Speech and language processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. Citado nas páginas 57, 58, 61, 62, 65, 66 e 129.
- Kay, M. e Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, 19(1):121–142. Citado na página 49.
- Kermes, H. (2012). A methodology for the extraction of information about the usage of formulaic expressions in scientific texts. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., e Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, páginas 2064–2068, Istanbul, Turkey. European Language Resources Association (ELRA). Citado nas páginas 93 e 94.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., e Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, páginas 177–180. Citado nas páginas 61 e 177.
- Koehn, P. e Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition - Volume 9*, ULA '02, páginas 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 68, 76, 78 e 176.

- Koehn, P., Och, F. J., e Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, páginas 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 59.
- Kondrak, G. (2001). Identifying cognates by phonetic and semantic similarity. In *Proceedings of NAACL 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, páginas 103–110. Citado na página 82.
- Kondrak, G. (2005). Cognates and word alignment in bitexts. In *Proceedings of the Tenth Machine Translation Summit*, (MT Summit X), páginas 305–312. Citado nas páginas 79 e 80.
- Kondrak, G. e Dorr, B. (2004). Identification of confusable drug names: A new approach and evaluation methodology. In *Proceedings of COLING 2004*, páginas 952–958. Citado nas páginas 78, 79, 81 e 176.
- Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4):377–439. Citado na página 106.
- Landis, J. R. e Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. Citado nas páginas xi, 23 e 24.
- Langlais, P., Simard, M., e Véronis, J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98, páginas 711–717, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 46 e 52.
- Larsen-Freeman, D. e Long, M. H. (1992). *An introduction to second language acquisition research*. London: Longman. Citado na página 7.
- Leacock, C., Chodorow, M., Gamon, M., e Tetreault, J. (2010). *Automated grammatical error detection for language learners*. Morgan and Claypool Publishers. Citado nas páginas 6, 7, 8, 23, 104, 105, 118, 119 e 154.
- Leacock, C., Gamon, M., e Brockett, C. (2009). User input and interactions on microsoft research esl assistant. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP

- '09, páginas 73–81, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 98.
- Lee, D. e Swales, J. (2006). A corpus-based eap course for nns doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25:56–75. Citado na página 67.
- Leedham, M. (2012). Combining intuition with corpus linguistic analysis: A study of marked lexical chunks in four chinese students' undergraduate assignments. *Nordic Journal of English Studies*, 11(3):155–187. Citado nas páginas 8, 90 e 92.
- Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, página 707. Citado nas páginas 62, 106, 107 e 223.
- Lewis, M. (1993). The lexical approach. *England: Language Teaching Production*. Citado nas páginas 9 e 69.
- Li, B., Gaussier, E., e Aizawa, A. (2011). Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, páginas 473–478, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 46, 51 e 68.
- Lins, R. D., Camelo, H. A. L., e Moura, R. S. (1999). Um sos para a língua portuguesa. In *IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR-99)*, páginas 115–128. Citado na página 114.
- Liu, A. L., Wible, D., e Tsao, N. (2009). Automated suggestions for miscollocations. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, páginas 47–50, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 112.
- Loucky, J. (2013). Using computerized bilingual dictionaries to help maximize english vocabulary learning at japanese colleges. *CALICO Journal*, 21(1):105–129. Citado nas páginas 10 e 45.
- Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317. Citado na página 207.

- Machado, D. J. (2009). Classificação de papéis retóricos baseada em expressões automaticamente extraídas de textos científicos. Relatório técnico, Universidade Estadual de Maringá, Brasil. Citado na página 91.
- Mackay, W. e 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*, CONLL '05, páginas 40–47, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 80.
- Mann, G. e Yarowsky, D. (2001). Multipath translation lexicon induction via bridge languages. In *NAACL*, páginas 1–8. Citado na página 78.
- Marquiafável, S. (2007). Avaliação automática da qualidade de escrita de resumos científicos em inglês. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Citado na página 99.
- Martins, R., Hasegawa, R., Nunes, D., Montilha, G., e De oliveira, J. (1998). Linguistic issues in the development of regra: A grammar checker for brazilian portuguese. *Natural Language Engineering*, 4(4):287–307. Citado nas páginas 5, 114 e 155.
- Mays, E., Damerau, F. J., e Mercer, R. L. (1991). Context based spelling correction. *Information Processing & Management*, 27(5):517–522. Citado na página 159.
- Maziero, E., Pardo, T., Di Felippo, A., e D da Silva, B. (2008). A base de dados lexical e a interface web do tep 2.0: Thesaurus eletrônico para o português do brasil. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, WebMedia '08, páginas 390–392, New York, NY, USA. ACM. Citado na página 185.
- McDonough, S. (2011). The psychology of second language acquisition. *ELT Journal*, 65(2):194–196. Citado na página 93.
- Melamed, I. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25:107–130. Citado nas páginas 78 e 80.
- Melamed, I. (2000). Pattern recognition for mapping bitext correspondence. In VÉRONIS, J. (ed.). *Parallel text processing: Alignment and use of translation corpora*, páginas 25–47. Kluwer Academic Publishers. Citado na página 50.

- Melamed, I. D. (1996). Automatic construction of clean broad-coverage translation lexicons. *CoRR*, cmp-lg/9607037:1–10. Citado na página 67.
- Meurers, D. (2013). Natural language processing and language learning. In Chapelle, C. A., editor, *Encyclopedia of Applied Linguistics*, páginas 4193–4205. Blackwell. Citado nas páginas 6, 7 e 98.
- Mitkov, R., Pekar, V., Blagoev, D., e Mulloni, A. (2007). Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21:29–53. Citado nas páginas 79, 81, 82 e 83.
- 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)*, páginas 372–387. Citado nas páginas 2, 3, 4, 31, 32 e 42.
- Morin, E. e Prochasson, E. (2011). Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, páginas 27–34, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 71, 72 e 75.
- Morley, J. (2015). Phrasebank: A university-wide online writing resource. Relatório técnico, School of Languages, Linguistics and Cultures, The University of Manchester. Citado nas páginas 85, 94 e 102.
- Moura, R. S. (1996). Sos: um sistema orto-sintático para a língua portuguesa. Dissertação de Mestrado, Universidade Federal de Pernambuco, Recife. Citado na página 114.
- Mulloni, A., Pekar, V., Mitkov, R., e Blagoev, D. (2007). Semantic evidence for automatic identification of cognates. In *Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons, Borovets, Bulgaria*, páginas 49–54. Citado nas páginas 194 e 200.
- Nagata, N. (2013). Robo-sensei's nlp-based error detection and feedback generation. *CALICO Journal*, 26(3):562–579. Citado na página 6.
- Nagata, R., Whittaker, E., e Sheinman, V. (2011). Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language*

- Technologies - Volume 1*, HLT '11, páginas 1210–1219, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 105.
- Nakov, S. (2009). Automatic identification of false friends in parallel corpora: statistical and semantic approach. *Serdica Journal of Computing*, 3(2):133–158. Citado nas páginas 77, 79, 81 e 83.
- Nattinger, J. R. e Decarrico, J. D. (1992). *Lexical phrase and language teaching*. Oxford: Oxford University Press. Citado na página 7.
- Nerbonne, J. (2000). Parallel texts in computer-assisted language learning. In Veronis, J., editor, *Parallel Text Processing*, páginas 354–369. University of Groningen, The Netherlands. Citado na página 47.
- Nóbrega, F. e Pardo, T. (2014). General purpose word sense disambiguation methods for nouns in portuguese. In *Computational Processing of the Portuguese Language - 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil, October 6-8, 2014. Proceedings*, páginas 94–101. Citado nas páginas 201 e 253.
- Nunes, M. G. V., Vieira, F. M. C., Zavaglia, C., Sossolote, C. R. C., e Hernandez, J. (1996). A construção de um léxico para o português do Brasil: Lições aprendidas e perspectivas. In *Anais do II Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR)*, páginas 61–70. Citado nas páginas 5, 114, 155, 209 e 223.
- Och, F. e Ney, H. (2000a). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, páginas 1086–1090, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 178.
- Och, F. e Ney, H. (2000b). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, páginas 440–447. Citado na página 57.
- Och, F. J. e Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51. Citado nas páginas 72 e 178.
- O'Donnell, M. B., Römer, U., e Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, 18(1):83–108. Citado nas páginas 90 e 92.

- Oparin, I. (2008). *Language models for automatic speech recognition of inflectional languages*. Tese de Doutorado, University of West Bohemia, Faculty of Applied Sciences. Citado na página 62.
- Paiva, J., Schwastz, W., Pedrini, H., e Minghim, R. (2012). Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data. *Computer Graphics Forum*, 31(3pt4):1345–1354. Citado nas páginas 135, 138 e 142.
- Papageorgiou, H., Cranias, L., e Piperidis, S. (1994). Automatic alignment in parallel corpora. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ACL '94*, páginas 334–336, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 50.
- Pardo, T. e Nunes, M. G. (2006). Review and evaluation of dizer - an automatic discourse analyzer for brazilian portuguese. In *Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006, Proceedings*, páginas 180–189. Citado nas páginas 87 e 213.
- Pinheiro, G. M., Rino, L. H. M., Martins, R. T., Di Felippo, A., Fillié, V. M., Hasegawa, R., e Nunes, M. G. V. (2001). Trasem: A investigação empírica sobre o problema da ambigüidade categorial. Relatório técnico, Série de Relatórios do NILC. NILC-TR-01-2, Abril 2001, 23p. Citado nas páginas 5, 114 e 155.
- Piperidis, S., Papageorgiou, H., e Boutsis, S. (2000). From sentences to words and clauses. In VÉRONIS, J. (ed.). *Parallel text processing: Alignment and use of translation corpora.*, páginas 117–138. Kluwer Academic Publishers. Citado na página 50.
- Pravec, N. (2002). Survey of learner corpora. *ICAME Journal*, 26:81–114. Citado nas páginas 26, 28 e 29.
- Prentice, M. (2012). A method for extracting formulaic sequences from a student corpus. Relatório técnico, Departmental Bulletin Paper. Citado nas páginas 92, 93 e 232.
- Ramisch, C. (2012). A generic framework for multiword expressions treatment: From acquisition to applications. In *Proceedings of the ACL 2012 Student Research Workshop*, páginas 61–66. Citado nas páginas 69, 71, 75 e 90.

- Ratnaparkhi, A. (1996). A maximum entropy part of speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, páginas 133–142. Citado na página 204.
- Ribeiro, M. (2012). Idéia ou ideia? *Folheie a edição especial de 50 anos da FAPESP*. Citado na página 114.
- Rojas, J. (2006). Processo de fossilização na interlíngua de hispanofalantes aprendizes de português no brasil: acomodação consentida? Dissertação de Mestrado, Programa de Mestrado em Linguística Aplicada do Departamento de Línguas Estrangeiras e Tradução da Universidade de Brasília. Brasília. Citado nas páginas 3, 18, 20, 25, 26, 28, 29, 30, 33, 132 e 145.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., e Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, páginas 1–15, London, UK, UK. Springer-Verlag. Citado na página 69.
- 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*. Citado nas páginas 3, 18, 19 e 31.
- Scaramucci, M. e Rodrigues, M. (2004). Compreensão (oral e escrita) e produção escrita no exame celpe-bras: análise do desempenho de candidatos hispanofalantes. In *Português para falantes de espanhol: artigos selecionados escritos em português e inglês, 2004*. In: SIMÕES, A. R. M. et al. (Org.Ed.). Citado nas páginas 3 e 32.
- Scarton, C. E. e Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: Adaptando as métricas do coh-metrix para o português. *Linguamática (Revista para o Processamento Automático das Línguas Ibéricas)*, 2:45–61. Citado na página 137.
- Schmid, H. (1995). Improvements in part of speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, páginas 47–50. Association for Computational Linguistics. Citado na página 204.
- Schoffen, J. (2009). *Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no exame Celpe-Bras*. Tese de Doutorado, Universidade Federal do Rio Grande do Sul. Instituto de Letras. Citado na página 1.

- Seno, E. (2010). *Um método para a fusão automática de sentenças similares em português*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Citado na página 63.
- Shaalán, K., Aref, R., e Fahmy, A. (2010). An approach for analyzing and correcting spelling errors for non-native arabic learners. In *The 7th International Conference on Informatics and Systems, INFOS2010*. Cairo University. Citado na página 106.
- Shanru, Y. (2012). Discourse markers? an area of confusion. Relatório técnico, Newcastle University, UK. Citado na página 87.
- Simard, M., Foster, G., e Pierre, I. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, páginas 67–81. Citado nas páginas 50, 77, 78, 79 e 80.
- Simpson-Vlach, R. e Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4):487–512. Citado nas páginas 8, 85, 90, 92, 93, 208 e 232.
- Sobrevilla Cabezudo, M. A. (2015). Investigação de métodos de desambiguação lexical de sentidos de verbos do português do brasil. Dissertação de Mestrado, Dissertação de Mestrado. Instituto de Ciencias Matemáticas e de Computacao, Universidade de Sao Paulo. Citado nas páginas 201 e 253.
- Sridhar, V. K. R., Barbosa, L., e Bangalore, S. (2011). A scalable approach to building a parallel corpus from the web. In *INTERSPEECH*, páginas 2113–2116. ISCA. Citado nas páginas 47 e 222.
- Stehouwer, H. e Van Zaanen, M. (2009). Language models for contextual error detection and correction. In *Proceedings of the EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference, CLAGI '09*, páginas 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 8, 63 e 107.
- Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, páginas 257–286. Citado na página 67.
- Stuker, S., Fay, J., e Berkling, K. (2011). Towards context-dependent phonetic spelling error correction in children's freely composed text for diagnostic and

- pedagogical purposes. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, páginas 1601–1604. Citado na página 107.
- Tagnin, E. (2006). A multilingual learner corpus in brazil. *Language and Computers*, 56(1):195–202. Citado nas páginas 20 e 22.
- Tagnin, S. e Fromm, G. (2011). Comaprend - a experiência da construção de um corpus de aprendizes para estudos. *Domínios de Lingu@gem*, 2(2):1–17. Citado nas páginas 21 e 22.
- Tang, J. (2012). An empirical study on the effectiveness of the lexical approach to improving writing in sla. *Journal of Language Teaching and Research*, 3(3):578–583. Citado na página 9.
- Tetreault, J., Foster, J., e Chodorow, M. (2010). Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers, HLT '11*, páginas 353–358, Uppsala, Sweden. Association for Computational Linguistics. Citado nas páginas 8, 105, 108, 109 e 122.
- Tiedemann, J. (2011). *Bitext alignment*. Number 14 in Synthesis Lectures on Human Language Technologies, Graeme Hirst (ed). Morgan & Claypool Publishers. Citado nas páginas 46, 47, 49, 52 e 56.
- Tiu, E. e Roxas, R. (2008). Automatic bilingual lexicon extraction for a minority target language. In *Proceedings of the 22nd PACLIC (2008)*, páginas 368–376. Citado nas páginas 46, 51, 68 e 125.
- Torres, L., Paiva, J., Minghim, R., e Aluísio, S. (2013). Visual classification of documents based on writing quality. In *Proceedings of X Encontro Nacional de Inteligência Artificial e Computacional, ENIAC 2013*, páginas 1–9. Citado na página 168.
- Torres, L., Rodrigues, R., e Aluísio, S. M. (2014). *Espanhol-Acadêmico-Br: A corpus of academic Portuguese learners produced by native speakers of Spanish*, volume 1, páginas 98–111. Cambridge, England: Cambridge Scholars Publishing. Citado na página 144.
- Torres, S. L. e Aluísio, 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, STIL 2011, páginas 67–76, Cuiabá/MT. Citado nas páginas 146 e 177.

Trnka, K. (2008). Adaptive language modeling for word prediction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop*, HLT-SRWS '08, páginas 61–66, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 63.

Tsao, N.-L. e Wible, D. (2009). A method for unsupervised broad-coverage lexical error detection and correction. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, páginas 51–54, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 8 e 107.

Tschichold, C. (1999). Grammar checking for call: Strategies for improving foreign language grammar checkers. *CALL: Media, Design and Applications*, páginas 203–221. Citado na página 7.

Tsvetkov, Y., Twitto, N., Schneider, N., Ordan, N., Faruqui, M., Chahuneau, V., Wintner, S., e Dyer, C. (2013). Identifying the 11 of non-native writers: The cmu-haifa system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, páginas 279–287. Citado na página 134.

Uzêda, V., Pardo, T., e Nunes, M. (2008). Evaluation of automatic text summarization methods based on rhetorical structure theory. *2008 Eighth International Conference on Intelligent Systems Design and Applications*, páginas 389–394. Citado na página 5.

Van Der Plas, L. e Tiedemann, J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, páginas 866–873, Sydney, Australia. Association for Computational Linguistics. Citado nas páginas 47 e 52.

Vogel, S., Ney, H., e Tillmann, C. (1996). Hmm-based word alignment in statistical translation. In *COLING 96: The 16th Int. Conf. on Computational Linguistics*, páginas 836–841. Citado nas páginas 56 e 178.

Wible, D. e Tsao, N. (2010). Stringnet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL*

- HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, EUCCCL '10, páginas 25–31, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 102.
- Wilcox-O'Hearn, A., Hirst, G., e Budanitsky, A. (2008). Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'08, páginas 605–616. Citado nas páginas 106 e 159.
- Witten, I. e Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. 2nd Edition. Morgan Kaufmann, San Francisco. Citado na página 197.
- Wold, H. (1985). Partial least squares. *Encyclopedia of Statistical Sciences*, 6:581–591. Citado na página 140.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: CUP. Citado na página 89.
- Wu, D. e Xia, X. (1994). Learning an english-chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, páginas 206–213. Citado nas páginas 68 e 76.
- Wu, H. e Zhou, M. (2003a). Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, páginas 72–79, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 71.
- Wu, H. e Zhou, M. (2003b). Synonymous collocation extraction using translation information. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, páginas 120–127, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado na página 187.
- Wu, Z. e Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, páginas 133–138. Citado na página 82.

- ZarrieB, S. e Kuhn, J. (2009). Exploiting translational correspondences for pattern-independent mwe identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, páginas 23–30, Stroudsburg, PA, USA. Association for Computational Linguistics. Citado nas páginas 72, 75 e 76.
- Zorraquino, M. e Portolés, L. (1999). *Los Marcadores del Discurso*, páginas 4051–4213. I. Bosque & V. Demonte (Eds.). Citado na página 88.

Medidas de Associação

Probabilidade Conjunta	$P(xy)$
Probabilidade Condicional	$P(y x)$
Pointwise mutual information	$\log\left(\frac{P(xy)}{P(x*)P(*y)}\right)$
Teste T	$\frac{f(xy) - \hat{f}(xy)}{\sqrt{f(xy)(1 - (\frac{f(xy)}{N}))}}$
Log likelihood ratio	$-2 \sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$
Coeficiente Dice	$\frac{2*P(xy)}{P(x*) + P(*y)}$
Jaccard	$\frac{a}{abc}$

$a = f(xy)$	$b = f(x\bar{y})$	$f(x*)$
$c = f(\bar{x}y)$	$d = f(\bar{x}\bar{y})$	$\bar{x}*$
$f(*y)$	$f(*\bar{y})$	N

A tabela de contingência mostra a frequências observadas e as relativas para os bigramas xy ; $\bar{w}$ representa qualquer palavra exceto w ; $*$ significa qualquer palavra; N é o número total de bigramas. Em algumas formulas as células da tabela de contingência foram referenciadas como f_{ij} . A frequência esperada utilizando a tabela de contingência é calculada como:

$$\hat{f}(xy) = f(x*)f(*y)/N.$$

Questionário Hispano Falantes

Principais problemas que nativos do espanhol enfrentam quando escrevem em português

*Obrigatório

Nome Completo:

Sexo: *

Nacionalidade: *

Curso *

Já você fez algum curso da língua portuguesa? Quantos anos? *

Segundo sua experiência prática, quais são os principais problemas que enfrenta quando escreve em português?

A seguir, enumeramos um conjunto de tipos de erros, com seus exemplos mais característicos, que hispanofalantes cometem quando escrevem em português.

Cada exemplo foi especificado considerando o erro e a forma correta como deve ser escrito (erro –> correção). Quando houver mais de um exemplo, eles estão separados por |||.

Cada tipo de erro que você considera realmente um problema na sua escrita em português, selecione "Sim".

1- Acentuação ortográfica equivocada

...formação economica... -> ...formação econômica...

☐ Sim

2- Uso da crase (Ausência ou uso inadequado)

...à meu entender... -> ...a meu entender...

☐ Sim

3- Marca equivocada de nasalização

...pessoas que vam fazer... -> ...pessoas que vão fazer...

☐ Sim

4- Separação e união de palavras (Palavras compostas)

(a) super resistente -> super-resistente ||| (b) arco íris -> arco-íris

☐ Sim

5- Separação e união de palavras (União de pronomes complementos ao verbo)

...é difícil darse conta... -> ...é difícil dar-se conta... ||| estudarlo -> estudá-lo

☐ Sim

6- Confusão de fonemas

...Tenho certeza do resultado sim nenhuma dúvida... -> ...Tenho certeza do resultado sem nenhuma dúvida...

☐ Sim

7- Confusão de grafemas para o mesmo fonema

...debido a... -> ...devido a...

☐ Sim

8- Uso de "falsos amigos" (palavras que existem nas duas línguas, mas com significados diferentes)

o rol do professor é importante na sociedade -> o papel do professor é importante na sociedade

☐ Sim

9- Uso de um vocabulário informal para o gênero científico

(a) ... se tem um monte de estruturas... -> tem-se um grande número de estruturas ||| (b) Juntar as características... -> Integrar as características...

☐ Sim

10- Mudanças de gênero

... as costumes e ideias ... -> ... os costumes e ideias ...

☐ Sim

11- Uso incorreto de artigos (Ausência, Presença)

(a) De todos os países só Brasil fala ... -> De todos os países só o Brasil fala ... ||| (b) Com a fronteira de Argentina ... -> Com a fronteira da Argentina ...

☐ Sim

12- Conjugações verbais

Se não foram duas línguas... -> Se não fossem duas línguas ...

☐ Sim

13- Uso incorreto de preposições (Erros de classes gramaticais)

(a) ... eu estive em Inglaterra ... -> ... eu estive na Inglaterra ... ||| (b) ... mas em este caso ... -> ... mas neste caso ...

☐ Sim

14- Problema da ordem das palavras e expressões

Além disso, a precisão pode-se aprimorar obtendo... -> Além disso, pode-se aprimorar a precisão obtendo ...

☐ Sim

15- Uso de palavras em Espanhol no texto

Falam solamente do tempo ... -> Falam somente do tempo ...

☐ Sim

Aproximadamente quantas revisões do texto são necessárias até ele ficar com a qualidade desejada? *

Comentários, críticas e sugestões:

Faça sua manifestação para que possamos compreender melhor as dificuldades na escrita do Português e conseguir criar um sistema focado nestas deficiências.

Enviar

Manual de Anotação

Manual de Anotação de Erros de hispano falantes e nativos do português com uso da ferramenta BRAT

O presente manual descreve a tarefa de anotação de diversas categorias de erros produzidos por hispano falantes aprendizes de português e por nativos do português. A razão de anotar os dois tipos de erros é que há um conjunto de erros cometidos tanto por hispano falantes aprendizes e por nativos do português. O processo de anotação é guiado por uma tipologia de erros previamente definida. A quantidade/diversidade dos erros identificados nos textos será utilizada para avaliar a qualidade da escrita de resumos de pesquisas científicas. Para guiar a avaliação dos textos foi definida uma rubrica de avaliação que é apresentada neste manual.

A partir dos aspectos considerados na anotação, será possível comparar a produção dos nativos e dos aprendizes de português, conhecendo algumas diferenças e/ou similaridades na escrita de textos científicos em português. Além disso, esta anotação vai permitir analisar as interferências da língua materna, o espanhol, na aprendizagem de uma segunda língua, o português. Finalmente, por meio desta anotação serão reconhecidos os erros ou pontos críticos que um aprendiz de espanhol enfrenta quando escreve ou fala em português.

1- Características dos textos que serão anotados

Os textos que serão anotados são resumos acadêmicos, produzidos por hispano falantes e nativos de português. Esses resumos foram produzidos por alunos matriculados em programas de pós-graduação das universidades brasileiras ou foram publicados em revistas científicas. Além disso, eles foram construídos utilizando um editor de texto que auxilia na correção de erros gramaticais de diversas categorias. Considerando os fatos anteriores poder-se-ia pensar que são poucos os erros presentes nos resumos, porém essa inferência não é verdadeira porque há múltiplos erros produzidos pelas interferências da língua espanhola que não são cobertos pelos corretores gramaticais mais utilizados. Além disso, ainda há erros produzidos pelos nativos que os corretores não conseguem identificar, muitas vezes estes erros são produzidos por descuidos ou presa na hora da escrita e embora o corretor aponte o erro o aluno não aceita as sugestões oferecidas para corrigir o problema.

O que é importante na explicação anterior é que é necessário “**prestar muita atenção**” na hora de analisar os resumos.

“Leia o texto antes da anotação”.

Não tente entender a pesquisa relatada. Concentre-se em identificar possíveis aspectos que afetem a qualidade da escrita. É importante ter uma visão geral do texto antes de focar em elementos específicos. Ao mesmo tempo, é importante fazer uma leitura detalhada de cada sentença, porque os hispano falantes produzem uma interlíngua que dada à similaridade entre o português e o espanhol é difícil identificar a fronteira do erro. Para avaliar a qualidade da escrita serão considerados os aspectos apresentados na rubrica de avaliação (Tabela 2).

2- Resumos selecionados

Um total de 100 resumos foi selecionado para anotação. Deste total, 60 resumos foram produzidos por hispano falantes e 40 por nativos do português. Na Tabela 1, são apresentados os resumos selecionados para anotar, assim como a fonte da qual eles foram extraídos.

	Nativos	Aprendizes
Resumos Revista	Computação Aplicada → 13 Informática na Educação → 13	Revista Educação → 15 Revista <i>Interciencia</i> → 16
Resumos Tese/Dissertações	SciPo → 14	Corpus HABLA → 29
Total Textos	40	60

Tabela 1: Resumos acadêmicos selecionados para anotação.

3- Rubrica de Avaliação

A anotação dos textos é guiada pela rubrica de avaliação, apresentada na Tabela 2. Esta rubrica foi construída seguindo as categorias de avaliação do exame CELPEBras e foca em avaliar a qualidade de textos acadêmicos seguindo vários critérios. Os erros etiquetados nesta anotação foram agrupados nas categorias da rubrica da Tabela 2.

Rubrica de Textos Científicos			
Nível	3-Avançado	2-Intermediário	1- Iniciante
Clareza e Coesão (Este nível sofre influência do tamanho da oração)	Texto bem desenvolvido → Pode ter até 1 simples deslize (Não deve ter nenhum problema de compreensão, mas se não for muito grave pode ser aceito 1)	Texto bem desenvolvido com poucos problemas → Pode ter até 2 deslizes (Pode ter um problema grave. Se o texto for extenso podem ser aceitos até 2 sentenças com problemas simples)	Texto pouco/mal desenvolvido → Até 3 problemas graves → 3 (Se o texto é pequeno não é aceito nem sequer uma sentença com problemas de coerência ou clareza)
Adequação Lexical e Gramatical (O mesmo erro é contabilizado uma única vez)	Algumas Inadequações → 1-2 (Pode ter até 2 erros. Claramente, depende do erro. Erros de ortografia por exemplo.	Inadequações Frequentes → 3-4 (Pode ter até 3 ou 4 erros de ortografia)	Inadequações Frequentes → 5 (Tem 5 ou mais erros)
Interferências da língua nativa (São os erros mais graves, porque o aluno não consegue diferenciar em que língua está escrevendo)	Adequado O texto não tem nenhuma interferência da língua nativa	Poucas Inadequações → 1-2 (Uma ou duas do mesmo tipo) Se o texto tiver duas interferências de tipos diferentes, já é um texto de “Iniciante”	Inadequações Frequentes → 2 (2 ou mais interferências no texto)

Tabela 2: Rubrica de avaliação dos resumos acadêmicos produzidos por hispano falantes aprendizes e nativos do português.

4- Ferramenta de anotação

Para a anotação dos erros vai ser utilizada a ferramenta web “Brat”. Esta ferramenta auxilia o processo de anotação de textos, permitindo adicionar etiquetas aos textos já existentes. A Figura 1 mostra a tela inicial da ferramenta e nas seções seguintes serão mostradas outras telas do Brat¹.

¹<http://143.107.183.175:12380/brat/index.xhtml/#/>

Figura 1: Tela inicial da ferramenta Brat.

A ferramenta permite personalizar um processo de anotação, a partir de uma tipologia definida pelo usuário e envolvendo vários anotadores. A Figura 2 mostra a tipologia de erros que será utilizada neste processo de anotação.

5- Processo de anotação

Na Figura 2 são apresentadas as categorias de anotação definidas na ferramenta Brat para anotar os resumos acadêmicos. Como mostrado na figura, o conjunto de etiquetas está agrupado em 3 categorias que representam os critérios de avaliação e em uma última categoria que é a avaliação geral (ou classe) dos textos. Para diminuir a complexidade da tarefa de anotação sugiro que cada categoria de erro seja anotada em turnos diferentes. Por exemplo, após uma primeira leitura do texto, devem ser analisadas as sentenças tentando identificar erros da categoria “Clareza-Coesao”. Após anotar a primeira categoria, novamente o avaliador deve ler o texto para reconhecer os erros da categoria seguinte, “Erros-Gramaticais”. Finalmente, em uma última leitura do resumo devem ser etiquetados os erros da categoria “Interferencias”.

Figura 2: Tipologia especificada na ferramenta Brat para anotar os resumos.

Na Figura 2, a primeira categoria “Clareza-Coesao” refere-se ao nível “Clareza e Coesão” da rubrica apresentada na Tabela 2. Os erros desta categoria, assim como os erros da segunda categoria, “Erros-Gramaticais” são erros que podem ser cometidos tanto por nativos quanto por aprendizes. Desta vez, a categoria “Erros-Gramaticais” refere-se ao nível de avaliação “Adequação Lexical e Gramatical”, apresentado na rubrica. Comumente, os aprendizes cometem erros gramaticais devido às interferências da língua espanhola, mas nem sempre essa é a natureza destes erros. Além disso, como estes erros também são encontrados em textos produzidos pelos nativos, foi criada a categoria “Erros-Gramaticais” para agrupar estes erros. Finalmente, os erros de “Interferências da Língua Nativa”, na rubrica, são considerados na categoria “Interferências”. Veja que cada categoria de erros aparece com uma cor diferente na tipologia de anotação.

A seguir são mostrados exemplos da anotação de cada uma das categorias de erros.

6- Anotação das categorias de erros

Clareza e Coesão

Nesta tarefa de anotação não será avaliada a clareza e a coesão ao nível do texto. Em textos científicos esta avaliação é feita considerando a adequação ao gênero do texto. Como não será considerado este nível de avaliação, somente avaliaremos a clareza e coesão no nível sentencial. A avaliação de clareza e coesão é uma das etapas mais subjetivas dentro da anotação. Tentando diminuir a subjetividade na anotação deste quesito, a seguir são listadas algumas dicas que podem indicar que uma sentença tem problemas de coerência ou de clareza das ideias.

- ✓ é necessário ler várias vezes a sentença para compreender a ideia do autor;
- ✓ em geral:
 - ✓ sentenças muito longas são difíceis de ler;
 - ✓ sentenças com poucos sinais de pontuação são difíceis de ler;
 - ✓ sentenças sem marcadores discursivos dificultam o entendimento da mensagem.

Por outro lado, o uso incorreto de marcadores discursivos também pode produzir falta de coerência dentro da sentença. A maioria das vezes fazendo uma troca do marcador utilizado melhora a coerência da sentença. A seguir apresentam-se exemplos de erros identificados nesta etapa. Veja que a categoria do erro, apresentado na tipologia aparece em amarelo.

Categoria Clareza-Coesao → **Incoerente** → Para etiquetar este tipo de erros, na ferramenta Brat, toda a sentença deve ser etiquetada.

De fato, SLSM constituem uma classe de sistemas lineares estocásticos que são relevantes para aplicações (vide por exemplo algumas referências) e apresentam um número de características paralelas àquelas dos sistemas lineares determinísticos, para uma abordagem com teoria de operadores para SLSM, como uma ilustração de trabalhos recentes e artigos inéditos; porém, controlabilidade foi considerada desde um ponto de vista único da trajetória.

Incoerente

número de características paralelas àquelas dos sistemas lineares determinísticos, para uma abordagem com teoria de operadores

Figure 3: Exemplo de sentença incoerente

O exemplo da Figura 3 mostra uma sentença muito longa, na qual é difícil entender a ideia que o autor quer transmitir. Acredita-se que este tipo de problemas seja comum tanto para os nativos, quanto para os aprendizes. Por outro lado, oferecer automaticamente sugestões para esse problema não é uma tarefa trivial, embora seja fácil automaticamente inferir que sentenças longas, sem conectivos e com falta de

símbolos de pontuação sejam candidatas a terem problemas de coerência. Seguindo essa linha, poder-se-ia pensar que não é necessário anotar essa categoria de erros, porém é importante fazer essa anotação, para avaliar se realmente o problema é compartilhado pelos nativos e/ou aprendizes da língua portuguesa e, além disso, para quantificar com que frequência ele acontece no corpus.

Categoria Clareza-Coesao → Conectivos-Errado

A Figura 4 mostra o uso incorreto de um marcador discursivo. No exemplo o marcador “*por tanto*” foi escrito de forma separada, quando devia ser escrito “portanto”. Embora o erro apresentado no exemplo, não represente uma falta de coerência ou clareza, este tipo de erro é incluído nesta categoria pelos motivos seguintes: (i) às vezes o uso incorreto do marcador discursivo influencia na coerência da sentença e (ii) os problemas relacionados com os marcadores discursivos podem ter natureza diversa (interferência da língua nativa, erros de ortográfica, etc.), desta forma esses itens léxicos são tratados na anotação como uma categoria independente.

Uma alternativa para anotar erros relacionados com marcadores discursivos poderia ser utilizar múltiplos rótulos, caso se identifique a natureza do erro relacionado com o marcador discursivo. Porém, optar por essa alternativa geraria muita ambiguidade da tarefa de anotação. Por exemplo, na anotação do exemplo da Figura 4, além de etiquetar “por tanto” com a etiqueta “Conetivos-Errados” ele também poderia ser etiquetado com a etiqueta “Erro-Ortografico”. Reconhecendo que no espanhol existe o marcador discursivo “por tanto”, acredito que o erro produzido pelo aluno representa uma interferência da sua língua nativa. Seguindo essa linha, o erro também poderia ser etiquetado como “uso de palavras em espanhol”. Desta forma, recomendo etiquetar o erro como “Conetivos-Errados” para diminuir as ambiguidades na anotação.

Entretanto , no contexto cultural dos países de língua espanhola e portuguesa , estes estudos se escasseam , e por tanto continuam silenciando todo um conjunto de aspectos relativos a referidas indidentidades .

Conetivos-Errados
e por tanto continuam silenciando

Figura 4: Exemplo de sentença, na qual um marcador discursivo foi utilizado de forma errada.

O exemplo da Figura 5 também mostra um erro relacionado com o uso de marcadores discursivos. Desta vez é mais fácil identificar que o erro foi produzido por uma interferência do espanhol. O aluno utilizou o marcador “*por uma parte*” do espanhol em vez de utilizar “por um lado”. Recomendo que a anotação deste exemplo seja a mesma que no exemplo anterior.

Destas tendências , por uma parte , derivam-se , de forma natural , algumas mudanças nos princípios metodológicos que deveriam guiar o ensino e a aprendizagem de nossos dias , o que se apresenta na SEÇÃO 4 , e , por outra , mudanças nos próprios conteúdos da nossa educação , mais acordes com as finalidades que hoje se pretende alcançar , tal como fica explicado na SEÇÃO 5 .

Conetivos-Errados
tendências , por uma parte , derivam-se ,

Figura 5: Exemplo de um marcador discursivo utilizado de forma incorreta.

Além dos exemplos apresentados nas Figuras 4 e 5, há outros erros que podem estar relacionados com o uso dos marcadores discursivos. Estes erros podem estar relacionados com a posição, dentro da sentença, onde o marcador foi utilizado. Além disso, também podem ser observados erros relacionados com o uso ou não das vírgulas com os marcadores. Acredito que estes tipos de erros podem ser produzidos tanto por nativos quanto por aprendizes da língua.

Adequação Lexical e Gramatical

A seguir, são detalhados os erros da categoria “Erros-Gramaticais” presente na tipologia definida na ferramenta Brat. Veja que os erros desta categoria são anotados na cor verde, assim que cada etiqueta da categoria comentada neste manual aparecerá nessa cor.

Erro de Ortografia (Acentuação ortográfica equivocada; Marca equivocada de nasalização; Confusão de fonemas; Confusão de grafemas para o mesmo fonema)

→ Erro-Ortografico

Consideram-se erros ortográficos aqueles relacionados com: (i) acentuação ortográfica equivocada; (ii) marca equivocada de nasalização; (iii) confusão de fonemas e (iv) confusão de grafemas. A seguir são apresentados exemplos de erros gramaticais, anotados com a ferramenta Brat.

Na Figura 6 observa-se um erro de ortografia relacionado com a falta de acentuação na palavra “*ultimas*”, ela deve ser escrita com acento na “u”, “última”. Por outro lado, a palavra “*parcias*”, mostra um erro relacionado com a confusão de grafemas, desta forma a grafia correta é “parciais”.

Figura 6: Exemplo da anotação de erros de ortografia. Falta de acentuação e confusão de grafemas.

Na Figura 7 a palavra “*classe*”, aparece escrita incorretamente, mostrando um erro de ortografia em que o fonema s foi trocado pelo fonema ss. Por isso, a palavra escrita corretamente é “classe”.

Figura 7: Exemplo da anotação de um erro de ortografia. Confusão do fonema s por ss.

Ausência ou uso inadequado da crase → Ausencia-Uso-Inadequado-Crase

O uso inadequado da crase na escrita da língua portuguesa é um erro bastante frequente. Exemplos deste tipo de erro serão observados no corpus analisado. No Anexo 1 é apresentada a regra que indica o uso correto da crase. A Figura 8 apresenta um exemplo de ausência de crase. A frase “... o tratamento de águas residuárias a altas taxa, ...” é escrita corretamente como a seguir “... o tratamento de água residuárias à altas taxa, ...”

Figura 8: Exemplo de ausência da crase.

Posição dos pronomes oblíquos → Colocacao-Pronominal

A colocação pronominal refere-se ao estudo dos pronomes oblíquos átonos (me, te, se, o, a, lhe, nos, vos, os, as, lhes) em relação ao verbo. Os pronomes átonos podem ocupar

as seguintes três posições: antes do verbo (próclise), no meio do verbo (mesóclise) e depois do verbo (ênclise).

A colocação pronominal é um dos problemas que mais aparece no corpus analisado. Como o corpus está composto por resumos acadêmicos, e este tipo de textos, em geral é produzido empregando a terceira pessoa, o pronome que mais aparece no corpus é o “se”. A gramática do português estabelece um conjunto de regras que indica o uso correto dos pronomes oblíquos. Por meio destas regras serão identificados os erros nos textos analisados. Uma explicação detalhada destas regras é apresentada no Anexo 2 deste documento.

Na Figura 9 observa-se um erro relacionado com a colocação pronominal. Segundo a regra mostrada na última parte da tabela do Anexo 2 (colocação pronominal nas locuções verbais), no exemplo o pronome “se” deve ser colocado depois do verbo (posição ênclise).

Figura 9: Exemplo de colocação pronominal incorreta.

Uso Incorreto de Artigos → Uso_Incorreto-de-Artigos

O uso incorreto de artigos é um erro observado frequentemente em textos produzidos por hispano falantes. Estes erros estão relacionados com: (i) ausência do artigo; (ii) uso incorreto do artigo e (iii) e uso do artigo quando ele não devia ser usado. O exemplo apresentado na Figura 10 refere-se ao último caso, no qual o artigo “a” não devia ser utilizado. No Anexo 3 é possível obter informações sobre o uso correto dos artigos na língua portuguesa.

Figura 10: Uso incorreto do artigo

Conjugações verbais → Conjugacoes-Verbais

Os hispano falantes enfrentam diversos problemas nas conjugações verbais do português. As categorias de erros que têm sido observadas com maior frequência são: (i) Uso do modo subjuntivo e do tempo futuro do pretérito e (ii) Uso da locução prepositiva “devido a” ou Infinitivo flexionado. A seguir são apresentados vários exemplos que mostram a ocorrência destes tipos de erros nos textos analisados. Na Figura 11 apresenta-se uma sentença que mostra um erro de conjugação do modo subjuntivo, na qual o verbo “desejasse” devia ter sido conjugado como “desejaria”. Por outro lado a Figura 12 mostra um erro do uso do infinitivo flexionado, em que no trecho “as etiquetas semânticas a serem empregadas”, o verbo “serem” devia ter sido usado em infinitivo “ser”.

Figura 11: Exemplos de sentenças com erros de uso do modo subjuntivo e do tempo futuro do pretérito.

Figura 12: Exemplos de sentenças com erros de uso do Infinitivo flexionado

Uso Incorreto de preposições → **Uso-Incorreto-de-Preposicoes**

O uso incorreto de preposições é um dos erros mais frequentes nos textos escritos pelos hispano falantes. A anotação deste tipo de erros às vezes é complicada porque o erro pode parecer um erro relacionado ao uso dos artigos. Em geral, os equívocos com as preposições estão relacionados com: (i) ausência da preposição; (ii) uso incorreto da preposição e (iii) e uso da preposição quando ela não devia ser usada. O exemplo apresentado na Figura 13 refere-se ao último caso. Neste exemplo, o autor utilizou a contração “na” (em + a), quando somente devia ser utilizado o artigo “a”. No Anexo 4 é possível obter informações sobre o uso correto das preposições na língua portuguesa.

Figura 13: Uso incorreto da preposição em.

Concordância Nominal → **Concordância-Nominal**

Acredito que a maioria das vezes, os erros de concordância nominal acontecem devido a descuidos na escrita dos textos. Caso aconteça esse tipo de erro nos textos analisados, ele será anotado. Na Figura 14 é apresentado um exemplo desta categoria de erro.

Figura 14: Exemplo de erro de concordância nominal de número.

Concordância Verbal → **Concordancia-Verbal**

A repartição agrária e as políticas públicas tem sido os principais motores da transformação dos bosques.

Figura 15: Exemplo de sentença com erro de concordância Verbal.

Mudança da classe gramatical da palavra por desconhecimento de sua ortografia → **Classe-Gramatical-Errada-Ortografia**

As técnicas de análise de DNA tais como eletroforese em gel de gradiente desnaturaste (DGGE), podem gerar perfis moleculares que permitem a análise de varias amostras ambientais de forma simultânea, podendo ser útil para o acompanhamento e compreensão das variações espaciais e temporais das comunidades microbianas.

Figura 16: Exemplos de erros ortográficos porque o aluno não conhece a classe gramatical da palavra.

Interferências do Espanhol

Exemplos de erros produzidos pela interferência do espanhol na escrita em Português acontecem de diversas formas nos textos analisados. As categorias de erros detalhadas a seguir referem-se a erros relacionados com a interferência de itens lexicais do espanhol nos textos em português. Os exemplos a seguir mostram palavras ou frases que foram escritas totalmente em espanhol ou escritas numa interlíngua. No último caso, a frase pode estar composta por palavras em português, mas a frase como uma unidade lexical, não acontece no português brasileiro. Esses últimos casos são classificados como colocações do espanhol. Os exemplos a seguir mostram, em negrito e itálico, sequências com as características mencionadas.

Exemplos:

- **Sequências que existem em português**, mas são pouco usadas e representam interferências do espanhol:

- O artigo termina propondo a necessidade de iniciar uma nova etapa que priorize a ação afirmativa e a discriminação positiva *además* propõe cinco opções para
- A análise está baseada nos resultados de estudos de casos realizados durante 1998 no *enquadre* do Programa Reformas Sociais ...

- **Sequências com interferências do espanhol, mas que não existem em português**

- **Foram escritas em espanhol**
 - *Asimismo*, com o objetivo de desenvolver um programa de capacitação intelectual, deve-se considerar todas as regras estabelecidas pelo governo ...
- **Foram escritas quase em português**
 - Sob uma denominação plural, modelos teóricos e enfoques contemporâneos, queremos **pôr de manifesto** que não há unanimidade nas práticas ...

A anotação dos erros da categoria interferência da língua requer uma análise muito detalhada das sentenças, porque em geral é necessário ter um conhecimento básico da língua espanhola para conseguir identificar possíveis interferências. Recomendo utilizar a ferramenta Linguee², pois ela pode trazer benefícios na hora de identificar os erros.

Uso de palavras ou expressões em espanhol e do espanhol no texto → **Uso-de palavras-em-espanhol**

Na Figura 17 se observam diversas palavras em espanhol que aparecem no texto em português. No lugar de “*las*” devia ter sido utilizada “as” e no lugar de “*la*”, “a”. Por outro lado, a palavra “concreção” é uma interferência do espanhol, ela existe em português, mas sua frequência é muito baixa, no exemplo a palavra deve ser trocada por consolidação.

Da mesma maneira, se tenta dotar o leitor de um ponto de vista amplo, com critérios rigorosos, que permita possível analisar e tornar visíveis las múltiplas indetidades, la diversidad de fins e a heterogeneidade de propósitos que há atrás de cada prática da eads.

²<http://www.linguee.com.br/portugues-espanhol?query=%E1guas+residu%E1rias&source=portuguese>

Figura 17: Sentenças com palavras do espanhol.

Flexão de gênero dos substantivos → Mudanca-Genero

Os erros de flexão dos substantivos geralmente acontecem produto da interferência da língua nativa. Há alguns substantivos femininos em espanhol que em português eles são masculinos e vice-versa. Essas mudanças interferem na escrita e consequentemente, aparece o erro nos textos. Na Figura 18 é observado um erro desta categoria. No Anexo 5 é possível obter informações sobre o gênero dos substantivos na língua portuguesa.

Figura 18: Exemplos de erros de flexão do gênero dos substantivos

Colocações do Espanhol → Colocacoes

Considerando que as colocações são sequências de palavras que convencionalmente acontecem em conjunto nos textos e com uma alta frequência. Nos textos produzidos por hispano falantes também é observada a presença de colocações do espanhol. Da mesma forma que as palavras, as colocações podem aparecer escritas em espanhol ou escritas na interlíngua, de forma que as palavras que formam a colocação existem em português, mas a sequência delas representa um erro da língua portuguesa. O exemplo da Figura 19 mostra como o aluno utilizou a colocação “*pôr de manifesto*” quando deve ter usado a palavra “evidenciar”.

Figura 19: Exemplo de uma colocação do espanhol.

Anexo 1. Uso da crase

A crase ocorre quando temos que juntar duas vogais iguais em uma frase. Por exemplo, a preposição "a" seguida do artigo "a", resultará no "a" com crase. Por ser formada pelo artigo "a", a crase só aparecerá antes de palavras femininas. Uma boa dica para verificar se existe crase em uma vogal, é trocar a palavra feminina para uma masculina. Se nessa troca, aparecer "ao(s)", significa que existe crase. Exemplo: Vou à igreja -> Vou **ao** shopping.

Neste exemplo é fundamental destacar que a verbo ir (vou) exige preposição, e igreja é uma palavra feminina. Quando trocada por "shopping", o "a + a" tornou-se "ao". Então existe crase.

A Tabela a seguir indica alguns casos em que o uso da crase é obrigatório e outros em que a crase não é empregada.

Nunca existe crase antes de:	Sempre ocorrerá crase antes de:	Pode ou não ocorrer crase antes
Palavras masculinas	Locuções prepositivas, adverbiais e conjuntivas: às vezes, à toa, à esquerda, à noite ..	Nomes de cidades, lugares, países, etc. Um bom truque para saber se vai crase ou não, é encaixar a palavra em questão na frase: "Vou a, volto da, crase há! vou a, volto de, crase pra quê?"
Entre palavras repetidas (dia-a-dia)	Antes de numeral cardinal indicando hora	
Artigos indefinidos (um, umas, uns, umas)		
Palavras no plural se o "a" estiver no singular		
Numeral cardinal (exceto se indicarem hora)		

Anexo 2. Colocação pronominal (Posição dos pronomes oblíquos)

Próclise	Mesóclise	Ênclise
(1) Com palavras ou expressões negativas: não, nunca, jamais, nada, ninguém, nem, de modo algum. Ex: - Nada <u>me</u> perturba. - Ninguém <u>se</u> mexeu. - De modo algum <u>me</u> afastarei daqui. - Ela <u>nem se</u> importou com meus problemas.	(1) Quando o verbo estiver no futuro do presente (vai acontecer – amarei, amarás, ...) ou no futuro do pretérito (ia acontecer mas não aconteceu – amaria, amarias, ...) Ex: - Convidar-me-ão para a festa. - Convidar-me-iam para a festa. Se houver uma palavra atrativa, a <i>próclise</i> será obrigatória. - Não (palavra atrativa) me convidarão para a festa.	(1) Ênclise de verbo no futuro e particípio está sempre errada. Ex: - Tornarei-me..... (errada) - Tinha entregado nos(errada)
(2) Com conjunções subordinativas: quando, se, porque, que, conforme, embora, logo, que. Ex: - Quando se trata de comida, ele é um "expert". - É necessário que a deixe na escola. - Fazia a lista de convidados, conforme me lembrava dos amigos sinceros.		(2) Ênclise de verbo no infinitivo está sempre certa. Ex: - Entregar-lhe (correta) - Não posso recebê-lo. (correta)
(3) Advérbios Ex: - Aqui se tem paz. - Sempre <u>me</u> dediquei aos estudos. - Talvez <u>o</u> veja na escola. OBS: Se houver vírgula depois do advérbio, este (o advérbio) deixa de atrair o pronome. - Aqui, <u>trabalha-se</u> .		(3) Outros casos: Ex: - Com o verbo no início da frase: Entregaram-me as camisas. - Com o verbo no imperativo afirmativo: Alunos, comportem-se. - Com o verbo no gerúndio: Saiu deixando-nos por instantes. - Com o verbo no infinitivo impessoal: Convém contar-lhe

		<p>tudo.</p> <p>OBS: se o gerúndio vier precedido de preposição ou de palavra atrativa, ocorrerá a próclise:</p> <ul style="list-style-type: none"> - Em se tratando de cinema, prefiro o suspense. - Saiu do escritório, não nos revelando os motivos.
<p>(4) Pronomes relativos, demonstrativos e indefinidos.</p> <p>Ex:</p> <ul style="list-style-type: none"> - <u>Alguém</u> <u>me</u> ligou? (indefinido) - A pessoa <u>que</u> <u>me</u> ligou era minha amiga. (relativo) - <u>Isso</u> <u>me</u> traz muita felicidade. (demonstrativo) 		
<p>(5) Em frases interrogativas.</p> <p>EX:</p> <ul style="list-style-type: none"> - <u>Quanto</u> <u>me</u> cobrará pela tradução? 		
<p>(6) Em frases exclamativas ou optativas (que exprimem desejo).</p> <p>Ex:</p> <ul style="list-style-type: none"> - Deus <u>o</u> abençoe! - Macacos <u>me</u> mordam! - Deus <u>te</u> abençoe, meu filho! 		
<p>(7) Com verbo no gerúndio antecedido de preposição EM.</p> <p>Ex:</p> <ul style="list-style-type: none"> - <u>Em se plantando</u> tudo dá. - <u>Em se tratando de beleza</u>, ele é campeão. 		
<p>(8) Com formas verbais proparoxítonas</p> <p>Ex:</p> <ul style="list-style-type: none"> - Nós <u>o censurávamos</u>. 		
<p>COLOCAÇÃO PRONOMINAL NAS LOCUÇÕES VERBAIS</p> <p>Locuções verbais são formadas por um verbo auxiliar + infinitivo, gerúndio ou particípio.</p> <p>AUX + PARTICÍPIO: o pronome deve ficar depois do verbo auxiliar. Se houver palavra atrativa, o pronome deverá ficar antes do verbo auxiliar.</p> <p>Ex:</p> <ul style="list-style-type: none"> - <u>Havia-lhe contado</u> a verdade. - <u>Não</u> (palavra atrativa) <u>lhe havia contado</u> a verdade. <p>AUX + GERÚNDIO OU INFINITIVO: se não houver palavra atrativa, o pronome oblíquo virá depois do verbo auxiliar ou do verbo principal.</p> <p>Infinitivo</p> <ul style="list-style-type: none"> - <u>Quero-lhe dizer</u> o que aconteceu. - <u>Quero dizer-lhe</u> o que aconteceu. <p>Gerúndio</p> <ul style="list-style-type: none"> - <u>Ia-lhe dizendo</u> o que aconteceu. - <u>Ia dizendo-lhe</u> o que aconteceu. <p>Se houver palavra atrativa, o pronome oblíquo virá antes do verbo auxiliar ou depois do verbo principal.</p> <p>Infinitivo</p> <ul style="list-style-type: none"> - <u>Não lhe quero dizer</u> o que aconteceu. - <u>Não quero dizer-lhe</u> o que aconteceu. <p>Gerúndio</p> <ul style="list-style-type: none"> - <u>Não lhe ia dizendo</u> a verdade. - <u>Não ia dizendo-lhe</u> a verdade. 		

Anexo 3. Uso de artigos

Os artigos são amigos inseparáveis do [substantivo](#), pois toda vez que for utilizado um artigo ele se estará se referindo a um substantivo. Esta referência do artigo poderá ser definida ou indefinida. Desta forma, os artigos são: Artigos definidos (o, os, a, as) e Artigos indefinidos (um, uns, uma, umas).

Mais informação do uso correto dos artigos na língua portuguesa está disponível no endereço: <http://www.infoescola.com/portugues/artigos/>.

Anexo 4. Uso de preposições

Mais informação sobre o uso correto das preposições na língua portuguesa está disponível no endereço: <http://www.infoescola.com/portugues/preposicao/>.

Anexo 5. Flexão de gênero dos substantivos

Quanto ao gênero, os substantivos podem ser flexionados em dois gêneros: masculino e feminino. Gênero é a propriedade que as palavras têm de indicar sexo real ou fictício dos seres. A gramática da língua portuguesa indica por regra que todo substantivo masculino é caracterizado pela desinência “o” e o feminino pela desinência “a”. No entanto, nem todos os substantivos masculinos terminam em “o” (ex. telefonema), assim como nem todos os substantivos femininos terminam em “a” (ex. mulher).

Portanto:

- **Pertencem ao gênero masculino os substantivos que podem vir precedidos dos artigos o, os, um, uns.**
- **Pertencem ao gênero feminino os substantivos que podem vir precedidos dos artigos a, as, uma, umas.**
- Exemplos
 - o estudante (masculino)
 - a estudante (feminino)
 - o cabeça (líder)
 - a cabeça (parte do corpo)

Uma explicação mais detalhada do gênero dos substantivos pode ser encontrada a seguir.

<http://www.infoescola.com/portugues/flexao-de-genero-nos-substantivos/>

Sufixos

ESPAÑHOL	PORTUGUÊS	ORIGEM	CATEGORIA
ADA	ADA	LATIM	SUBSTANTIVO
ADA	ADO	LATIM	ADJETIVO
AJE	AGEM	LATIM	SUBSTANTIVO
AL	AL	LATIM	SUBSTANTIVO
ALÓN	ALHÃO	LATIM	AUMENTATIVO
ANZA	ANÇA	LATIM	SUBSTANTIVO
ANO	ANO	LATIM	ADJETIVO
ANTE	ANTE	LATIM	SUBSTANTIVO
ANTE	ANTE	LATIM	ADJETIVO
ÓN	ÃO	LATIM	AUMENTATIVO
ERÍA	ARIA	LATIM	SUBSTANTIVO
ARIA	ÁRIA	LATIM	SUBSTANTIVO
ARIO	ÁRIO	LATIM	SUBSTANTIVO
ABLE	ÁVEL	LATIM	ADJETIVO
CIÓN	ÇÃO	LATIM	SUBSTANTIVO
COLA	CULA	LATIM	DIMINUTIVO
COLO	CULO	LATIM	DIMINUTIVO
DAD	DADE	LATIM	SUBSTANTIVO
DOR	DOR	LATIM	SUBSTANTIVO
DURA	DURA	LATIM	SUBSTANTIVO

EAR	EAR	LATIM	VERBO
ECER	ECER	LATIM	VERBO
ERA	EIRA	LATIM	SUBSTANTIVO
ERO	EIRO	LATIM	SUBSTANTIVO
ENZA	ENÇA	LATIM	SUBSTANTIVO
EÑO	ENSE	LATIM	ADJETIVO
ENTE	ENTE	LATIM	SUBSTANTIVO
ENTE	ENTE	LATIM	ADJETIVO
ENTO	ENTO	LATIM	ADJETIVO
ERIA	ERIA	LATIM	SUBSTANTIVO
ERIO	ÉRIO	LATIM	SUBSTANTIVO
ÊS	ÊS	LATIM	ADJETIVO
	ESCER	LATIM	VERBO
	ESTE	LATIM	ADJETIVO
ESTRE	ESTRE	LATIM	ADJETIVO
EZ	EZ	LATIM	SUBSTANTIVO
EZA	EZA	LATIM	SUBSTANTIVO
ÍA	IA	GREGO	SUBSTANTIVO
ICA	ICA	LATIM	ADJETIVO
ICIA	ÍCIA	LATIM	ADJETIVO
ICIO	ÍCIO	LATIM	ADJETIVO
ICO	ICO	LATIM	ADJETIVO
IL	IL	LATIM	ADJETIVO
ILLA	INHA	LATIM	DIMINUTIVO
ILLO	INHO	LATIM	DIMINUTIVO
ÍSIMO	ÍSSIMO	LATIM	SUPERLATIVO
ÍSIMA	ÍSSIMA	LATIM	SUPERLATIVO
ANZA	INO	LATIM	ADJETIVO
	IO	LATIM	ADJETIVO
ISMO	ISMO	GREGO	SUBSTANTIVO
ISTA	ISTA	GREGO	SUBSTANTIVO
	ITE	GREGO	SUBSTANTIVO
IBLE	ÍVEL	LATIM	ADJETIVO
IVO	IVO	LATIM	ADJETIVO
IZAR	IZAR	GREGO	VERBAL
IZO	IÇO		ADJETIVO
IZA	IÇA		ADJETIVO

MENTE	MENTE	LATIM	ADVÉRPIO
MIENTO	MENTO	LATIM	SUBSTANTIVO
ÓN	ADOR	LATIM	ADJETIVO
OSA	OSA	LATIM	ADJETIVO
	OSE	GREGO	SUBSTANTIVO
OSO	OSO	LATIM	ADJETIVO
CIÓN	SÃO	LATIM	SUBSTANTIVO
SIÓN	SÃO	LATIM	SUBSTANTIVO
SOR	SOR	LATIM	SUBSTANTIVO
TOR	TOR	LATIM	SUBSTANTIVO
TUD	TUDE	LATIM	SUBSTANTIVO
	TURA	LATIM	SUBSTANTIVO
OLA	ULA	LATIM	DIMINUTIVO
OLO	ULO	LATIM	DIMINUTIVO
	UME	LATIM	SUBSTANTIVO
UELA	ZINHA	LATIM	DIMINUTIVO
UELO	ZINHO	LATIM	DIMINUTIVO

Verbos

Verbo	Sinônimos	Combinação Linear
ressaltar	ressaltar	1.5/-0.222349
	sublinhar	0/0.845089
	ressalvar	0/0.390693
	destacar	0/0.0810311
	marcar	0/0.0794653
	acentuar	0/-0.096638
	ênfatizar	0/-0.169512
	notar	0/-0.172496
	salientar	0/-0.594893
investigar	perguntar	1.5/0.677571
	investigar	1.5/0.392254
	indagar	1.5/0.156558
	pesquisar	1.5/0.112605
	buscar	1.5/0.0350758
	conferir	0/0.0777676
compilar	compilar	1.5/-0.514283
	levantar	0/-0.085696
aprofundar	aprofundar	1.5/0.221223
	recorrer	1.5/0.0711164
	mergulhar	0/0.698634

	reforçar	0/0.336684
	estabelecer	0/0.0664598
	fixar	0/-0.334424
instaurar	instaurar	1.5/0.345437
	criar	1.5/0.250581
	estabelecer	1.5/0.0664598
	instituir	1.5/-0.404193
	fazer	0/0.255752
	desenvolver	0/0.248608
	inaugurar	0/0.234574
	adotar	0/0.0842294
	inserir	0/-0.123445
outorgar	conceder	1.5/0.385415
	outorgar	1.5/-0.480761
	dar	0/0.170017
	ter	0/0.113013
	conferir	0/0.0777676
citar	dar	1.5/0.170017
	citar	1.5/-0.318402
	mencionar	1.5/-0.499762
	falar	0/0.371823
	dizer	0/0.213834
ver	avistar	1.5/1.18477
	antever	1.5/0.651468
	ler	1.5/0.485889
	enxergar	1.5/0.457442
	ver	1.5/0.361784
	conceber	1.5/0.243608
	observar	1.5/0.198403
	assistir	1.5/0.19833
	perceber	1.5/0.166082
	sentir	1.5/0.162316
afirmar	afirmar	1.5/0.294451
	sustentar	1.5/0.235201
	dizer	1.5/-0.222349
	gostar	0/0.749448
	ponderar	0/0.490018

	ressalvar	0/0.390693
	reafirmar	0/0.386259
	aceitar	0/0.279171
	reconhecer	0/0.216225
corrigir	corrigir	1.5/-0.302313
	retificar	1.5/-0.424671
	ressalvar	0/0.390693

Marcadores Discursivos

Espanhol	Português
<i>por otro lado; por otra parte; de otro lado; de otra parte</i>	<i>por outro lado; por outra parte; de outro lado</i>
<i>por consiguiente; por tanto; por lo tanto</i>	<i>por conseguinte; portanto; conseqüentemente; consequentemente</i>
<i>ya que; puesto que; porque; pues; dado que</i>	<i>porque; por; pois; uma vez que</i>
<i>por último; en resumen; al fin; en fin; por fin; en definitiva; en síntesis; finalmente; al fin y al cabo</i>	<i>por último; enfim; finalmente; por fim; afinal</i>
<i>así y todo; de todos modos; aunque; pero; sino; más bien; sin embargo; con todo; en resumen; no obstante; por ello</i>	<i>mas; contudo; porém; não obstante; no entanto</i>
<i>dicho sea de paso; a propósito; mejor dicho; por cierto; además</i>	<i>aliás; ou melhor; a propósito</i>
<i>acto seguido; a continuación; luego; al cabo</i>	<i>em seguida; após; depois; a seguir</i>

<i>asimismo; de este modo; así las cosas; siendo así; ni; desde luego; por eso; así; para ello; por ello; en todo caso; entonces; de ahí que; de ahí</i>	<i>assim; para isso; então; daí</i>
<i>de todos modos; siendo así; aun así</i>	<i>de qualquer forma; sendo assim; ainda assim</i>
<i>igualmente; asimismo; así mismo; aun así; además</i>	<i>igualmente; mesmo assim; além disso</i>
<i>es decir; o sea; esto es</i>	<i>quer dizer; ou seja; isto é</i>
<i>sobre todo; por encima de todo</i>	<i>sobretudo; por cima de tudo; acima de tudo</i>
<i>pues bien; ahora; ahora bien</i>	<i>pois bem; enquanto; agora</i>
<i>por todo ello; para ello; por tanto; gracias a eso; por eso mismo; por ello; por eso; por lo tanto</i>	<i>por tudo isso; para tanto; por isso</i>
<i>en suma; en síntesis; en resumen</i>	<i>em suma; em resumo; em síntese</i>
<i>incluso; inclusive; asimismo; igualmente; hasta; para empezar</i>	<i>inclusive; mesmo; para; até</i>
<i>claro que; claro; por supuesto; ciertamente</i>	<i>claro que; claro; certamente; é claro</i>
<i>en realidad</i>	<i>na verdade; na realidade</i>
<i>primeramente; primero</i>	<i>primeiro; primeiramente</i>
<i>de lo contrario; por el contrario; al contrario; muy por el contrario; al revés; en cambio; en caso contrario</i>	<i>do contrário; pelo contrário; ao contrário; caso contrário</i>
<i>dado que; tal como; sino; también</i>	<i>como; conforme</i>
<i>en contraste; ahora bien; pues</i>	<i>em contraste; já</i>
<i>de este modo</i>	<i>dessa maneira;dessa forma;desse modo</i>
<i>bien; por lo tanto; bueno; así; pues bien</i>	<i>logo; bom; bem</i>
<i>en conclusión</i>	<i>em conclusão;conclusão</i>
<i>incluso; aunque; si bien; también; además</i>	<i>ainda; mesmo que; embora; ainda que</i>
<i>tampoco; ni; no; también</i>	<i>também não; tampouco; nem; também</i>

<i>en efecto; de hecho; evidentemente; sí; claro</i>	<i>de fato; evidentemente; sim</i>
<i>así como; así que</i>	<i>bem como; assim que</i>