

ALINE EVERS

**PROCESSAMENTO DE LÍNGUA NATURAL E NÍVEIS DE
PROFICIÊNCIA DO PORTUGUÊS: UM ESTUDO DE PRODUÇÕES
TEXTUAIS DO EXAME CELPE-BRAS**

**Porto Alegre
2013**

**UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
ÁREA: ESTUDOS DA LINGUAGEM
LINHA DE PESQUISA: TEORIAS LINGUÍSTICAS DO LÉXICO: RELAÇÕES
TEXTUAIS**

**PROCESSAMENTO DE LÍNGUA NATURAL E NÍVEIS DE
PROFICIÊNCIA DO PORTUGUÊS: UM ESTUDO DE PRODUÇÕES
TEXTUAIS DO EXAME CELPE-BRAS**

ALINE EVERS

ORIENTADORA: Profa. Dra. Maria José Bocorny Finatto

Dissertação de Mestrado em *Teorias Linguísticas do Léxico: Relações Textuais*, apresentada como requisito para obtenção do título de Mestre pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.

**Porto Alegre
2013**

CIP - Catalogação na Publicação

Evers, Aline

Processamento de língua natural e níveis de proficiência do português: um estudo de produções textuais do exame Celpe-Bras / Aline Evers. -- 2013. 174 f.

Orientadora: Maria José Bocorny Finatto.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, BR-RS, 2013.

1. Língua Portuguesa. 2. Processamento de Língua Natural. 3. Exame Celpe-Bras. 4. Linguística Textual. 5. Linguística de Corpus. I. Bocorny Finatto, Maria José, orient. II. Título.

ALINE EVERS

**PROCESSAMENTO DE LÍNGUA NATURAL E NÍVEIS DE
PROFICIÊNCIA DO PORTUGUÊS: UM ESTUDO DE PRODUÇÕES
TEXTUAIS DO EXAME CELPE-BRAS**

Dissertação de Mestrado em *Teorias Linguísticas do Léxico: Relações Textuais*, apresentada como requisito para obtenção do título de Mestre pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.

Aprovada em 24 de maio de 2013.

BANCA EXAMINADORA

**Profa. Dra. Juliana Roquele Schoffen – UFRGS
Profa. Dra. Gladis Maria de Barcellos Almeida – UFSCar
Profa. Dra. Lucelene Lopes – PUCRS**

**Porto Alegre
2013**

AGRADECIMENTOS

À professora Maria José Bocorny Finatto, por todas as apostas compradas ao longo desses anos de trabalho conjunto. Muito obrigada pelos incentivos não só acadêmicos, mas profissionais e pessoais, pela amizade, pelo tempo, pelo exemplo. Agradeço de antemão pelos trabalhos em andamento: ainda estamos longe de concluir a nossa parceria.

À Secretaria de Educação a Distância da Universidade Federal do Rio Grande do Sul, pela bolsa de pós-graduação concedida ao longo desses dois anos de pesquisa.

Às diretoras e idealizadoras da Escola Bem Brasil de português para estrangeiros: Simone Paula Kunrath, Juliana Roquele Schoffen, Letícia Grubert dos Santos e Graziela Hoerbe Andrighetti. Agradeço pela formação docente, pelas trocas e por todas as oportunidades diretas e indiretas que a porta aberta da escola me trouxe.

Aos parceiros de pesquisa do Instituto de Informática desta universidade, Aline Villavicencio e Rodrigo Wilkens, por apontarem caminhos para o desenvolvimento desta pesquisa.

Ao Núcleo Interinstitucional de Linguística Computacional, especialmente nas figuras da professora Sandra Maria Alúcio e de suas orientandas, Carolina Scarton e Lianet Sepúlveda-Torres, pelas propostas de abordagens e intercâmbios de pesquisa.

Às parceiras de pesquisa da Pontifícia Universidade Católica do Rio Grande do Sul, Renata Vieira e Lucelene Lopes, pelas contribuições que resultaram em discussões diluídas neste trabalho.

À Bianca Franco Pasqualini, pelas tardes no Skype, pela leitura, pela amizade. Também agradeço às colegas Juliana Andrade Feiden e Bruna Dahm, por mostrarem que há espaço na academia e no ensino para todo mundo.

Ao Felipe, por sequer transparecer qualquer indício de incômodo durante as minhas infundáveis tentativas de explicar este trabalho. E por discutir e enriquecer conceitos de Língua e Linguagem do jeito que só um Filósofo é capaz de fazer.

À minha família que, mesmo estando separada pela distância neste último ano e sem entender exatamente o que faço, vem apostando em mim todos os dias, há muitos e muitos anos. Mãe, Pai, Henrique, Marli, Laís: obrigada.

Mas dificuldades há, embora nem sempre sejam mencionadas nos livros de introdução à biologia. Uma delas é o ornitorrinco. Esse estranho bicho australiano bota ovos, mas amamenta os filhotes; tem a temperatura do corpo parcialmente dependente da temperatura ambiente e tem pêlos. Será um mamífero, um réptil ou outra categoria qualquer? Isso depende de darmos mais importância a um ou outro dos critérios que definem as classes. De qualquer forma, é necessário admitir que as categorias "mamífero" e "réptil", embora convenientes e úteis, não são perfeitas. A maioria dos animais se coloca claramente em uma ou outra das diversas classes reconhecidas pelos zoólogos; mas há alguns, como o ornitorrinco, que ficam mais ou menos no meio.

Mário Perini

RESUMO

Este trabalho trata dos temas da proficiência em português como língua adicional e da detecção de padrões lexicais e coesivos a partir de um enfoque computacional, situando o tema em meio à descrição de textos produzidos no contexto do exame de proficiência Celpe-Bras de 2006-1. Fazendo uso de pressupostos teórico-metodológicos da Linguística de Corpus, da Linguística Textual e do Processamento de Língua Natural, investigou-se a hipótese de que seria possível classificar, de modo automático, textos submetidos ao exame conforme níveis de proficiência pré-estabelecidos. Por meio do processamento de 177 textos previamente avaliados por corretores humanos em seis níveis (Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior), usou-se o Aprendizado de Máquina (AM) supervisionado para cotejar padrões lexicais e coesivos capazes de distinguir os níveis sob estudo. Para o cotejo dos padrões, a ferramenta Coh-Metrix-Port – que calcula parâmetros de coesão, coerência e inteligibilidade textual – foi utilizada. Cada um dos textos foi processado na ferramenta; para o AM, os resultados da ferramenta Coh-Metrix-Port foram usados como *atributos*, os níveis de proficiência como *classes* e os textos como *instâncias*. As etapas de processamento do *corpus* foram: 1) digitação do *corpus*; 2) processamento individual dos textos na ferramenta Coh-Metrix-Port; 3) análise usando AM – Algoritmo J48 – e os seis níveis de proficiência; 4) nova análise usando AM e duas novas classes: textos sem certificação (Iniciante e Básico) e com certificação (Intermediário, Intermediário Superior, Avançado e Avançado Superior). Apesar do tamanho reduzido do *corpus*, foi possível identificar os seguintes atributos distintivos entre os textos da amostra: número de palavras, medida de riqueza lexical, número de parágrafos, incidência de conectivos negativos, incidência de adjetivos e Índice Flesch. Chegou-se a um classificador capaz de separar dois conjuntos de texto (SEM e COM CERTIFICAÇÃO) através das métricas utilizadas (*f-measure* de 70%).

Palavras-chave: português como língua adicional; português como língua estrangeira; *corpora* de aprendizes de português; Celpe-Bras; proficiência em língua portuguesa; Coh-Metrix; processamento de língua natural.

ABSTRACT

This research analyzes Portuguese proficiency from a computational perspective, studying texts submitted to the Brazilian Portuguese proficiency exam Celpe-Bras (Certificate of Proficiency in Portuguese for Foreigners). The study was based on Corpus Linguistics, Textual Linguistics, and Natural Language Processing. We investigated the hypothesis that it would be possible to predict second language proficiency using Machine Learning (ML), measures given by a NLP tool (Coh-Metrix-Port), and a *corpus* of texts previously classified by human raters. The texts (177) were previously classified as Beginner, Elementary, Intermediate, Upper Intermediate, Advanced, and Upper Advanced. After preparation, they were processed by Coh-Metrix-Port, a tool that calculates cohesion, coherence, and textual readability at different linguistic levels. The output of this tool provided 48 measures that were used as *attributes*, the proficiency levels given by raters were considered *classes*, and the 177 were considered *instances* for ML purposes. The algorithm J48 was used with this set of texts, providing a Decision Tree that classified the six levels of proficiency. The results for this analysis were not conclusive; because of that, we performed a new analysis with a new set of texts: two classes, one with texts that did not receive certificate (Beginner and Elementary) and the other with texts that did receive the certificate (Intermediate, Upper Intermediate, Advanced, and Upper Advanced). Despite the small size of the *corpus*, we were able to identify the following distinguishing attributes: number of words, type token ratio, number of paragraphs, incidence of negative connectives, incidence of adjectives, and Flesch Index. The classifier was able to separate these two last sets of texts with a F-measure of 70%.

Keywords: Portuguese as an additional language; Portuguese as a foreign language; portuguese learner *corpus*; Celpe-Bras; proficiency in Portuguese; Coh-Metrix; natural language processing.

SUMÁRIO

INTRODUÇÃO	14
TRABALHOS ANTERIORES – EXPERIMENTOS PRÉVIOS	18
OBJETIVOS, QUESTÕES E HIPÓTESES DE PESQUISA	21
ORGANIZAÇÃO DA DISSERTAÇÃO	23
PARTE I – OBJETO DE ESTUDO	25
1 O EXAME CELPE-BRAS	26
1.1 CONSTRUTO TEÓRICO DO EXAME CELPE-BRAS	28
1.2 APLICAÇÃO DO EXAME CELPE-BRAS	33
1.3 NÍVEIS CERTIFICADOS PELO EXAME CELPE-BRAS	36
1.4 PROCESSO DE CORREÇÃO DO EXAME CELPE-BRAS	38
2 O CORPUS DE ESTUDO	46
2.1 DADOS DE COMPOSIÇÃO DO CORPUS	46
2.2 TAREFAS QUE ORIGINARAM O CORPUS	48
2.3 INFORMAÇÕES BÁSICAS DO CORPUS	49
PARTE II – REFERENCIAIS TEÓRICO-METODOLÓGICOS	52
3 LINGUÍSTICA TEXTUAL	53
3.1 O PAPEL DOS RECURSOS LINGUÍSTICOS NA AVALIAÇÃO DE PROFICIÊNCIA	53
3.2 LINGUÍSTICA TEXTUAL	55
3.2.1 PRINCÍPIOS DE TEXTUALIDADE	58
3.2.2 COESÃO E COERÊNCIA	60
3.3 MEDIDAS DE INTELIGIBILIDADE	64
4 LINGUÍSTICA DE CORPUS	67
4.1 LINGUÍSTICA DE CORPUS E USO DA LINGUAGEM	72
4.2 LINGUÍSTICA DE CORPUS E CORPORA DE APRENDIZES	73
4.3 PROJETOS E FERRAMENTAS COM CORPUS	76
4.3.1 CORPORA INTERNACIONAIS	77
4.3.2 CORPORA DO PORTUGUÊS	77
4.3.3 FERRAMENTAS PARA COMPILAÇÃO DE TEXTOS	78
4.3.4 ANOTADORES MANUAIS	78
4.3.5 ANOTADORES AUTOMÁTICOS	79
4.3.6 FERRAMENTAS PARA ACESSO A CORPUS	80
4.3.7 FERRAMENTAS DE EXTRAÇÃO DE CONHECIMENTO	81
4.4 LIMITES DA LINGUÍSTICA DE CORPUS: POSSIBILIDADES NO PLN	82

5 PROCESSAMENTO DE LÍNGUA NATURAL	86
5.1 PLN: BREVE HISTÓRICO E MODO DE TRABALHO	89
5.2 COH-METRIX	95
5.3 COH-METRIX-PORT	97
5.3.1 ÍNDICE FLESCH	100
5.3.2 TYPE-TOKEN RATIO (TTR) OU MEDIDA DE RIQUEZA LEXICAL	101
5.4 APRENDIZADO DE MÁQUINA	102
5.4.1 WEKA	104
5.4.2 ÁRVORE DE DECISÃO, MATRIZ DE CONFUSÃO E AVALIAÇÃO DE AM	106
6 POSICIONAMENTO DESTE TRABALHO	109
PARTE III – PROCEDIMENTOS, RESULTADOS E CONCLUSÕES	111
7 PROCEDIMENTOS	112
8 RESULTADOS E DESCRIÇÃO DOS DADOS OBTIDOS	116
8.1 RESULTADOS LEXICAIS	116
8.2 RESULTADOS COESIVOS	122
8.3 RESULTADOS DO APRENDIZADO DE MÁQUINA	123
9 DISCUSSÃO DOS RESULTADOS	128
9.1 RESULTADOS LEXICAIS	128
9.2 RESULTADOS COESIVOS	130
9.3 RESULTADOS DO APRENDIZADO DE MÁQUINA	131
10 RETOMADA DAS QUESTÕES E HIPÓTESES DE PESQUISA	132
10.1 QUESTÕES DE PESQUISA	132
10.2 HIPÓTESES DE PESQUISA	135
11 LIMITES DO TRABALHO, PERSPECTIVAS E CONSIDERAÇÕES FINAIS	137
REFERÊNCIAS	143
ANEXO I – TAREFAS E TEXTOS-BASE	149
ANEXO II – GRADES DE CORREÇÃO	155
ANEXO III – ARQUIVOS ARFF	167

ÍNDICE DE FIGURAS

Figura 1 Faixas do Índice Flesch.....	20
Figura 2 Enunciado da Tarefa I de 2006-1.....	32
Figura 3 Parte do texto-base da Tarefa I de 2006-1.	32
Figura 4 Grade de Correção e orientação de leitura.	44
Figura 5 Etapas estratégicas para o desenvolvimento de um trabalho em PLN.	92
Figura 6 Ferramentas e recursos linguísticos disponíveis para o português: de 0 (muito baixo) a 6 (muito alto).....	94
Figura 7 Tela inicial do Coh-Metrix 3.0.....	96
Figura 8 Cabeçalho de arquivo produzido pelo sistema Coh-Metrix-Port.....	99
Figura 9 Parte do arquivo contendo métricas e contagens.	99
Figura 10 Todos os personagens do desenho animado Os Simpsons.....	103
Figura 11 Personagens de Os Simpsons classificados em diferentes grupos.	103
Figura 12 Exemplo de arquivo ARFF.....	105
Figura 13 Interface do Weka.....	106
Figura 14 Árvore de Decisão para o problema de espera para jantar em um restaurante.	107
Figura 15 Etapas de processamento do <i>corpus</i> e 1ª Etapa de AM.....	114
Figura 16 Etapas de processamento do <i>corpus</i> e 2ª Etapa de AM.....	115
Figura 17 Similaridade de vocabulário entre os grupos.....	121
Figura 18 Árvore de Decisão para os seis níveis (Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior).	125
Figura 19 Novas classes para os textos: SEM CERTIFICAÇÃO e COM CERTIFICAÇÃO... ..	127
Figura 20 Árvore de Decisão para as duas classes (SEM CERTIFICAÇÃO e COM CERTIFICAÇÃO).	127

ÍNDICE DE GRÁFICOS

Gráfico 1 Examinandos: de 127 à 7.748 em 14 anos. Fonte: Adaptado de Damazo (2012) e atualizado com dados do INEP (2012).....	27
Gráfico 2 Postos aplicadores no Brasil e no Exterior: de 8 postos a 22.	28
Gráfico 3 Número TOTAL de palavras por Tarefa e Nível.....	117
Gráfico 4 Número de palavras DISTINTAS por Tarefa e Nível.	118
Gráfico 5 Palavras totais (linha azul) e palavras distintas (linha vermelha).....	119
Gráfico 6 Matriz de Confusão do classificador.	126

ÍNDICE DE TABELAS

Tabela 1 Métricas do Coh-Matrix-Port para cinco textos de um mesmo estudante.	19
Tabela 2 Grade de Correção da Tarefa I de 2006-1: Fundação Darcy Ribeiro.	40
Tabela 3 Classes e números brutos de textos por Tarefa e Nível.....	50
Tabela 4 Número absoluto de palavras por Tarefa e Nível.....	50
Tabela 5 Número normalizado de palavras por Tarefa e Nível	50
Tabela 6 Exemplo de Matriz de Confusão.	108
Tabela 7 Número TOTAL de palavras por Tarefa e Nível.....	116
Tabela 8 Número de palavras DISTINTAS por Tarefa e Nível.	117
Tabela 9 Intersecção de vocabulário entre os conjuntos de textos.....	120
Tabela 10 Similaridade de vocabulário entre os grupos.	120
Tabela 11 Atributos (medidas do Coh-Matrix-Port) que demonstraram comportamento não aleatório.....	122
Tabela 12 Matriz de Confusão do classificador.....	126

INTRODUÇÃO

Retomando a citação de Mário Perini, este trabalho seria comparável a um ornitorrinco. Digo isso não porque ele não se encaixe em nenhum lugar dentro da complexa árvore conceitual da Linguística, mas porque transita e se faz em diferentes áreas – ora está com os pés na Linguística Aplicada (LA), coloca a mente nas crenças sobre a língua da Linguística de Corpus (LC), ora dá as mãos ao Processamento de Língua Natural (PLN). Por fazer uso de áreas tão diferentes do ponto de vista epistemológico, cabe lembrar aqui que a Linguística é ainda uma ciência em construção, especialmente porque não possui um paradigma aceito e não há consenso entre linguistas sobre as teorias de língua e linguagem. Entendo, portanto, que a Linguística ainda está em um estágio de História Natural, tal como bem coloca Perini, em recente entrevista concedida à ReVEL:

O trabalho científico se compõe de observação e teorização, e nenhum desses aspectos é dispensável. Mas nem a observação sem teoria nem a teorização sem dados tem utilidade. No momento, acredito que se tem teorizado excessivamente, e em certos setores percebo quase que um desprezo pelo trabalho descritivo. Não acredito que nosso conhecimento da linguagem esteja avançado a ponto de permitir a elaboração de teorias abrangentes e detalhadas como algumas das teorias atualmente correntes; acho que a linguística está, em grande parte, no estágio da “história natural”, em que a prioridade é o levantamento de dados confiáveis e sua sistematização segundo princípios rigorosos. Vou repetir: o problema não é a teorização, mas a teorização prematura, isto é, sem fundamentação suficiente dos dados. (2010, p. 11-12)

Sendo uma ciência em construção, a Linguística está comprometida com a coleta de fatos da língua. A nossa responsabilidade, enquanto linguistas, está em sermos fiéis a esses fatos e em descrevê-los da forma mais completa possível, fazendo uso das melhores teorias parciais de que dispomos. Esta dissertação, nesse sentido, dá apenas mais um passo em direção ao processo de coleta de fatos da língua, e outro mais em direção ao processo de descrição da língua portuguesa.

Voltando ao texto da epígrafe, Perini afirma que é difícil categorizar o ornitorrinco como “réptil” ou “mamífero” pois tal classificação, como qualquer outra, depende de critérios e de objetivos assumidos antes da tarefa de classificar. Essas observações são relevantes neste trabalho porque aqui lidamos com dois tipos de “classificação”: a classificação automática,

uma tarefa do PLN, e a “classificação” avaliação de proficiência, uma tarefa da LA. Nesta dissertação, são descritos experimentos e é relatado um cotejo feito entre parâmetros avaliativos de um exame de proficiência em português – o exame Celpe-Bras – e parâmetros coesivos e lexicais manifestados em um conjunto de textos produzido no contexto desse mesmo exame, detectados automaticamente por ferramentas de PLN. Por meio desse cotejo, busca-se problematizar a avaliação automática de textos e, por extensão, a avaliação automática de proficiência em língua portuguesa, apresentando resultados que não são absolutos e que não encerram, de forma alguma, a questão, mas demonstrando possibilidades para o uso de ferramentas computacionais que podem vir a auxiliar o processo de correção de outros textos escritos em português.

Além disso, por apontar elementos coesivos e lexicais que, guardadas as devidas ressalvas, foram capazes de categorizar automaticamente produções textuais em diferentes níveis de proficiência, é apresentada, ao longo desta descrição, uma metodologia para o processamento automatizado de textos em língua portuguesa, que pode ser aplicada a outros trabalhos que tenham diferentes objetivos. Ao propor tal descrição e metodologia, é importante marcar de onde a autora deste trabalho fala e fazer uma breve digressão, a título de esclarecimentos e de orientação de leitura para o que aqui está exposto, pois, remontando ao(s) pensamento(s) Saussuriano(s), a língua:

É [...] um produto social da faculdade de linguagem e um conjunto de convenções necessárias, adotadas pelo corpo social para permitir o exercício dessa faculdade nos indivíduos. [...] Não se deixa classificar em nenhuma categoria de fatos humanos, pois não se sabe como inferir sua unidade. A língua, ao contrário, é um todo por si e um princípio de classificação. (SAUSSURE, 2006, p. 17)

Sendo a língua um princípio de classificação, questões relacionadas aos métodos e abordagens de pesquisa recaem nas mãos do observador, visto que “é o ponto de vista que cria o objeto” (p. 15), ou seja, é a perspectiva sobre o fato linguístico que determinará o método por meio do qual ele será investigado. A perspectiva aqui adotada, assim, é a da **probabilidade**, da **estatística** e da **quantificação**, em detrimento de tantas outras possíveis.

Essa escolha se deu porque sou tradutora e revisora de textos desde 2006 e pesquisadora em Linguística Textual e de Corpus desde 2007, e há muito tempo venho, junto a grupos de pesquisa¹, encontrando resultados que têm ajudado na descrição da língua portuguesa, na construção de ferramentas para o auxílio de produção textual e na otimização

¹ Informações sobre pesquisas anteriores, projetos e ferramentas podem ser obtidas nos sites dos projetos PorPopular, PorLexBras, Textecc e Dicionário Colaborativo (disponíveis em <<http://www.ufrgs.br/textecc/porlexbras/porpopular/>>, <<http://www.ufrgs.br/textecc/porlexbras/>>, <<http://www.ufrgs.br/textecc/>> e <<http://www.ufrgs.br/textecc/porlexbras/di/>>, respectivamente).

de tarefas de tradutores e trabalhadores do texto através da aplicação de processos automatizados. A ideia desta pesquisa foi desencadeada a partir de uma série de novos questionamentos nesse sentido levantados no início de minha prática profissional como professora de PLA, em 2010. Embora minha trajetória como professora seja relativamente curta, foi a partir da junção dessa nova prática com a prática antiga, e a coincidente entrada no mestrado, na linha de pesquisa *Teorias Linguísticas do Léxico: Relações Textuais*, que este trabalho foi delineado. Neste trabalho, portanto, imbuída de um conhecimento longamente construído e aportado teoricamente, lido com o aspecto quantitativo da língua a partir de uma perspectiva do texto.

Embora pesquisadores insistam em associar estudos quantitativos e léxico-estatísticos a estudos de vocabulário, ou em encarar trabalhos desse teor como apenas “trabalhos quantitativos com *corpus*”, quero aqui propor uma reflexão. Gostaria de deixar marcado que não é de hoje que se afirma que o aspecto quantitativo é uma das propriedades do léxico e que a frequência é uma característica típica das palavras, afirmações registradas, por exemplo, nos pioneiros trabalhos realizados em língua portuguesa por Maria Tereza Camargo Biderman (1978, 1996).

É na esteira do legado de Biderman que esta pesquisa identifica-se como um estudo em *léxico-estatística textual*. Esse enfoque sobre o funcionamento da linguagem entende que a língua em uso é observável em textos e “sua pretensão teórica consiste em **modelar a comunicação linguística como um processo de probabilidades**” (HOFFMANN, 2007, p. 61-62; *grifos meus*). O diferencial dessa *léxico-estatística textual*, quando contraposto à LC, por exemplo, conforme entendo, centra-se no papel do texto e sua ambiência, no qual um dado léxico se observe, de modo que **o texto** não seja subsumido em meio ao todo de um *corpus*. Nesse sentido, importa alertar que o enfoque estatístico deve ser entendido neste trabalho como uma **referência**, e não como um fim em si mesmo. O resultado da análise estatística deve ser compreendido como um **auxílio** e o que se obtém como resultado não pode ser tomado como medida absoluta, que imponha determinadas ações ou cerceie escolhas a despeito de quaisquer necessidades práticas ou de opções teóricas dos pesquisadores que venham a ler esta dissertação.

Imbuída dessas crenças, usei minha experiência como linguista de *corpus* que lidava com textos científicos e suas convencionalidades tentando relacionar algum tratamento estatístico dos textos produzidos por estudantes de PLA. Note-se que não há nada sobre o

assunto na literatura² e que são raras as pesquisas³ que fazem uso de *corpora* de estudantes de PLA. Entre os trabalhos existentes não há qualquer aproximação com metodologias de LC ou PLN ou quaisquer estudos probabilísticos e léxico-estatísticos sobre o tema.

De longa data, há muitos trabalhos em LC no Brasil e mais ainda no exterior que descrevem a língua produzida por estudantes com diferentes *backgrounds*, mas a ênfase tem sido para interfaces com o ensino de inglês como língua adicional, foco que não se repete no português, como é possível ver no recente livro lançado por Viana e Tagnin (2011), *Corpora no Ensino de Línguas Estrangeiras*. Assim, na tentativa de trazer a perspectiva do trabalho com *corpus* somado a técnicas de PLN, busquei avaliar como um enfoque quantitativo e computacional seria capaz de contribuir com os estudos sobre a produção escrita de estudantes de português.

Para isso, trabalho aqui com um *corpus* composto por uma amostra de produções textuais realizadas no contexto do exame de proficiência Celpe-Bras de 2006-1. Esse exame confere o Certificado de Proficiência em Língua Portuguesa para Estrangeiros e é detalhado no Capítulo 1 desta dissertação. Esse mesmo *corpus* já foi analisado a partir de diferentes pontos de vista (SCHOFFEN, 2009), considerando-se uma revisão da noção de proficiência subjacente às grades de avaliação de compreensão oral e escrita e produção escrita do exame. Ele foi útil para mostrar e problematizar como os critérios presentes nas grades de correção se relacionavam com o construto explicitado no Manual do Examinando (à época chamado de “Manual do Candidato”) e como esses critérios e construto se refletiam na avaliação dos textos.

Dito isso, é importante frisar que este trabalho não tem como objetivo discutir a confiabilidade, a validade e o construto teórico do exame Celpe-Bras, muito embora algumas questões sobre o exame e sua metodologia de avaliação naturalmente venham à tona à medida que se contrastam os dados obtidos do *corpus* e as avaliações dadas aos textos. Acredito, entretanto, que as ferramentas computacionais e os *corpora* podem trazer novas visões, mais panorâmicas, sobre a produção de escrita e uso da língua nessa situação de avaliação, e quem sabe esta pesquisa possa instigar outros trabalhos nesse sentido.

² Apenas a coleta de *corpora* de estudantes de português europeu realizada pelo Centro de Linguística da Universidade de Lisboa. Não há documentação de análises realizadas com esse *corpus*. Informações disponíveis em <<http://www.clul.ul.pt/en/resources/314-corpora-of-ple>>. Acesso em 24 nov 2012.

³ Trabalhos conhecidos desenvolvidos com *corpora* de estudantes de português brasileiro são os de Sun Yuqi (2011), sobre a produção de *hedges* por estudantes chineses, e o de Sepúlveda-Torres (2010, em andamento), sobre o desenvolvimento de ferramentas para auxiliar a escrita acadêmica de hispano-falantes em contexto de pesquisa brasileiro.

A opção por abordar os textos produzidos pelos examinandos surgiu após realizar experimentos prévios, os quais relato na próxima seção. Esses experimentos, de caráter muito inicial, forneceram pistas e a base para o desenvolvimento do restante da pesquisa. Os textos utilizados fizeram parte de estudos piloto, a fim de prospectar a viabilidade do estudo com um conjunto maior de produções textuais.

TRABALHOS ANTERIORES – EXPERIMENTOS PRÉVIOS

Durante a graduação, tive experiência com pesquisas envolvendo Linguística de Corpus, Gêneros Textuais e Linguagens Especializadas (FINATTO et al., 2010), cuja metodologia julguei interessante explorar com o PLA. Em um primeiro estudo piloto (EVERS, FINATTO e PASQUALINI, 2011), coletei produções textuais de 16 estudantes postadas no blog de uma escola privada voltada ao ensino de PLA. Processei essas produções, uma a uma, com a ferramenta de análise textual Coh-Metrix-Port⁴ (SCARTON, ALMEIDA e ALUÍSIO, 2009). Essa ferramenta reconhece e quantifica diferentes elementos presentes em um texto, e sua compreensão é fundamental neste trabalho. Ela será mais bem explicada no Capítulo 5, seções 5.2 e 5.3, mas, em linhas gerais, quantifica: número de palavras, variedade lexical, número e variedade de conectivos, entre outros.

Para a realização desse primeiro ensaio, fiz um recorte de cinco textos que foram produzidos por um mesmo estudante e que tinham semelhança temática (viagens pelo Brasil), de gênero (relato para o blog da escola) e de tamanho (número aproximado de palavras). A ferramenta Coh-Metrix-Port, adaptada dos trabalhos de Graesser et al. (2004), gera automaticamente valores para métricas e índices lexicais e coesivos de cada texto, mostrando, em um conjunto de medidas, características desse texto. Poderíamos comparar, por analogia, o resultado da ferramenta Coh-Metrix-Port ao resultado de um exame de Raios-X: o médico ortopedista precisa analisar o resultado do exame a fim de prover o melhor diagnóstico e tratamento ao seu paciente. Da mesma forma, os resultados da ferramenta Coh-Metrix-Port podem servir para que o professor analise e compare as produções escritas de seus alunos, verificando alterações de medidas, métricas e índices ao logo do tempo ou na produção de diferentes gêneros. Com esses resultados, é possível fornecer algum tipo de retroalimentação ao aluno, não em formato de prescrição, do tipo “escreva mais” ou “use mais conectores”,

⁴ Disponível em <www.nilc.icmc.usp.br/cohmetrixport>. Acesso em 22 jan 2013.

mas de análise textual conjunta. A partir da interpretação e comparação desses valores, é possível fazer inferências com relação a esse desempenho textual.

Devido ao grande número de métricas de que dispõe essa ferramenta, nesse primeiro estudo piloto, optei por trabalhar apenas com as medidas mais básicas ou superficiais, que seriam as contagens lexicais e o Índice Flesch⁵. Essas medidas e o Índice Flesch foram os elementos mais salientes na observação. Isso é o que se pode ver na Tabela 1, que mostra o desempenho da ferramenta analisando cinco textos de um mesmo estudante, produzidos ao longo de 3 meses (maio, agosto e setembro de 2011):

Métrica	Significado	Texto 01 (mai)	Texto 02 (mai)	Texto 03 (ago)	Texto 04 (ago)	Texto 05 (set)
Índice Flesch	Índice Flesch	47.07	52.19	58.49	61.06	69.99
Número de Palavras	Número de palavras do texto	397.0	438	425.0	299.0	434
Número de Sentenças	Número de sentenças de um texto	24.0	20	23.0	20.0	34
Número de Parágrafos	Número de parágrafos de um texto (quebra de linha)	10.0	11	9.0	7.0	13
Palavras por Sentença	Número de palavras dividido pelo número de sentenças	16.54	21.9	18.47	14.95	12.76

Tabela 1 Métricas do Coh-Metrix-Port para cinco textos de um mesmo estudante.

A partir desse primeiro estudo, em que acompanhei um estudante de forma qualitativa e longitudinal, foi possível perceber que os resultados apresentados pela ferramenta apontaram uma mudança de desempenho da escrita, especialmente em dois aspectos: o Índice Flesch e o número de palavras por sentença.

O Índice Flesch (FLESCH, 1948) é dado por uma fórmula matemática que gera como resultado um índice, organizado em uma escala que indica a maior ou menor facilidade de leitura de um texto. Esse índice é amplamente utilizado nos Estados Unidos para a tarefa de escolha de livros e textos direcionados a diferentes séries da escola regular. No Brasil, sua fórmula adaptada é utilizada pelos Comitês de Ética das universidades a fim de mensurar a dificuldade de leitura de formulários de consentimento livre e esclarecido. A escala do Índice Flesch possui sete faixas que determinam o grau de inteligibilidade dos textos, reproduzidas a seguir, na Figura 1:

⁵ Proposto na década de 40 pelo austríaco Rudolph Flesch, que havia fugido da Europa nazista durante a guerra e se repatriado nos Estados Unidos. Como estrangeiro, Flesch acreditava que um inglês simplificado, o *plain English*, deveria ser utilizado em documentos oficiais. Criou uma fórmula, que foi adaptada para o português em 1996 pelo Instituto de Ciências Matemáticas e de Computação da USP, campus São Carlos, para classificar textos em mais ou menos difíceis. O Índice Flesch é explicado com maior profundidade no Capítulo 5.

Valor do Índice	Leitura do texto
90-100	Muito fácil
80-90	Fácil
70-80	Razoavelmente fácil
60-70	Padrão
50-60	Razoavelmente difícil
40-50	Difícil
0-30	Muito difícil

Figura 1 Faixas do Índice Flesch.

De acordo com os dados da Tabela 1 e da Figura 1, os textos do estudante, de maio a setembro, vão passando da faixa 40-50 (de difícil compreensão) para a faixa 60-70 (de compreensão padrão). Isso significa que, de acordo com o índice, os textos ficaram mais fáceis ou, conforme documentação sobre o índice, ficaram mais inteligíveis. Além disso, o número de sentenças de seus textos aumentou, ao mesmo tempo em que o número de palavras nessas sentenças diminuiu.

É possível inferir, a partir da lógica de funcionamento dessa ferramenta, que o estudante passou a fazer frases mais curtas e aumentou os seus textos, possivelmente apresentando mais ideias e as explicando ou desenvolvendo mais, tornando o texto mais compreensível e claro do ponto de vista da observação dos valores gerados pela ferramenta.

Embora esses resultados parecessem interessantes, a quantidade de dados era insuficiente e não foi possível fazer generalizações. Para prosseguir a pesquisa, havia um grande problema a ser sanado: a falta de um *corpus* de tamanho significativo e que tivesse sido, de preferência, previamente estudado. Compilar um *corpus* de português produzido por estudantes não seria uma tarefa tão simples como fora a de coletar textos em um blog, visto que precisaria, primeiro, dos sujeitos de pesquisa e, na sequência, criar formulários para o estabelecimento de perfis de alunos e, por fim, elaborar propostas de produção textual que fossem delimitadas o suficiente para serem comparáveis entre si.

Foi a partir da busca de um *corpus* que o escopo deste trabalho foi sendo definido. A partir da descoberta de trabalhos que envolviam as produções escritas do exame Celpe-Bras, encontrei apenas dois que lidavam com uma quantidade expressiva desses textos. Em contato com os autores⁶, Schoffen (2009), que utiliza a amostragem coletada no exame de 2006-1 para o ajuste da grade de correção, ainda estava de posse dos textos estudados. Os mesmos foram gentilmente cedidos e, posteriormente, por mim digitados, para a realização deste

⁶ Outros trabalhos que analisam textos submetidos ao exame Celpe-Bras, e que apresentam descrições detalhadas da construção do exame e de suas tarefas, são os de Sidi (2002) e Gomes (2009). Damazo (2012) recentemente analisou, do ponto de vista enunciativo, produções submetidas à Tarefa I do exame de 2010-2, mapeando “comportamentos” modais e relacionando-os aos níveis de proficiência avaliados pelo exame.

trabalho. A partir desse *corpus* e de seu processamento pela ferramenta Coh-Metrix-Port, empreendi um cotejo entre características identificadas automaticamente e os escores de avaliação atribuídos a cada um dos textos da amostra. Nesse cotejo está o cerne deste trabalho.

Com os textos reunidos em um *corpus* e com os indícios obtidos no primeiro estudo piloto, foi possível elaborar questões e hipóteses de pesquisa, bem como delimitar os objetivos a alcançar. Assim, havendo *corpus*, embora não seja um *corpus* robusto em termos de tamanho – pois está bem abaixo do tamanho mítico de 1 milhão de palavras (nosso *corpus* tem um pouco mais do que 27 mil palavras) sempre citado como ideal em LC –, há o que explorar e o que mostrar.

OBJETIVOS, QUESTÕES E HIPÓTESES DE PESQUISA

O objetivo geral desta pesquisa é, conforme apresentado no início desta Introdução, verificar a viabilidade de avaliar proficiência automaticamente através de parâmetros de coesão e coerência apontados por ferramentas de PLN em textos produzidos por estudantes de PLA. O ponto de vista adotado para a descrição do *corpus* guia-se pelos princípios teóricos da LC, da Linguística Textual e do que chamamos de *léxico-estatística textual*, mas é executado com metodologias de PLN. A partir dessa junção teórica e metodológica, foi possível fazer afirmações e inferências sobre o funcionamento da *língua em uso* com base na observação de amostras de textos autênticos, que constituem o *corpus* selecionado para este estudo.

Além disso, parte-se da crença da LC de que a língua é um sistema probabilístico de combinatórias. E, como futuramente pretendo trabalhar com um número maior de textos e cruzar diferentes instrumentos avaliativos de língua, acredito ser fundamental produzir um estudo capaz de ser replicado com outros *corpora* e para outras finalidades e capaz de mostrar o que, em termos de descrição linguística, é mais ou menos custoso. Assim, por exemplo, os próprios textos das propostas de produção textual do exame poderiam compor outro *corpus* para estudo; o mesmo poderia ser feito com um *corpus* de redações de vestibular ou redações submetidas a outros processos avaliativos; ou poderíamos propor a construção de um *corpus* de estudantes de português no ensino médio, em turmas de educação de jovens e adultos; enfim, poderíamos propor o início da construção de diversos *corpora* cujos autores seriam estudantes de português com os mais diferentes *backgrounds*, com a finalidade de acompanhar seu desempenho longitudinalmente e quantitativamente.

As questões de pesquisa investigadas são as seguintes:

1. Como é a configuração lexical e gramatical de textos submetidos a um exame de proficiência do português brasileiro que tem como construto teórico avaliar o desempenho do examinando através da aplicação de um teste comunicativo?
2. Qual relação podemos estabelecer entre a configuração do léxico empregado nas produções dos examinandos e os níveis de proficiência previamente avaliados no exame?
3. A partir de um enfoque da LC e do PLN, quais são os melhores pontos de aproveitamento para pesquisas sobre a produção escrita em PLA?
4. Partindo de textos previamente avaliados por avaliadores humanos, é possível apontar e formalizar, através do Aprendizado de Máquina (AM) elementos que se correlacionam e diferenciam os níveis de proficiência?

Aproveitando trabalhos anteriores que abordam *corpora* de estudantes de língua inglesa e outras línguas⁷ que não o português (GRANGER, 1994 e 2002; TAGNIN e FROMM, 2008, CROSSLEY e MCNAMARA, 2012)⁸, algumas hipóteses formuladas e confirmadas nesses trabalhos estão aqui repetidas, agora testadas para o PLA. Embora comparar línguas não seja um objetivo deste trabalho, e seja algo que consideramos improdutivo e até inviável gramaticalmente, estudos realizados sobre a língua inglesa forneceram pistas interessantes a serem confirmadas ou refutadas em português. Uma hipótese a ser verificada, conforme apontado no estudo piloto, seria a de que o aumento do Índice Flesch, por exemplo, é um ponto distintivo de níveis de proficiência escrita, visto que quanto maior é o seu valor, maior ficaria a inteligibilidade do texto do estudante e, como consequência, sua proficiência.

Outro ponto levantado em trabalhos que envolvem *corpora* de estudantes de língua inglesa é a questão da repetição de palavras: a maioria das pesquisas aponta que, quanto mais repetitivo é um texto, ou seja, quanto menor for a sua variação lexical, maior é seu nível de proficiência e clareza (FERRIS, 2002; JARVIS et al., 2003; FINATTO et al., 2008; CROSSLEY e MCNAMARA, 2012). Esses são alguns pontos a serem tratados e testados, em português, nesta dissertação.

Assim, levando em conta os objetivos e as questões de pesquisa elencados, pretendo verificar a validade e a abrangência das seguintes hipóteses:

⁷ Projeto CoMAprend (Corpus Multilíngue de Aprendizizes), disponível em <<http://www.fflch.usp.br/dlm/comet/comaprend.html>>.

⁸ Projeto Br-ICLE (Brazilian Portuguese Sub-Corpus do ICLE, The International Corpus of Learner English), disponível em <<http://www2.lael.pucsp.br/~tony/2001bricle-interc.pdf>>.

- I. Padrões de frequência e distribuição das palavras ao longo dos textos e do *corpus* são capazes de colaborar para distinguir níveis de proficiência em português.
- II. Textos maiores tendem a ser avaliados como mais proficientes.
- III. A relação entre número de palavras e palavras diferentes presentes em um dado texto (*Type Token Ratio*) é um bom indicativo para distinção de níveis de proficiência.
- IV. O Índice Flesch é capaz de distinguir, combinado com outros índices e métricas, níveis de proficiência de português.

Diante das evidências obtidas, espero ser capaz de produzir dados úteis para fomentar a descrição do PLA produzido por estudantes com diferentes desempenhos e, indiretamente, subsidiar a avaliação das produções textuais submetidas ao exame Celpe-Bras, de modo a facilitar o trabalho dos corretores.

ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está organizada em onze capítulos, divididos em três blocos, chamados de Parte I – Objeto de Estudo, Parte II – Referenciais Teórico-Methodológicos e Parte III – Procedimentos, resultados e conclusões. Na seção de Anexos estão as tarefas do exame Celpe-Bras realizado em 2006-1 e os arquivos .ARFF, os quais são manipulados pelo *software* Weka (que contém algoritmos de aprendizado de máquina para tarefas de mineração de dados), tendo como instâncias as métricas da ferramenta Coh-Matrix-Port.

A **Parte I**, que trata do objeto de estudo, o *corpus*, possui dois capítulos: o Capítulo 1, com breve descrição do exame Celpe-Bras, incluindo dados históricos, construto teórico do exame e detalhamentos sobre o modo de aplicação e de correção das provas; e o Capítulo 2, que descreve o modo como o *corpus* foi reunido, as tarefas que deram origem a ele e informações básicas quantificáveis do *corpus*, como número de textos e de palavras.

Já a **Parte II**, que trata dos pressupostos teórico-metodológicos da pesquisa, está dividida em quatro capítulos. Esses capítulos apresentam uma discussão sobre o papel dos recursos linguísticos na avaliação de proficiência, fazendo uma relação com os conceitos trazidos pela Linguística Textual e pela LC. Segue, ainda, uma apreciação da contribuição do PLN nessa discussão, especialmente possibilitada através do uso das ferramentas Coh-Matrix-Port e do AM.

Na parte final da dissertação, a **Parte III**, estão os procedimentos para a realização da análise computacional, os resultados dessa análise, separados em lexicais e coesivos e, por fim, a discussão dos resultados e as conclusões, bem como a retomada das questões e hipóteses de pesquisa. Finalizando, comento as limitações do trabalho e expectativas de pesquisas futuras.

PARTE I – OBJETO DE ESTUDO

O objeto de estudo desta pesquisa é um *corpus* de produções textuais submetidas ao exame Celpe-Bras e os parâmetros utilizados para avaliá-los em níveis de proficiência do português. Os textos que compõem o *corpus* foram previamente avaliados⁹ por corretores humanos de acordo com os critérios de avaliação postos pelo exame. São trazidas aqui informações relacionadas ao *corpus* e suas condições de produção e de correção. Explica-se brevemente como o exame é formulado, o que faz parte de seu construto teórico e como são descritos os níveis de certificação.

⁹ Previamente realizada pelos corretores do exame Celpe-Bras, edição 2006-1.

1 O EXAME CELPE-BRAS

O exame Celpe-Bras foi elaborado e testado nos anos 90 devido à necessidade de criação de um exame que certificasse a proficiência de estrangeiros que tivessem a necessidade de se integrar à vida brasileira ou de usar o português do Brasil no exterior. Tendo como meta não somente conferir o certificado de proficiência em língua portuguesa, os idealizadores do exame o pensaram de modo que este servisse também como ferramenta para atualizar métodos e técnicas de ensino ao redor do mundo. Dessa forma, esse exame não deve ser visto apenas como um instrumento avaliativo: ele é um importante difusor da língua portuguesa, propagando um determinado modo de pensar o ensino de línguas, também refletindo concepções sobre a cultura brasileira.

O Celpe-Bras confere o Certificado de Proficiência em Língua Portuguesa para Estrangeiros; é aplicado oficialmente desde 1998, ano em que passou a fazer parte das exigências do Ministério da Educação do Brasil (MEC) para o ingresso de alunos estrangeiros nos cursos de graduação e de pós-graduação das universidades brasileiras. Os examinandos devem ser estrangeiros não lusófonos maiores de 16 anos e precisam possuir escolaridade equivalente ao Ensino Fundamental brasileiro. Desde 2001, quando o Conselho Federal de Medicina passou a exigir o certificado a fim de validar diplomas de Medicina no Brasil, é possível verificar o aumento significativo no número de inscritos no exame e a importância que seus resultados tomaram no mercado de trabalho e no meio acadêmico.

O exame era aplicado apenas uma vez ao ano até que, em 2002, passou a ter duas edições anuais. A partir de 2003, as inscrições passaram a ser feitas on-line; desde 2009, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é responsável por sua elaboração e aplicação. Para se ter uma ideia das dimensões e da relevância deste exame, o Gráfico 1 a seguir mostra o crescimento do número de examinandos inscritos desde a primeira edição, em 1998, até a segunda edição de 2012.

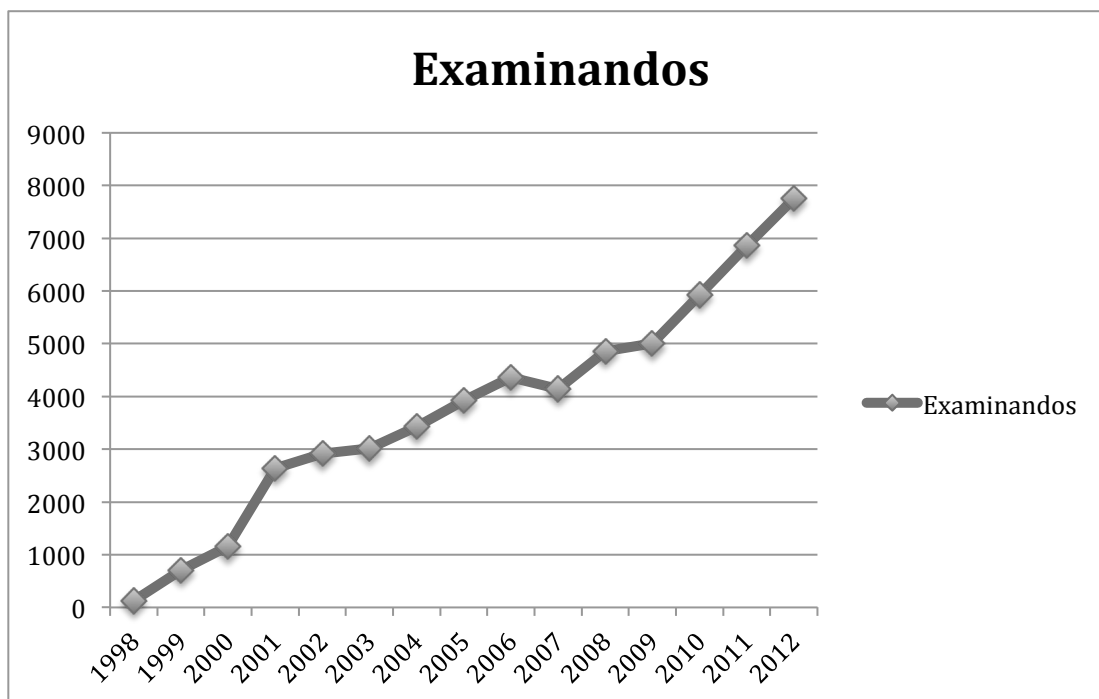


Gráfico 1 Examinandos: de 127 a 7.748 em 14 anos. Fonte: Adaptado de Damazo (2012) e atualizado com dados do INEP (2012).

Com um número cada vez maior de inscritos e de provas a serem corrigidas, ao longo dos anos, o processo de correção dos textos foi sendo modificado. Antes, era um evento organizado por e para um pequeno grupo de corretores. Tratava-se de uma comissão de professores com experiência na área de ensino PLA provenientes de diversas regiões do Brasil, na maioria das vezes vinculados a instituições credenciadas para aplicar o exame (COURA-SOBRINHO, 2009) e selecionados nos postos de aplicação. Hoje, após o anúncio de um edital em 2010 e outro em 2012 para a seleção de elaboradores de itens e de corretores, aberto a todo o país, professores de todo o Brasil e até de fora têm a oportunidade de participar dos treinamentos, trocar informações com outros profissionais da área de PLA e de participar do evento de correção dos textos produzidos para o exame Celpe-Bras. Assim, o exame Celpe-Bras de fato acaba promovendo uma verdadeira ação de formação de professores e de divulgação do construto teórico desejável ao ensino de PLA.

Além do número cada vez maior de inscritos, a relevância do exame também pode ser constatada através da observação do crescimento do número de postos aplicadores no Brasil e no exterior. É interessante notar, no Gráfico 2 a seguir, como o número de postos aplicadores praticamente duplicou desde a primeira aplicação oficial do exame em 1998. Outro fator interessante que chama a atenção é que o número de postos aplicadores fora do Brasil é, hoje, praticamente duas vezes maior do que o número de postos aplicadores que temos dentro do país.

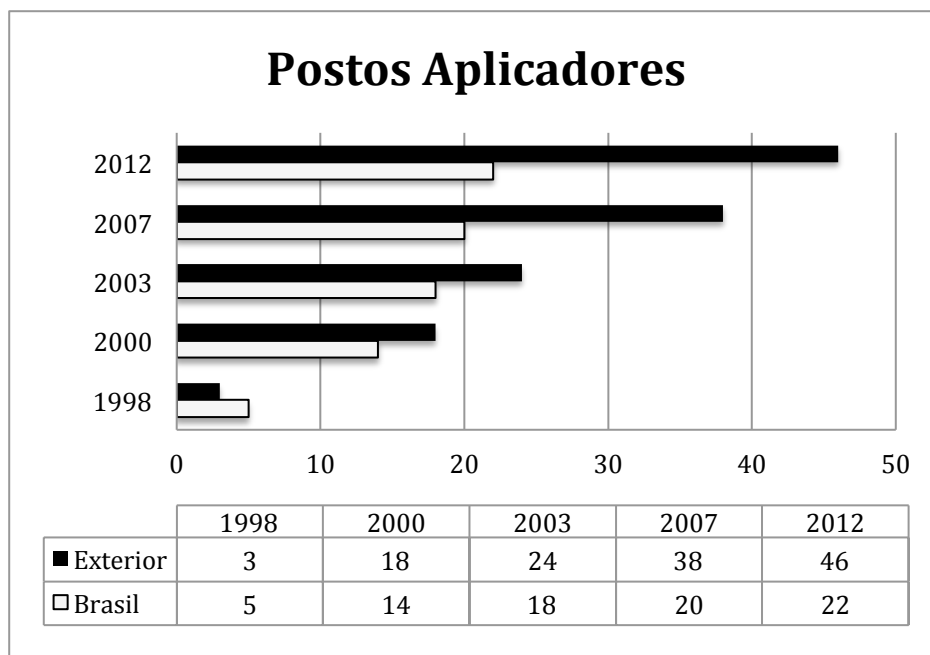


Gráfico 2 Postos aplicadores no Brasil e no Exterior: de 8 postos a 22.¹⁰

Assim, fica bastante clara a dimensão desse exame e de sua importância, bem como o volume de trabalho que a Comissão Técnica tem para elaborá-lo e coordenar as equipes de correção. Tendo em vista esse crescimento contínuo, tanto com relação ao número de inscritos quanto com relação ao número de postos aplicadores credenciados, é possível dizer que o exame está em potencial expansão. Ainda, o Brasil sediará importantes eventos esportivos nos próximos anos e a economia está em franco crescimento, fatores que acabam atraindo o interesse de estrangeiros que desejam aprender o português brasileiro para trabalhar no país ou fazer negócios em/com as empresas brasileiras e governo. Esses fatores apontam para o crescimento de examinandos em futuras edições do exame, o que resultará em mais cada vez mais textos a serem corrigidos.

1.1 CONSTRUTO TEÓRICO DO EXAME CELPE-BRAS

Como já dito antes, os textos submetidos ao exame Celpe-Bras, como em qualquer exame, têm suas produções influenciadas por uma série de critérios avaliativos. Por isso, é importante explicar aqui alguns elementos teóricos que perpassam a elaboração e aplicação deste instrumento de avaliação. O Celpe-Bras é um exame que avalia o desempenho de seus examinandos ao executarem tarefas que envolvem compreensão oral e escrita e produções

¹⁰ Fonte: Disponível em <http://portal.mec.gov.br/dmdocuments/celpbras_institcredenciadas.pdf>. Acesso em 12 fev 2013.

textuais e orais através de um único instrumento. Isso quer dizer que o examinando não escolhe o nível de proficiência em que deseja ser testado, como acontece em outros exames, conforme aponta Amaral (2011). Por exemplo, existem exames muito conhecidos que testam proficiência em língua inglesa e que oferecem duas ou mais modalidades de inscrição e testagem. Um deles é o *International English Language Testing System* (IELTS), utilizado para ingresso em universidades, mercado de trabalho ou para fins de imigração, que pode ser feito em dois formatos: *Academic* ou *General Training*, dependendo do objetivo do examinando. Outro bom exemplo é o *Certificate of Proficiency in English* (CPE), o mais avançado de um conjunto de exames para avaliar diferentes níveis de proficiência em inglês. Seu objetivo é mais genérico e é muito procurado na Europa por aqueles que querem demonstrar proficiência em inglês voltada ao mercado de trabalho. As modalidades são:

- KET (*Key English Test*) – Cambridge Level One
- PET (*Preliminary English Test*) – Cambridge Level Two
- FCE (*First Certificate in English*) – Cambridge Level Three
- CAE (*Certificate in Advanced English*) – Cambridge Level Four
- CPE (*Certificate of Proficiency in English*) - Cambridge Level Five

Nesses dois exemplos de teste, o examinando escolhe em qual nível deseja ser testado e qual tipo de teste preencherá seus objetivos pessoais – quer demonstrar proficiência para estudar, para trabalhar, para obter outra nacionalidade/cidadania, entre outros. No caso do exame Celpe-Bras, essas separações não existem. De acordo com Gomes (2009), existem tarefas mais ou menos difíceis implícita e estrategicamente selecionadas e colocadas em determinada ordem na composição do exame, mas não há modalidades diferentes do exame para quem deseja ou estudar ou trabalhar ou viver no Brasil. As duas primeiras tarefas do exame, por exemplo, são de compreensão oral e produção textual, e as produções de texto solicitadas geralmente são menos longas. Já as duas últimas tarefas do exame são de compreensão escrita e produção textual, o que resulta em produções muitas vezes mais longas, que envolvem, por exemplo, certo grau de cópia do texto provocador que faz parte da tarefa.

Assim, o desempenho dos examinandos (de compreensão e escrita) é avaliado a partir da realização de quatro tarefas e, ao final do processo, são conferidos certificados em quatro níveis de proficiência: Intermediário, Intermediário Superior, Avançado e Avançado Superior. Essas tarefas, de acordo com o Manual do Examinando, são colocadas como

um convite para interagir com o mundo, usando a linguagem com um propósito social, em outras palavras, uma tarefa envolve basicamente uma ação, com um propósito, direcionada a um ou mais interlocutores. [Um exemplo de tarefa seria]

Ler uma coluna de aconselhamento de uma revista (ação) para escrever uma carta (ação) à seção ‘Cartas do Leitor’ dessa revista (interlocutor), opinando sobre as respostas do colunista aos leitores (propósito). (BRASIL, 2013, p. 5)

Portanto, todas as tarefas do exame envolvem compreender um texto (oral ou escrito) e produzir um texto em resposta a uma determinada solicitação. Esse é um grande diferencial do exame, que avalia competências de forma integrada. Os testes mais conhecidos de proficiência normalmente avaliam o conhecimento de língua de seus candidatos de modo estruturado e compartimentado, muitas vezes não por meio de tarefas, muito menos tarefas que integrem compreensão e produção. Existem nesses testes, geralmente, questões pontuais, que têm como objetivo verificar conhecimentos gramaticais e de vocabulário. São estes conhecimentos que, via de regra, indicam em qual nível de proficiência o examinando se encaixa. Esse modelo de testagem, embora venha sendo substituído por outras abordagens (TURNER e UPSHUR, 2002), ainda é muito comum, mas não é algo que aconteça no contexto do exame Celpe-Bras.

O que o Manual do Examinando (BRASIL, 2013) informa com clareza é aquilo que o exame busca e não busca aferir:

[...] não [...] busca aferir conhecimentos a respeito da língua, por meio de questões sobre a gramática e o vocabulário, mas sim a capacidade de uso dessa língua. A competência do examinando é [...] avaliada pelo seu desempenho em tarefas que se assemelham a situações que possam ocorrer na vida real. (p. 4)

A partir desse excerto, passamos a ver que, para avaliar o que o exame chama de “uso da língua”, a avaliação é perpassada pela ideia de que aferir somente conhecimento puramente linguístico (entendendo linguístico aqui como gramatical: saber fazer concordâncias, saber utilizar a ortografia vigente, saber usar sinais de pontuação) não é indicador, sozinho, de maior ou menor proficiência. Na verdade, esta testagem procura relativizar a importância que se dá a esses indicadores e está preocupada em observar os graus de adequação das produções dos examinandos em contextos de uso, e com isso calibrar a relevância da apresentação de informações e de estruturas linguísticas de acordo com o que é proposto para a realização das tarefas que constituem o exame. Por exemplo, se a tarefa solicitar a escrita de um *e-mail* para um amigo o convidando a conhecer Porto Alegre, o examinando irá lançar mão de estruturas linguísticas que sejam adequadas a esse uso de língua, ou seja, irá produzir um texto com elementos mais ou menos formais/informais, usará o vocabulário de acordo com essa solicitação e criará um formato que corresponda ao gênero textual e-mail (COURA-SOBRINHO e DELL’ISOLA, 2009). Tudo isso entra no cálculo complexo do desempenho em uso da língua que está sendo avaliado no exame Celpe-Bras.

Além desses fatores, como mencionado, a avaliação de desempenho do exame se dá de forma integrada. Isso quer dizer que a avaliação envolve não só a produção escrita, mas a produção escrita como resultado e relacionada à compreensão de um texto oral ou escrito. Toda a produção escrita no contexto do exame Celpe-Bras é uma resposta a um texto oral ou escrito e demonstra o desempenho do examinando nessas duas modalidades. Essa questão, no entanto, é bastante complexa, como salienta Schoffen ao apontar que apenas exigir que determinadas informações estejam expressas nas produções não basta, já que

[...] para uma avaliação integrada das práticas de compreensão e produção, seria necessário que o contexto de produção construído pela tarefa justificasse a necessidade da explicitação de determinadas informações ou as considerasse como pressupostas se assim indicadas no texto produzido. Precisariamos de tarefas em que a explicitação da compreensão fosse de fato necessária para a produção do texto, e não apenas uma forma de demonstrar que o texto-base foi compreendido. (2009, p. 72)

Assim, o construto teórico do exame Celpe-Bras traz como pressuposto o conceito de que ser proficiente é usar adequadamente a língua – o que significa compreender e fazer uso de recursos linguísticos levando em consideração para quem e para quê se está fazendo esse uso. Dessa forma, o exame avalia o desempenho dos examinandos em suas práticas ao simularem uma situação de comunicação autêntica de uso da língua (compreensão e escrita), e prioriza isso em detrimento de meramente testar conhecimento linguístico ou gramatical e de vocabulário.

Isso significa que, por exemplo, se um examinando produziu textos perfeitamente adequados do ponto de vista da expressão linguística, sendo coesos e coerentes, fazendo uso impecável de tempos verbais, esse examinando não pode ser considerado plenamente proficiente apenas por dominar a norma da língua. Se seu texto estava impecável, normativamente, mas não fez aquilo que o enunciado da tarefa propunha em termos de praticar uma ação solicitada, se não demonstra compreensão do que foi lido ou ouvido no texto desencadeador da tarefa e se não conseguiu considerar interlocutor, informações a serem apresentadas, entre outros elementos, há problemas nessa produção.

Para avaliar o que o Celpe-Bras entende por proficiência, há que se considerar aspectos como propósito, interlocutor, formato, informações e outros elementos que digam respeito ao gênero e às ações demandadas pelo contexto das tarefas. Essa forma de entender proficiência de língua tem por trás a compreensão de que, sempre que falamos ou escrevemos, produzimos língua em um determinado momento, no formato de um gênero específico e dentro de um contexto que viabiliza a comunicação entre os interlocutores. Não saber realizar

essas ações adequadamente em português, no contexto do exame Celpe-Bras, significa não ser proficiente.

Avaliar proficiência dessa forma é, no entanto, uma tarefa difícil. Para entendermos como isso é operacionalizado no exame, apresentamos a seguir trechos da Tarefa I – que pode ser vista integralmente em Anexos, ao final da dissertação –, uma das quatro tarefas que faziam parte do exame aplicado em 2006-1 e que deu origem à parte dos textos do nosso *corpus*. Este é o enunciado da tarefa, que envolve assistir ao vídeo e produzir um texto específico a partir daquilo que compreender desse vídeo:

Tarefa I - FUNDAÇÃO DARCY RIBEIRO

Você vai assistir duas vezes a uma entrevista com Tatiana Memória, Presidente da Fundação Darcy Ribeiro (Documentário *O Povo Brasileiro*, Superfilmes, 2000), podendo fazer anotações enquanto assiste.

Imagine que você tenha sido convidado/a para fazer um **texto de apresentação** da Fundação Darcy Ribeiro, para ser publicado em um guia sobre centros culturais do Rio de Janeiro.

Seu texto deverá conter informações sobre **Darcy Ribeiro** e sobre a **criação e objetivos da Fundação**.

Figura 2 Enunciado da Tarefa I de 2006-1.

Já no enunciado da tarefa é possível verificar quais informações são consideradas relevantes e que, portanto, devem estar implícita ou explicitamente postas na produção textual: informações sobre Darcy Ribeiro e informações sobre a criação e objetivos da fundação que leva seu nome. Na sequência, apresento a transcrição do vídeo-base desta tarefa.

Transcrição da fala de Tatiana Memória:

O grande desejo da vida do Darcy era ser imortal. Ele desejava demais a imortalidade. Na impossibilidade de conseguir isso, ele criou em 1996 a Fundação Darcy Ribeiro com o objetivo principal de continuar trabalhando as idéias dele, as obras dele. Ele se considerava um homem de fazimentos e era. Eu acho que poucos intelectuais brasileiros tiveram uma atuação tão diversificada em tantas áreas e foram capazes de executar tanta coisa, tantas obras quanto ele foi. Ele teve um apoio indispensável pra que isso tudo acontecesse que foi o apoio dos dois governos Brizola. A Fundação foi criada com a intenção de trabalhar principalmente educação, antropologia, as obras literárias dele, manter os livros dele, principalmente, à disposição. Ele não teve nenhuma intenção filantrópica. Ele não pretendeu com a fundação exercer nenhuma atividade filantrópica. O Darcy acreditava principalmente no trabalho, e muito pouco na caridade. Ele achava que o indispensável era dar ao povo trabalho, porque com trabalho o povo teria dignidade e auto-suficiência. A Fundação é uma instituição de direito privado, sem fins lucrativos. Ela nunca recebeu valor nenhum em termos de doação. Tudo o que a fundação é hoje, esse prédio em que ela ta instalada, com dois andares, com restaurante, com salão multimídia, tudo isso foi conquistado com o trabalho de uma equipe que acredita realmente que é preciso realizar alguma coisa, que trabalha com uma enorme dedicação. São quase todas elas aqui dentro pessoas que trabalharam com o Darcy depois que ele voltou do exílio.

Figura 3 Parte do texto-base da Tarefa I de 2006-1.

A partir desse vídeo, era proposta a produção, conforme apresentada na Figura 2. Assim, o Celpe-Bras entende que ter

Proficiência implica efetivamente agir mediante o uso da linguagem. Nesta perspectiva, ler, por exemplo, significa mais do que compreender as palavras do texto. Uma leitura proficiente e crítica envolve atribuir sentidos autorizados pelo texto, selecionar informações relevantes, relacioná-las e usá-las para propósitos específicos solicitados pela tarefa do Exame. A proficiência na escrita significa usar a informação relevante e adequar a linguagem ao propósito da escrita (reclamar, opinar, argumentar etc.) e ao interlocutor (amigo, chefe, leitores de um jornal etc.), tendo-se em conta os parâmetros de textualização de diferentes gêneros discursivos: mensagem eletrônica, cartas do leitor, texto publicitário etc. (BRASIL, 2013, p. 7)

Ao fazer essa relação entre adequação do desempenho dos examinandos ao contexto de uso da língua, o desempenho no uso da língua é o que conta no momento de avaliar os diferentes níveis de proficiência. A proficiência, então, passa a ser compreendida como “capacidade de produzir enunciados adequados dentro de determinados gêneros do discurso, configurando a interlocução de maneira adequada ao contexto de produção e ao propósito comunicativo” (BAKHTIN *apud* SCHOFFEN, 2009, p. 163).

A noção de gênero, nesse sentido, está imbricada no construto teórico do exame Celpe-Bras. A noção de gênero circula, hoje, por diversas áreas do conhecimento, e esse fato corrobora uma abordagem cada vez mais multidisciplinar dos estudos de gênero. Atualmente, quando se fala em analisar gêneros, isso pode implicar em analisar textos, discursos, realizar categorização, entre outros tipos de estudos. No contexto do exame Celpe-Bras, os gêneros são os gêneros do discurso, o que significa que são “reflexos e refrações do processo de relações dialógicas estabelecidas dentro das esferas da atividade humana e necessitam ser entendidos como processos, não como produtos” (SCHOFFEN, 2009, p. 100). Em outras palavras, os gêneros do discurso no contexto do exame são as tarefas de compreensão e produção, que envolvem o estabelecimento de uma interlocução entre aquele que lê e produz e aquele que lerá o produto final, a produção textual, que é a materialidade do discurso e que é a fonte da avaliação realizada pelos corretores.

1.2 APLICAÇÃO DO EXAME CELPE-BRAS

Em 2013, a aplicação do exame Celpe-Bras está separada em duas etapas, chamadas de Parte Escrita (3 horas de duração) e Parte Oral (20 minutos de duração). Na época da coleta do *corpus*, em 2006, as etapas e faixas de proficiência diferiam um pouco. A Parte Escrita era chamada de Parte Coletiva (pois era e é realizada em uma sala, junto com outros

examinandos) e a Parte Oral era chamada de Parte Individual (pois consistia na interação entre somente um examinando e o entrevistador). Esses dois momentos tinham duração de duas horas e meia e 20 minutos, respectivamente. Como é possível perceber, em 2013 o tempo da Parte Escrita aumentou e o da Parte Oral permanece o mesmo.

Essas etapas ocorrem em momentos diferentes durante o período de aplicação do exame, e a nota mais baixa entre elas determina o desempenho dos examinandos. Essa nota final geral do examinando é definida, portanto, pela menor nota entre a nota final da Parte Escrita e a nota final da Parte Oral. Isso significa que, ao receber avaliação correspondente ao certificado Avançado na Parte Escrita e Intermediário Superior na Parte Oral, o examinando receberá o certificado Intermediário Superior porque o Celpe-Bras certifica a proficiência de forma global e entende-se que o examinando ainda não alcançou o mesmo nível de desempenho oral que tem na parte escrita. De acordo com a nota geral obtida, o examinando será encaixado em uma das seguintes faixas de proficiência mostrados no Quadro 1:

Quadro 1 Pontuação e diferença entre os níveis certificados e não certificados.

Nível	Pontuação Obtida
Sem Certificação	0,00 a 1,99
Intermediário	2,00 a 2,75
Intermediário Superior	2,76 a 3,50
Avançado	3,51 a 4,25
Avançado Superior	4,26 a 5,00

Na edição de 2006-1, contexto em que as produções textuais que constituem o *corpus* deste estudo foram produzidas, haviam 6 faixas avaliadas em vez de 5, e eram, portanto, separadas de maneira diferente. A faixa “Sem Certificação” era subdividida em “Iniciante” e “Básico”.

A Parte Oral avalia a compreensão e produção oral do examinando através de uma interação face a face realizada com a participação de um entrevistador – é uma conversa guiada por elementos desencadeadores. Este trabalho está considerando apenas a Parte Escrita do exame, a que avalia compreensão oral – através da audição de textos orais, em vídeo e áudio – e compreensão de leitura – através de tarefas de leitura de textos –, além de avaliar a habilidade de produção escrita. Sendo conformado dessa forma, o exame avalia a habilidade de compreensão através da realização de tarefas que resultam em textos de um determinado gênero discursivo solicitado pelos enunciados. Por meio dessa disposição de tarefas, as quatro habilidades dos examinandos são testadas de forma integrada, de modo que um melhor desempenho nas habilidades de produção escrita e oral depende de um melhor desempenho nas habilidades de compreensão oral e escrita. O Quadro 2 a seguir mostra o número de

tarefas, tipos de insumo presentes a cada nova edição do exame e o tempo que os examinandos possuem para respondê-las. O Quadro 2 corresponde às edições recentes do exame (2012, 2013), mas em 2006 a aplicação diferia. Em vez das tarefas I e II possuírem 30 minutos, deviam ser completadas em 25 minutos. Já as tarefas III e IV, que hoje podem ser realizadas em até 120 minutos, precisavam ser executadas em 100 minutos em 2006.

Quadro 2 Distribuição do tempo de prova entre as tarefas do exame.

Aplicação da Prova Escrita		
Insumo	Tarefa	Duração
Vídeo	I	30 minutos
Áudio	II	30 minutos
Documento escrito	III e IV	120 minutos

Em todas as tarefas, o examinando deve ser capaz de produzir textos que levem em conta os interlocutores, o propósito, as informações coerentes à proposta, a clareza, a coesão e a adequação lexical e gramatical. Como vimos anteriormente, não há questões pontuais sobre gramática e vocabulário, mas essas dimensões estão presentes e distribuídas entre os critérios de avaliação descritos na grade de correção das tarefas.

Para instruir o examinando sobre o que ele pode fazer para preparar-se para a prova, o Manual do Examinando traz algumas indicações, que são bastante amplas e genéricas, com relação aos conteúdos a serem abordados, resumidas no Quadro 3 a seguir:

Quadro 3 Relação de operações, propósitos, interlocutores, gêneros e tópicos que podem ser solicitados no exame Celpe-Bras.

Operações	<ul style="list-style-type: none"> • Reconhecer a situação de comunicação • Identificar a ideia principal do texto • Distinguir pontos principais e detalhes • Identificar o objetivo do texto • Destacar pontos relevantes • Acompanhar e registrar o desenvolvimento de um argumento • Decidir se o texto é baseado em fato, opinião, pesquisa • Reescrever informação no mesmo estilo ou em estilo diferente • Reconhecer marcas linguísticas características de diferentes gêneros
Propósitos	<ul style="list-style-type: none"> • Narrar, relatar, argumentar, expor, instruir, agradecer, pedir, opinar, comentar, expressar atitudes, confirmar, desculpar-se, informar, reclamar, justificar, persuadir, aconselhar, avisar
Interlocutores	<ul style="list-style-type: none"> • Falantes de português em geral, em situações que requerem registro formal e informal
Gêneros	<ul style="list-style-type: none"> • Textos escritos: de periódicos (jornais e revistas) – editorial, notícia, entrevista, reportagem, anúncio classificado, publicidade, cartas de leitores, horóscopo, cartuns, quadrinhos; de telegramas, cartas, bilhetes, e-mails, cartões-postais; de documentos, formulários, questionários, instruções; de mapas, roteiros, quadro de horários, calendários, programas, cardápios, recibos; de dicionários, catálogos, listas telefônicas, letras de música, legendas de filme • Textos orais: entrevistas, depoimentos, noticiários, debates, reportagens, documentários

Tópicos	<ul style="list-style-type: none"> • Indivíduo: dados pessoais (profissão, características, preferências etc.); vida familiar e social (relações entre gerações, aspectos relativos à divisão de responsabilidades, ao trabalho doméstico, à amizade, à vizinhança etc.) • Habitação (tipo de habitação e de hospedagem, localização, cômodos, móveis, utensílios, eletrodomésticos, ferramentas, serviços domésticos, consertos, compra e aluguel de imóvel) • Trabalho e estudo (características, local, instalações, deveres, direitos, horário, salário, relações entre superiores e subordinados, qualificação profissional, mercado de trabalho, entrevistas, reuniões, viagens de negócios, férias e aposentadoria, escola, universidade, bolsa de estudos, exames, estágios, profissões, perspectivas de trabalho, informatização, globalização)
----------------	---

Nota-se que, de forma alguma e em nenhum ponto, estão mencionadas questões de gramática a serem verificadas ou vocabulário específico a ser estudado.

1.3 NÍVEIS CERTIFICADOS PELO EXAME CELPE-BRAS

O desempenho do examinando na Prova Escrita e na Prova Oral é avaliado com relação aos descritores das faixas correspondentes aos seguintes níveis: Iniciante; Básico; Intermediário; Intermediário Superior; Avançado e Avançado Superior. Esses descritores buscam espelhar e graduar a qualidade do desempenho do examinando nas tarefas de acordo com os critérios de adequação ao contexto (que leva em conta o cumprimento do propósito de compreensão e de produção, o gênero discursivo e o interlocutor), adequação discursiva (coesão e coerência) e adequação linguística (uso adequado de vocabulário e de estruturas gramaticais).

O certificado somente é dado ao examinando caso seu desempenho corresponda aos níveis Intermediário, Intermediário Superior, Avançado ou Avançado Superior. A obtenção dessa certificação está condicionada ao equilíbrio entre o desempenho na Prova Escrita e na Prova Oral. Isso significa que, mesmo apresentando um desempenho oral avançado, se o examinando não alcançar um desempenho de nível intermediário na Prova Escrita, ele não obterá certificação. De acordo com o Manual do Examinando (BRASIL, 2013, p. 6), os níveis de proficiência certificados são descritos da seguinte forma:

- O **Nível Intermediário** é conferido ao examinando que evidencia um domínio operacional parcial da língua portuguesa, demonstrando ser capaz de compreender e produzir textos orais e escritos sobre assuntos limitados, em contextos conhecidos e situações do cotidiano; trata-se de alguém que usa estruturas simples da língua e vocabulário adequado a contextos conhecidos, podendo apresentar

inadequações e interferências da língua materna e/ou de outra(s) língua(s) estrangeira(s) mais frequentes em situações desconhecidas.

- O **Nível Intermediário Superior** é conferido ao examinando que preenche as características descritas no nível Intermediário. Entretanto, as inadequações e as interferências da língua materna e/ou de outra(s) língua(s) estrangeira(s) na pronúncia e na escrita devem ser menos frequentes do que naquele nível.
- O **Nível Avançado** é conferido ao examinando que evidencia domínio operacional amplo da língua portuguesa, demonstrando ser capaz de compreender e produzir textos orais e escritos, de forma fluente, sobre assuntos variados em contextos conhecidos e desconhecidos. Trata-se de alguém, portanto, que usa estruturas complexas da língua e vocabulário adequado, podendo apresentar inadequações ocasionais na comunicação, especialmente em contextos desconhecidos. O examinando que obtém este certificado tem condições de interagir com desenvoltura nas mais variadas situações que exigem domínio da língua-alvo.
- O **Nível Avançado Superior** é conferido ao examinando que preenche todos os requisitos do nível Avançado; porém, as inadequações na produção escrita e oral devem ser menos frequentes do que naquele nível.

Como se está lidando com textos produzidos no contexto do exame de 2006, os níveis utilizados neste trabalho englobam os níveis Iniciante e Básico. Esses níveis não aparecem no Manual do Examinando por não garantirem a certificação. Apesar disso, são descritos nas grades de correção das tarefas do exame, de modo que seja possível discriminar com maior clareza os diferentes níveis de desempenho. Neste trabalho, além dos quatro níveis de proficiência certificados, os textos classificados como Iniciante e Básico em 2006 (correspondentes aos Sem Certificação de 2012 e 2013) também fazem parte da análise. Esses textos foram considerados relevantes para a diferenciação global entre textos mais e menos proficientes e estavam presentes na amostra coletada, de modo que complementam a avaliação como um todo. Além disso, como poderiam ser fontes importantes para a detecção de algum fenômeno referente aos estudantes de níveis mais básicos de PLA quando comparados aos mais avançados, esses textos não foram descartados.

1.4 PROCESSO DE CORREÇÃO DO EXAME CELPE-BRAS

Tendo em vista que esta pesquisa pretende gerar dados que possam colaborar com o processo de correção do exame Celpe-Bras – verificando a possibilidade de automatização de algumas tarefas que atualmente são de competência plena dos avaliadores –, é importante explicar como se dá o trabalho humano neste processo. Nosso foco recairá na Parte Escrita, que é a fonte do *corpus* de estudo.

Na fase de correção da Parte Escrita do exame Celpe-Bras, os postos aplicadores enviam as provas para o Ministério da Educação, em Brasília. Essas provas são desidentificadas e são formadas quatro equipes, cada uma ficando responsável por uma das tarefas do exame. O primeiro passo na etapa de correção é separar a amostra que servirá para o ajuste das grades de correção da prova. Essa amostra é formada por aproximadamente 40 textos, que são avaliados pela Comissão Técnica e classificados de acordo com os critérios do exame. Essa amostra serve para que as expectativas da comissão elaboradora do Celpe-Bras possam ser readequadas aos textos produzidos de fato pelos examinandos. Em geral, dois especialistas compõem a equipe de cada tarefa e corrigem os textos produzidos. Depois dessa pré-avaliação, a Comissão Técnica desenvolve a grade de correção de cada uma das tarefas e que serão posteriormente utilizadas no treinamento dos demais corretores, servindo de guia durante todo o processo de correção.

No Evento de Correção do Exame Celpe-Bras, que dura cerca de uma semana, os corretores são reunidos presencialmente e recebem treinamento, primeiro, no grande grupo, e, logo depois, nas equipes formadas para cada tarefa com a respectiva grade de correção. Hoje, os corretores trabalham em frente a um computador, fazendo a leitura dos textos escaneados e atribuindo notas em um sistema informatizado. Cada texto é lido por dois corretores, que dão uma nota de 0 a 5 ao texto. Caso haja discrepância de 1 ponto entre as notas dadas, o texto entra na fila de correção do coordenador da equipe, que discute o texto com sua equipe quando julga necessário, até que se chegue a um consenso sobre a nota que o texto do examinando deve receber. Ao final desse processo, garante-se que a prova escrita do examinando, que contém ao todo 4 textos, seja corrigida por, no mínimo, 8 avaliadores diferentes.

Os textos produzidos na Parte Escrita do exame são avaliados a partir das grades de correção propostas pela Comissão Técnica para cada uma das tarefas. Essa grade avalia Adequação Contextual (o propósito, o interlocutor, as informações necessárias para cumprir a tarefa – que precisam ser selecionadas do texto base –, e o formato adequado ao gênero

solicitado), Adequação Discursiva (coesão e coerência) e Adequação Linguística (adequação lexical e gramatical) dos textos. Para cada item dessa escala há uma descrição do desempenho esperado em cada nível (os níveis vão de 0 a 5, sendo 0 o menos proficiente e 5 o mais proficiente) em cada uma das tarefas. Essa grade busca fazer o que Weigle (2002, p. 112) afirma que as avaliações holísticas fazem, ou seja, atribuem “uma única nota para um texto [com base] na impressão geral desse texto. Em uma sessão de avaliação holística típica, cada texto é lido rapidamente e então julgado através de uma escala que apresenta os critérios de avaliação”. Esse procedimento mimetizaria o que fazemos enquanto leitores ao lermos de fato um texto no nosso dia-a-dia.

A Tabela 2 a seguir mostra parte da grade de correção utilizada para a Tarefa I de 2006-1, tarefa que pertence ao *corpus* de estudo deste trabalho. As grades das tarefas possuem, normalmente, de 3 a 4 páginas, como é possível ver em Anexos. Cada tarefa possui uma grade que é feita após a aplicação do exame e coleta de amostras das provas realizadas. Cada grade traz o enunciado de sua tarefa, um esquema de resposta esperada elaborado pela equipe de ajuste de grade, além de algumas observações que a equipe julgou importantes de serem tratadas no treinamento dos demais corretores e observadas durante o processo de correção.

NÍVEL	5 – Avançado Superior	4 – Avançado	3 – Intermediário Superior	2 – Intermediário	1 – Básico	0 – Iniciante
GÊNERO: Formato: Texto de apresentação para guia (com título ou não). Interlocutor: Público em geral.	Adequado	Adequado	Adequado (Pode apresentar, também, interferência do gênero notícia, desde que o formato geral do texto seja de apresentação).	Adequado/Parcialmente adequado (formato híbrido: pode começar com data e nome de outra cidade, mas resgata o texto de apresentação).	Parcialmente Adequado (Pode apresentar formato híbrido como no nível 2).	Inadequado (se apresentar o formato de outro gênero. EX: carta).
GÊNERO: Propósito: Texto de apresentação sobre a Fundação Darcy Ribeiro. Informações: 1, 2, 3, 4 (informações extras/opcionais).	Adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2, 3 e 4 (Mesmo que falte algum detalhe de algum tópico).	Adequado	Adequado OU Parcialmente adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2 e 3 ou 1, 3 e 4 (Pode apresentar alguns equívocos quanto às informações OU apresentar informações incompletas).	Adequado OU Parcialmente adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2 e 3 ou 1, 3 e 4 (com frequentes problemas nessas informações OU 2 e 3 – fala apenas sobre a Fundação).	Adequado OU Parcialmente adequado OU Inadequado: Apresenta Fundação e Darcy Ribeiro com muitas (ou algumas) informações, porém equivocadas.	Inadequado Apresenta Fundação Darcy Ribeiro com muitas (ou algumas) informações, porém equivocadas. OU Não apresenta informações contidas no vídeo.
Clareza e Coesão -Uso de articuladores (conjunção, advérbio) -Concordância (verbal e nominal) -Referência (pronomes)	Texto muito bem desenvolvido com clareza e coesão.	Texto bem desenvolvido com clareza e coesão, com deslizos na construção da coesão.	1. Texto bem desenvolvido com poucos problemas de clareza e coesão. OU 2. Texto pouco desenvolvido com clareza e coesão.	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão (é necessário um mínimo de articulação do texto).	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão.	Texto pouco desenvolvido com muitos problemas graves de clareza e coesão.
Adequação Lexical e Gramatical	Raras inadequações e/ou interferências da língua materna.	Algumas inadequações e/ou interferências da língua materna.	1. Inadequações e/ou interferências da língua materna frequentes no uso de estruturas mais complexas e algumas nas elementares. OU 2. Poucas inadequações e/ou interferências da língua materna.	Inadequações e/ou interferências da língua materna frequentes no uso de estruturas complexas e elementares.	Inadequações e/ou interferências da língua materna mais frequentes.	Muitas inadequações e/ou interferências da língua materna.

Tabela 2 Grade de Correção da Tarefa I de 2006-1: Fundação Darcy Ribeiro.

Dessa forma, um texto Avançado Superior, levando em consideração a Tarefa I de 2006-1 (Darcy Ribeiro) e a grade apresentada na Tabela 2 e nos Anexos, seria o texto abaixo:

A fundação Darcy Ribeiro: para alcançar a imortalidade do seu trabalho...
 Entre todos os centros culturais do Rio a fundação Darcy Ribeiro destaca-se pela sua originalidade.
 O Darcy Ribeiro nasceu em 1922 no Minas Gerais. Ele foi um intelectual brasileiro com atuação muita diversificada nas áreas de educação e antropologia. Ele também foi um escritor e publicou livros tais como “Maíra” e “Os Índios e a Civilização”. Seu trabalho recebeu o apoio do governo. O desejo mais forte do Darcy era a imortalidade. Diante da impossibilidade de ser imortal, ele resolveu a criar uma fundação em 1996 para dar continuidade às suas ideias e ao seu trabalho, manter os seus livros.
 A fundação é uma entidade de direito privado sem objetivo lucrativo. Nunca recebeu nenhuma doação, nenhuma verba. O patrimônio da fundação consiste principalmente em um prédio de dois andares localizado na cidade do Rio e onde ela é sediada.
 A fundação não tem objetivo filantrópico. O Darcy Ribeiro acreditava pouco a caridade, mas acreditava sim ao trabalho. Ele achava que o essencial era dar trabalho ao povo. Com trabalho o povo tem autosuficiência.
 Hoje a equipe da fundação é constituída principalmente de antigos colegas do Roberto Darcy que trabalhavam com ele e empolgam-se em continuar o seu trabalho.
 Vale uma visita!

Texto 1 T01E03.

E um texto Iniciante seria:

Florianópolis, 26 de abril de 2006
 Texto: Informação sobre criação e objetivos da fundação Darcy Ribeiro
 Fundação Darcy Ribeiro foi criado em 1996. Objetivo da fundação mostrar povo brasileiro que pode conquistar tudo, através do trabalho. Fundação executo obras filantropicas, sem reseber nem um doação, so teve apoio de duas governo de Brizol. Tudo foi conquistada pelo grupo de trabalho.
 Hoje fundação tem um prédio de dois andares, onde grupo continua trabalhando, para que objetivos da Fundação sejam realizadas.

Texto 2 T01E42.

As grades dos exames aplicados em 2012-2 e 2013-1¹¹ não apresentam diferença na elaboração dos descritores da parte de baixo da grade, embora seja solicitado exaustivamente que os corretores atentem ao fato de que justamente esses descritores devem ser levados em consideração no todo dos textos produzidos e no todo da realização da tarefa de compreensão e produção. Evidentemente, os itens separados que são alterados entre as tarefas são os itens do eixo gênero, que compreendem o formato, o interlocutor, o propósito e as informações. Essas características são particulares de cada tarefa, assim como Clareza e Coesão e Adequação Lexical e Gramatical, muito embora os últimos sejam apresentados sempre da mesma forma na grade.

A gradação entre os níveis permanece igual na descrição das grades de todas as tarefas naqueles eixos que ficam mais abaixo da grade (Coesão e Coerência e Adequação Lexical e Gramatical). A gradação descrita referente a esses aspectos é subjetiva mas, ao mesmo tempo,

¹¹ Observado *in loco*.

tende ao analítico se considerarmos termos como “texto muito bem desenvolvido”, “texto pouco desenvolvido”, “alguns problemas”, “muitos problemas” ou “problemas graves”, presentes na grade de 2006-1, que não podem ser mensurados objetivamente e que não devem ser contados, já que esse tipo de procedimento não faz parte do escopo do exame. O entendimento desses termos espelha o gênero discursivo solicitado na tarefa, é relativo à extensão do texto produzido e, embora as visões sejam afinadas nas formações de corretores, esses termos estão inegavelmente dependentes da formação do corretor e das discussões feitas pelas equipes de correção das tarefas a cada novo evento de correção.

Existem boas sugestões para lidar com essa subjetividade em outros trabalhos, inclusive que esses eixos sejam excluídos das grades, mas versões de 2012 e 2013 da grade ainda reproduzem os mesmos eixos e termos: “algumas”, “raras” e “frequentes”, expressões como “pouco” ou “estruturas complexas e elementares”, o que, pelo lado positivo, mantém a subjetividade e ideia de se discutir o que seriam esses termos para cada nova tarefa proposta mas que, por outro lado, parece tornar lícita a contagem de elementos.

É evidente que a subjetividade é inerente a uma avaliação holística como a pretendida pelo exame Celpe-Bras. No entanto, há que se gerir essa subjetividade, e parece que a grade de correção não está explícita o suficiente para auxiliar a tarefa. Apesar de ser considerada uma grade de correção holística, o que significa estar orientada de modo a não separar os critérios ou observá-los e pontuá-los isoladamente, é possível perceber uma clara separação dessa grade em eixos. De acordo com Schoffen (2009), a grade está de fato dividida em três eixos (Adequação Contextual, Adequação Discursiva e Adequação Linguística), de modo a facilitar a correção, mas a orientação aos corretores é a de que não interpretem esses eixos como notas a serem dadas em separado, ou itens a serem analisados e aferidos.

Fica claro que a grade não deve ser preenchida com pontuações, mas utilizada como um guia, que orienta a tarefa dos corretores. O papel do corretor é interpretar esse guia e relacionar a produção do examinando à descrição da grade de correção que melhor reflita o desempenho dessa produção, o que vai ser traduzido em uma pontuação global de 0 a 5, correspondente aos seis níveis avaliados pelo exame.

Por ser um guia, a disposição dos eixos e o próprio construto teórico do exame apontam para uma hierarquia de importância de eixos. Uso “hierarquia” não por entender que determinados aspectos são mais importantes do que outros, mas para apontar que sua disposição no papel direciona a leitura que se espera que os corretores façam. Ao observarmos as grades de correção e retomarmos os pressupostos teóricos do exame, fica bastante claro que o eixo nevrálgico a ser avaliado na grade de correção, e o responsável por

envolver a avaliação de uso da língua em situações de comunicação, é o eixo de Adequação Contextual, o primeiro dos três dispostos nas tabelas. Esse eixo é tão complexo e relevante que possui subdivisões, de modo a impossibilitar que o corretor fuja da observação do texto como um todo singular. Esse tipo de leitura tende, também, a mimetizar a leitura real que fazemos de qualquer texto, na tentativa de entendê-lo e de estabelecer uma comunicação com quem o escreveu. Fica em segundo plano, portanto, a observação de questões de expressão linguística.

Assim, o primeiro eixo recebe destaque na grade quando comparado aos demais eixos. Os eixos que estão abaixo da grade são, por outro lado, descritos de forma estática e não variam de uma grade a outra. Entende-se, nesse sentido, que seu gradual pode ser uma constante capaz de fornecer pistas sobre a constituição textual mais ou menos proficiente, para fins de descrição linguística e formalização pensando em uma tarefa de automatização. Retomando os eixos:

- **O Eixo de Adequação Contextual ou Gênero**, o primeiro encontrado de cima para baixo na grade, dá espaço para a observação do desempenho do examinando ao cumprir com o propósito da tarefa, ao endereçar o interlocutor esperado e solicitado no enunciado da tarefa, ao selecionar as informações necessárias e descartar as informações nem tão necessárias para o cumprimento do propósito solicitado e o formato do gênero. Este eixo é SEMPRE adaptado às tarefas de cada nova aplicação do exame, e tem seus descritores modificados de tarefa a tarefa.
- **O Eixo de Adequação Discursiva ou Clareza e Coesão**, o segundo de cima para baixo na grade, orienta a observação da clareza e da coerência do texto produzido, salientando a observação do uso de articuladores como conjunções e advérbios, a concordância (verbal e nominal) e a coesão referencial, por meio de pronomes.
- **O Eixo de Adequação Linguística ou Adequação Lexical e Gramatical**, o último eixo presente na grade, orienta para a observação da adequação lexical e gramatical do texto produzido, salientam a presença ou ausência de interferência de outras línguas na produção em português do examinando.

A orientação de leitura da grade de correção é de que seja feita na diagonal, da esquerda para a direita e de cima para baixo. Dessa forma, observando a grade da Tarefa I a seguir (Figura 4), vemos que o corretor é orientado a partir da leitura daquilo que é positivo no texto, verificando o grau de adequação daquele texto, em primeiro lugar, com relação à contextualização da produção. Isso quer dizer que, se o texto é adequado no primeiro eixo, ele começará como Avançado Superior, com a possibilidade de “ir caindo” de acordo com a

observação dos outros eixos, que são complementares, nesse caso, decisivos para a classificação do texto.

NÍVEL	5 – Avançado Superior	4 – Avançado	3 – Intermediário Superior	2 – Intermediário	1 – Básico	0 – Iniciante
GÊNERO: Formato: Texto de apresentação para gêneros (com título ou não). Interlocutor: Público em geral.	Adequado	Adequado	Adequado (Pode apresentar, também, interferência do gênero notícia, desde que o formato geral do texto seja de apresentação).	Adequado/Parcialmente adequado (formato híbrido: pode começar com data e nome de outra cidade, mas resgata o texto de apresentação).	Parcialmente Adequado (Pode apresentar formato híbrido como no nível 2).	Inadequado (se apresentar o formato de outro gênero. EX: carta).
GÊNERO: Propósito: Texto de apresentação sobre a Fundação Darcy Ribeiro. Informações: 1, 2, 3, 4 (informações extras/opcionais).	Adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2, 3 e 4 (Mesmo que falte algum detalhe de algum tópico).	Adequado	Adequado OU Parcialmente adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2 e 3 ou 1, 3 e 4 (Pode apresentar alguns equívocos quanto às informações OU apresentar informações incompletas).	Adequado OU Parcialmente adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2 e 3 ou 1, 3 e 4 (com frequentes problemas nessas informações OU 2 e 3 – fala apenas sobre a Fundação).	Adequado OU Parcialmente adequado OU Inadequado: Apresenta Fundação e Darcy Ribeiro com muitas (ou algumas) informações, porém equivocadas.	Inadequado Apresenta Fundação Darcy Ribeiro com muitas (ou algumas) informações, porém equivocadas. OU Não apresenta informações contidas no vídeo.
Clareza e Coesão -Uso de articuladores (conjunção, advérbio) -Concordância (verbal e nominal) -Referência (pronomes)	Texto muito bem desenvolvido com clareza e coesão.	Texto bem desenvolvido com clareza e coesão, com deslizes na construção da coesão.	1. Texto bem desenvolvido com poucos problemas de clareza e coesão. OU 2. Texto pouco desenvolvido com clareza e coesão.	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão (é necessário um mínimo de articulação do texto).	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão.	Texto pouco desenvolvido com muitos problemas graves de clareza e coesão.
Adequação Lexical e Gramatical	Raras inadequações e/ou interferências da língua materna.	Algumas inadequações e/ou interferências da língua materna.	1. Inadequações e/ou interferências da língua materna frequentes no uso de estruturas mais complexas e algumas nas elementares. OU 2. Poucas inadequações e/ou interferências da língua materna	Inadequações e/ou interferências da língua materna frequentes no uso de estruturas complexas e elementares.	Inadequações e/ou interferências da língua materna mais frequentes.	Muitas inadequações e/ou interferências da língua materna.

Figura 4 Grade de Correção e orientação de leitura.

Partindo dessa orientação, entende-se que existe uma relação de importâncias entre os eixos, muito embora a avaliação seja holística. O eixo determinante, de acordo com essa orientação de leitura, é o eixo Adequação Contextual, que engloba Gênero, Formato, Propósito e Informações. Esse eixo é fundamental para determinação da certificação ou não certificação dos examinandos, visto que os textos inadequados, de acordo com os critérios presentes neste eixo, partirão diretamente dos níveis 0 e 1, ou seja, sequer serão considerados no espectro de certificação e não poderão subir na diagonal.

Como foi possível verificar, o exame Celpe-Bras, conforme já mencionado, não tem como objetivo fornecer um diagnóstico de proficiência, mas sim testar as habilidades de forma integrada de modo que seja possível medir a proficiência dos examinandos em situações de uso da língua. Para isso, faz a avaliação do texto como um todo, o que é possível através da avaliação realizada com a orientação de uma grade de correção holística. A avaliação holística, dessa forma, faz o corretor atentar para os pontos positivos do texto, e não para os erros, inadequações ou deslizes que o examinando possa estar cometendo. Isso é o que fazemos quando lemos um texto em situações comunicativas reais: tentamos compreender, passando por cima de eventuais problemas linguísticos.

Dessa forma, as produções textuais do exame são lidas e avaliadas, em primeiro lugar, pensando se o texto está adequado ou não ao Gênero Textual indicado pela tarefa, e se essa adequação é total ou parcial. A partir daí, é fundamental, então, observar os outros dois eixos, de Adequação Discursiva e Linguística, porque são eles que vão determinar e diferenciar significativamente os níveis de proficiência dos examinandos ao praticarem aquele ato comunicativo.

Na prática, isso significa que a Adequação Contextual, no processo de correção, é observada e sua avaliação aplicada durante a leitura do texto antes da Adequação Discursiva e da Adequação Linguística. Essa ordem de aplicação dos critérios apresentados na grade faz com que o nível de proficiência determinado pela Adequação Contextual só possa ser diminuído ao se levar em conta os recursos discursivos e linguísticos presentes no texto, não sendo possível elevar o nível avaliado na Adequação Contextual somente por conta desses recursos. Essa ordem também reflete a priorização da função que o texto desempenha no contexto de produção proposto ao invés da sua adequação composicional e linguística. Nesse sentido, podemos observar também que, para os níveis certificados, é exigido que o texto cumpra pelo menos parcialmente o critério Adequação Contextual. Um texto que apresenta inadequação neste eixo é considerado Básico ou Sem Certificação.

Essa grade ou escala é entendida, conforme exposto, como um esquema, que descreve os diferentes níveis de proficiência do português no contexto do exame Celpe-Bras. Esse esquema desenha um *continuum*, que abarca desde o uso menos proficiente da língua até o uso mais proficiente. Para preencher esse esquema, durante a etapa de ajuste da grade de correção, os membros da comissão técnica buscam na amostra de textos coletada exemplos textuais e concretos que sirvam de modelo às descrições dessa grade.

Assim, os modelos e a sugestão de modo de leitura desse esquema faz com que o leitor-avaliador procure o que há de positivo no texto, em detrimento de marcar erros. Fazendo isso, acredita-se que os examinandos estejam recebendo uma nota por aquilo que conseguem fazer bem em suas produções textuais a partir de uma impressão geral que o avaliador tem do texto, guiada pela grade de correção.

2 O CORPUS DE ESTUDO

2.1 DADOS DE COMPOSIÇÃO DO CORPUS

Segundo Sánchez e Cantos (1996, p. 260) uma definição completa de *corpus* seria

Conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos, de maneira que sejam representativos do uso linguístico, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise.

O Corpus Celpe-Bras 2006, como já dito, foi previamente estudado por Schoffen (2009). O objetivo do trabalho mencionado era o de analisar a validade de construto da avaliação de compreensão oral, leitura e produção escrita do exame Celpe-Bras, verificando a noção de proficiência operacionalizada na Parte Coletiva do exame. A finalidade do estudo era a de apontar caminhos para o aumento dessa validade, e para isso foram analisados os textos. A análise problematizou os critérios dos pontos de corte para os níveis de proficiência, em que constatou-se que a recuperação de informações do texto base e a adequação lexical e gramatical foram avaliadas muitas vezes desvinculadas dos demais aspectos avaliados. De acordo com Schoffen,

a avaliação da compreensão deu-se de forma relativamente autônoma em relação à avaliação da produção, e os recursos linguísticos foram muitas vezes avaliados dissociados da sua contribuição para o cumprimento do propósito comunicativo, revelando limitações na operacionalização do conceito de uso da linguagem por não conseguir relacionar todos os componentes do contexto de recepção e de produção sugeridos pela tarefa na grade de avaliação e no processo de correção. (2009, p. 5)

Assim, selecionar textos de acordo com determinados critérios e avaliar se esses textos seriam suficientemente extensos e representativos de um dado uso da língua foi uma questão resolvida naquele trabalho e, para além dele, pela Comissão Técnica do exame, que considerou a amostra suficiente para realizar o ajuste da grade de correção.

Antes do processo de correção dos textos submetidos ao exame Celpe-Bras, ocorre o ajuste das grades, o que significa verificar se os descritores nelas contidos são adequados para avaliar os textos de fato produzidos pelos examinandos em cada nova edição do exame. As

provas são desidentificadas, as tarefas são separadas e os textos são colocados em ordem numérica, divididos em pacotes. É formada uma amostra dos textos de cada tarefa (cerca de 40 textos, provenientes de diferentes postos aplicadores) e uma equipe de corretores especialistas procede à testagem e ao ajuste da grade de avaliação. As grades ajustadas pela Comissão Técnica são testadas, posteriormente, pelas equipes de correção da prova escrita, e podem inclusive passar por uma juste mais fino, como forma de diminuir a subjetividade do processo avaliativo, a partir da compreensão coletiva dos critérios e parâmetros de atribuição de notas aos desempenhos dos examinandos.

Nesse sentido, parto do pressuposto de que, se esses textos foram utilizados para ajustar a grade de correção do exame, tanto seu número quanto seu conteúdo são suficientes para descrever com certa segurança as características das produções textuais oriundas do exame Celpe-Bras de 2006-1 e o que foi considerado ser proficiente naquela edição. Dessa forma, essa amostra de textos foi considerada suficiente para representar os níveis certificados e não certificados do exame, esse *continuum* de proficiência exposto na grade de correção, já que a própria Comissão Técnica assim os julgou.

Os textos da amostra coletada para o ajuste da grade de correção do exame Celpe-Bras de 2006-1 foram digitados por Evers et. al. (2011) para a realização desta dissertação. As produções textuais foram digitadas em 2011, de modo que as inadequações¹² existentes nos textos foram mantidas e seus autores não foram identificados. As inadequações não foram marcadas com qualquer etiqueta, e o *corpus* também não foi etiquetado por qualquer *parser*¹³. Atualmente, estão disponíveis nos formatos .doc, .txt e .html. Cada arquivo foi identificado como o exemplo: T01E01, em que “T” significa TAREFA, seguido pelo número da tarefa (que pode ser 01, 02, 03 ou 04); e “E” significa EXAMINANDO (variando de 01 até 46). Os “E” de numeração mais baixa – como E01, E02, E03 – correspondem aos textos classificados pelos corretores humanos como mais avançados (Avançado Superior, Avançado). Os “E” mais altos – como E41, E42, E43 – correspondem aos textos avaliados pelos corretores como textos menos proficientes (Iniciante e Básico).

Este *corpus* é composto por uma amostra de textos previamente utilizada pela comissão avaliadora do exame de proficiência Celpe-Bras referente à edição do primeiro

¹² No processo de digitação do *corpus* foram mantidas as inadequações de ortografia, gramática e semântica. Essas inadequações não receberam nenhum tipo de marcação. Também mantivemos a separação entre parágrafos e eventuais incompletudes dos textos, tal como se apresentaram originalmente.

¹³ Um *parser* é um analisador automático que fornece informações morfossintáticas sobre as palavras de um *corpus*. A frase “O chimarrão está pelando!”, por exemplo, após ser processada por um *parser*, recebe etiquetas e fica assim: **o** [o] <artd> **DET M S @>N** **chimarrão** [chimarrão] <n> **ADJ M S @SUBJ>** **está** [estar] <fmc> **V PR 3S IND VFIN @FAUX pelando** [pelar] **V GER @IMV @#ICL-AUX<!**. Um *parser* disponível on-line para testagem é o do projeto VISL <<http://beta.visl.sdu.dk/visl/pt/parsing/automatic/parse.php>>.

semestre de 2006. Essa amostra foi utilizada por Schoffen (2009) para a realização do ajuste das grades de avaliação e para o treinamento dos corretores do exame daquela edição, de modo que os textos foram avaliados por corretores humanos e classificados de acordo com seis blocos: Iniciantes, Básicos, Intermediários, Intermediários Superiores, Avançados e Avançados Superiores.

2.2 TAREFAS QUE ORIGINARAM O CORPUS

Os textos que compõem o *corpus* foram produzidos na Parte Escrita do exame, que se dá, conforme explicado anteriormente, através da aplicação de quatro tarefas que englobam compreensão oral e escrita e produção de texto. Na edição de 2006-1 do exame, que originou o *corpus* sob estudo, as tarefas foram:

- **Tarefa I:** os candidatos assistiram a um vídeo sobre a Fundação Darcy Ribeiro, a qual era apresentada pela presidente da Fundação, Tatiana Memória. A partir desse vídeo, foi proposta a seguinte tarefa: “Imagine que você tenha sido convidado para fazer um texto de apresentação da Fundação Darcy Ribeiro, para ser publicado em um guia sobre centros culturais do Rio Janeiro. Seu texto deverá conter informações sobre Darcy Ribeiro e sobre a criação e objetivos da Fundação”.
- **Tarefa II:** os candidatos ouviram o trecho de uma entrevista de um programa da rádio “Revista CBN”, cujo tema era a divulgação do programa “Famílias Acolhedoras”, desenvolvido por uma organização não-governamental (ONG). A partir dessa audição, foi proposta ao candidato a seguinte tarefa: “Convencido/a da importância do projeto, escreva uma carta aberta aos moradores de seu bairro, estimulando o cadastramento de famílias acolhedoras. Sua carta deverá explicar o que é e como funciona o projeto e quais são as motivações das famílias que dele participam”.
- **Tarefa III:** os candidatos leram uma reportagem da “Revista Época” intitulada “Você sabe o que está comendo?”, que trata sobre os mitos aos quais estão expostas as pessoas que se alimentam exclusivamente de alimentos naturais, produtos não industrializados, produtos *light* e *diet* e que não comem carne. A partir desse texto, foi proposta ao candidato a seguinte tarefa: “Com base na reportagem da ‘Revista Época’, de 25 de julho de 2005, escreva um e-mail a uma amiga que consome apenas alimentos naturais, alertando-a sobre os mitos a respeito desse tipo de alimentação”.

- **Tarefa IV:** os candidatos leram um texto da revista “Você S.A.” que falava sobre o hábito das empresas de controlarem o e-mail dos funcionários e mencionava uma decisão do Tribunal Superior do Trabalho que autorizou as empresas a agirem dessa maneira. A partir desse texto, era proposta a seguinte tarefa: “Com o intuito de estimular a discussão sobre a decisão do Tribunal Superior do Trabalho (TST) apresentada na reportagem da ‘Revista Você S.A.’ (junho de 2005), escreva um texto para ser afixado no quadro de avisos de sua empresa, argumentando contra a invasão de privacidade”.

O conteúdo dessas tarefas pode ser visto em Anexos. No total, são 177 produções textuais resultantes dessas tarefas. As produções textuais foram corrigidas pela comissão técnica do exame, que as classificou de acordo com os seis níveis estabelecidos pelo instrumento avaliativo (de menor proficiência a maior proficiência: Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior). Sempre lembrando que a certificação apenas é dada a partir do nível Intermediário. Dessa forma, os textos avaliados como Iniciante e Básico existem enquanto níveis, são diferentes entre si e servem como amostragem para que os corretores compreendam o que são textos completamente inadequados do ponto de vista Contextual, embora, ainda assim, seja níveis de expressão.

2.3 INFORMAÇÕES BÁSICAS DO CORPUS

Ao final da digitação do *corpus*, algumas contagens básicas, a fim de levantar dados preliminares e quantificáveis, foram levantadas para ter-se uma ideia da extensão do conjunto de dados e de sua distribuição. A Tabela 3 a seguir mostra o que convencionamos chamar neste trabalho de Classe (níveis de proficiência avaliados pelos corretores no ajuste da grade de correção do exame de 2006-1), Tarefa (referente às tarefas aplicadas no exame em questão) e número bruto de produções textuais por tarefa e por nível de proficiência. Além disso, os números entre parêntesis – como em Iniciante (6) – serão relevantes para as demais tarefas de classificação descritas ao longo deste trabalho, servindo de etiquetas para as classes. Os números mais altos representam os níveis menos proficientes e os mais baixos os mais proficientes:

Classe	Tarefa 1	Tarefa 2	Tarefa 3	Tarefa 4	Total
Iniciante (6)	2	0	1	1	4
Básico (5)	9	7	9	5	30
Intermediário (4)	9	9	23	8	49
Intermediário Superior (3)	11	13	6	13	43
Avançado (2)	5	11	4	9	29
Avançado Superior (1)	8	6	2	6	22
TOTAL	44	46	45	42	177

Tabela 3 Classes e números brutos de textos por Tarefa e Nível.

Assim, o *corpus* possui um total de **177 textos**, sendo que, quando agrupados, uma minoria de textos faz parte da Classe Iniciantes. A grande concentração de textos fica entre os níveis intermediários (Intermediário e Intermediário Superior), sofrendo uma queda, novamente nos níveis mais avançados.

Com relação ao número de palavras do *corpus*, a Tabela 4 a seguir mostra as quantidades absolutas de palavras nas classes (níveis de proficiência) e tarefas analisadas na amostra coletada:

Classe	Tarefa 1	Tarefa 2	Tarefa 3	Tarefa 4	Total
Iniciante (6)	295	0	50	146	491
Básico (5)	999	1023	1191	955	4168
Intermediário (4)	1065	1186	4095	675	7021
Intermediário Superior (3)	1474	1994	1105	1972	6545
Avançado (2)	685	2074	750	1546	5055
Avançado Superior (1)	1343	1209	425	1066	4043
TOTAL	5861	7486	7616	6360	27323

Tabela 4 Número absoluto de palavras por Tarefa e Nível.

Para uma melhor visualização, a Tabela 5 a seguir mostra os mesmos números acima, porém, normalizados¹⁴, para fins de comparação:

Classe	Tarefa 1	Tarefa 2	Tarefa 3	Tarefa 4	Total
Iniciante (6)	147.5	0.0	50.0	146.0	122.8
Básico (5)	111.0	146.1	132.3	191.0	138.9
Intermediário (4)	118.3	131.8	178.0	84.4	143.3
Intermediário Superior (3)	134.0	153.4	184.2	151.7	149.9
Avançado (2)	137.0	188.5	187.5	171.8	174.3
Avançado Superior (1)	167.9	201.5	212.5	177.7	183.8

Tabela 5 Número normalizado de palavras por Tarefa e Nível

Dessa forma, o total de palavras deste *corpus* é de **27.323** palavras, sendo que temos uma disparidade entre o número de palavras em classes mais avançadas e menos avançadas.

¹⁴ A normalização de dados é utilizada, geralmente, quando existem números muito diferentes que precisam ser comparados entre si. A normalização reduz a redundância e as chances de dados se tornarem inconsistentes.

Tendo em vista essa disparidade, procedemos a normalização desses números, de modo que as classes e resultados pudessem ser comparáveis. As etapas de normalização serão explicadas no Capítulo 7, que relata os procedimentos do trabalho.

PARTE II – REFERENCIAIS TEÓRICO-METODOLÓGICOS

Compreendendo de modo geral como funciona o exame Celpe-Bras e o que está envolvido na constituição do *corpus* de estudo desta dissertação, passo agora a apresentar os referenciais teórico-metodológicos. São teorias e métodos de pesquisa considerados relevantes para o desenho de análise que se buscou fazer nesta pesquisa. Esses pressupostos também justificam as escolhas de pesquisa feitas e sustentam as conclusões apresentadas ao final do estudo.

3 LINGUÍSTICA TEXTUAL

3.1 O PAPEL DOS RECURSOS LINGUÍSTICOS NA AVALIAÇÃO DE PROFICIÊNCIA

Como visto anteriormente, a proficiência em língua portuguesa no contexto do exame Celpe-Bras é aferida a partir de um determinado uso da linguagem. Ser proficiente, no contexto do exame, não é demonstrar conhecimentos gramaticais, fazer uso de um vocabulário amplo ou utilizar construções complexas e raras em língua portuguesa. Ser proficiente, nesse contexto, é usar adequadamente esses recursos linguísticos do português a fim de desempenhar uma *ação no mundo*. Assim, os recursos linguísticos estão a serviço do ato de conseguir desempenhar essa *ação no mundo*.

No entanto, para executar qualquer tarefa que pretenda subsidiar um processo de avaliação de modo automático precisa encontrar algum ponto que possa ser quantificado, neste caso, na materialidade da língua, o uso da língua. Este parece ser o grande desafio ao lidar com um *corpus* de textos de um exame com o perfil como o do Celpe-Bras. Não se pretende desconfigurar ou desvirtuar o construto teórico construído em torno deste instrumento avaliativo para esse fim. Esse grande desafio é inerente aos processos de avaliação, que muito dificilmente podem ser feitos em sua totalidade automaticamente. Julgamentos que envolvam subjetividade vão ser inevitavelmente complexos e envolver a capacidade de interpretabilidade do avaliador. Essa interpretabilidade sempre vai estar sujeita a discordâncias e nunca vai ser absoluta.

Em virtude dessas dificuldades postas, nos momentos iniciais desta pesquisa chegou-se a fazer uma tentativa de marcação das inadequações presentes nos textos. A princípio, essas marcações estavam mais voltadas à questão das regularidades dos gêneros solicitados pelas tarefas do exame de 2006-1. Depois, passou-se à tentativa de marcar as interferências de outras línguas através do etiquetamento de palavras que apresentavam grafia híbrida, um enfoque voltado para a ortografia e a semântica.

Porém, após avaliar resultados prévios desta análise, notou-se que, fazendo esses tipos de anotações, estar-se-ia criando novos critérios para a avaliação dos textos, fazendo algo que

os próprios corretores do exame não fazem. Por meio da marcação desses tipos de inadequações, compreendi que estaria observando “problemas” que factualmente não fazem parte do escopo de correção do exame e da avaliação de desempenho proposta pela Comissão Técnica do Celpe-Bras e, dessa forma, estaria fugindo dos objetivos principais deste trabalho – por exemplo, apontar erros de ortografia, de pontuação, de concordância, de semântica, etc.

Após essas primeiras tentativas, passou-se a atentar mais para a leitura e interpretação das grades de correção das tarefas do exame. Notou-se que essas grades possuem uma separação explícita. Muito embora a avaliação do exame Celpe-Bras seja considerada uma avaliação holística¹⁵ e a orientação seja de que a grade é um guia e não um quadro a ser preenchido com pontuações, há uma separação (SCHOFFEN, 2009). Essa separação, para este trabalho, forneceu pistas para uma tentativa de formalização que permitisse aferir a proficiência dos examinandos em um processo semi-automatizado.

Conforme já mencionado no Capítulo 1, os itens que se encontram na parte inferior da grade de correção são itens que podem ser interpretados como mais estáveis do que os itens que pertencem ao eixo de **Adequação Contextual**. Por “mais estáveis” quero dizer itens que não são alterados entre as diferentes tarefas e edições do exame Celpe-Bras. Esses itens, pertencentes aos eixos de **Clareza e Coesão** e de **Adequação Lexical e Gramatical**, permanecem inalterados nas grades de correção¹⁶.

Essa imutabilidade dos descritores nos eixos **Clareza e Coesão** e **Adequação Lexical e Gramatical** pode ser verificada em uma comparação entre as grades de correção das diferentes tarefas e edições do exame, como fizemos, a título de exemplo, entre as grades de correção dos exames aplicados em 2006-1 e 2012-2 e 2013-1. Se são itens que permanecem inalterados, é possível entender que esses seriam os itens mais capazes de fornecer alguma pista com relação às diferenças mais genéricas possíveis de uso da linguagem (dos recursos linguísticos) entre textos mais ou menos proficientes. Considerando que é possível generalizar esses dois itens, para a realização de um experimento, poder-se-ia abstrair o comando para as diferentes tarefas comunicativas aplicadas no exame de 2006-1, deixando o eixo de **Adequação Contextual** de lado por um momento, com a finalidade de comparar as produções textuais apenas utilizando os critérios **Clareza e Coesão** e **Adequação Lexical e Gramatical**.

¹⁵ “A avaliação holística consiste na atribuição de uma única nota para um texto baseada na impressão geral desse texto. Em uma sessão de avaliação holística típica, cada texto é lido rapidamente e então julgado através de uma escala que apresenta os critérios de avaliação.” (WEIGLE, 2002, p. 112)

¹⁶ Ver exemplos em Anexos.

Entendo que deixar de lado o eixo que comanda, no contexto do exame Celpe-Bras, os demais eixos da grade de correção pode ser entendido como algo incoerente ou prejudicial para o processo de avaliação. Em se tratando de *uso adequado da língua para praticar ações*, o essencial para a avaliação da produção textual oral ou escrita é, no caso desse exame, justamente o aspecto comunicativo, isto é, a adequação ao contexto, que é representada pelo eixo **Adequação Contextual**. Isso quer dizer que, mesmo que os textos apresentem coesão e adequação linguística, a produção é sempre julgada como inadequada se não cumprir o que foi solicitado na tarefa (BRASIL, 2013, p. 6).

Porém, para os fins deste trabalho, o que estava procurando descobrir era se esses dois eixos, poderiam, sozinhos, dizer algo a respeito da classificação dos textos em níveis diferentes de proficiência. Essa avaliação de textos mais e menos avançados já estava dada pela Comissão Técnica do exame – lembrando que o *corpus* de estudo foi previamente avaliado por humanos. Além disso, procurava poder dizer até que ponto eles funcionariam isoladamente nesse tipo de instrumento avaliativo e até, talvez, justificar sua presença ou sua retirada da grade de correção.

Para responder a essas perguntas, busquei na literatura não somente pesquisas que tivessem uma abordagem que corroborasse essas ideias, mas pressupostos teóricos que permitissem a reflexão sobre a importância de um eixo chamado **Clareza e Coesão** na determinação da qualidade e da proficiência de um texto. Para isso, a Linguística Textual ofereceu uma discussão interessante daquilo que compõe um texto, especialmente sobre os componentes do que se conhece por *tessitura textual*. Considerando que diversos elementos da tessitura do texto desempenham diferentes funções na sua organização e que esses diferentes elementos podem, inclusive, sinalizar diferentes graus de competência do redator, passamos, a seguir, a uma apresentação de alguns fundamentos da concepção de texto que atualmente são prestigiados no âmbito dos estudos da linguagem, especialmente em Linguística Textual.

3.2 LINGUÍSTICA TEXTUAL

A Linguística Textual ou Linguística do Texto (LT) estuda o texto como produto ou processo. É Koch (2001) quem diz que a LT, atualmente, é um subdomínio linguístico, mostrando que cada vez mais se tornará um fio transdisciplinar que passa por texto e discurso. O texto, nesse sentido, é entendido como um conjunto de *condições* que conduz um leitor e

um autor à produção de um evento comunicativo. Enquanto *conjunto de condições*, o texto é um todo de significação e não apenas uma soma de sentenças que estão alinhadas: o texto é uma unidade de comunicação que tem um fim em si mesmo.

Portanto, neste trabalho, em que textos oriundos do exame Celpe-Bras são observados, *textos* são entendidos como produtos de tarefas ao mesmo tempo em que são produtos de um processo avaliativo. Por serem textos que foram submetidos a um exame de proficiência, há dimensões diferentes de leitura que se interpelam: a dimensão do interlocutor enquanto público-alvo da tarefa (a quem o texto se destinaria) e a dimensão do interlocutor enquanto avaliador (a quem o texto de fato se destina). As tarefas solicitadas no exame são elaboradas de modo a mimetizar tarefas realizadas por nós cotidianamente quando precisamos escrever, portanto, há gêneros (e-mail, carta, texto para afixar em um mural) dentro de outro gênero (teste de proficiência).

O enfoque dado nesta pesquisa é para questões referentes à observação do texto como um produto, e não do processo percorrido por esse produto. Em outras palavras, o cerne do trabalho recai sobre elementos tais como a coesão, mais superficiais, e alguns elementos mais complexos, como coerência, organização de sentenças e inteligibilidade dos textos. O julgamento do que é mais adequado já foi efetuado pela equipe de avaliação do exame. Considerando uma proposta para automatização de alguns passos da avaliação, a centralização em elementos mais superficiais ou mais formalizáveis torna-se uma solução em detrimento de outras possíveis.

Quando se fala em proficiência escrita e texto, Widdowson (1991) afirma que ser proficiente é utilizar orações de forma que se atinja o efeito comunicativo desejado. Para isso, é necessário que as frases sejam organizadas de modo a conduzir o leitor do texto a compreender o que motivou sua escrita e a responder de alguma forma a esse texto produzido. Os produtos desses eventos comunicativos, conforme apontam Beaugrande e Dressler (1981), podem ser diferentes gêneros – que podem partir de um e-mail informal a um amigo próximo e chegar à redação de um contrato jurídico entre instituições privadas. O conhecimento que possuímos da nossa língua nos permite distinguir o que é uma poesia, uma narração, uma bula de remédio. Ainda de acordo com os autores, a ciência do texto deveria estar apta a descrever tanto as características que tornam os textos semelhantes entre si, quanto as características que os diferem e que tornam possível a existência de diferentes gêneros textuais, como, por exemplo, uma placa de trânsito, uma notícia de jornal, um texto didático, um bilhete, uma receita de bolo. É preciso entender, segundo os autores, quais os padrões exigidos por determinados textos, além de saber que pessoas os lerão e com quais objetivos.

No sistema de avaliação do Celpe-Bras, como vimos, o desempenho linguístico – no caso, a proficiência dos examinandos – é mensurado a partir das *habilidades* linguísticas que são concretizadas por *recursos linguísticos*. No contexto de aplicação e de correção do exame, entende-se que o recurso linguístico serve como um instrumento que sustenta a habilidade sendo testada; entende-se, também, que no momento em que esses recursos são utilizados de maneira inadequada, a leitura do texto fica comprometida, podendo causar a incompletude do objetivo comunicativo de escrita.

Dito isso, esses elementos que não são observados analiticamente na avaliação desempenham um papel crucial na construção textual. Sem eles, não haveria materialidade linguística para aferir a proficiência dos examinandos, nem para constituir um texto ou estabelecer a interlocução. Por isso ainda estão presentes na grade de correção. No entanto, os recursos linguísticos, como já dito, são itens separados dos itens que avaliam a adequação contextual (de gênero) e são tratados como secundários ou de menor relevância. Na verdade, ao retomar o que dizem Beaugrande e Dressler (1981), o *status* dos recursos linguísticos pode ser entendido como o mesmo de todos os outros eixos, porque revelam padrões que correspondem às características particulares de cada gênero textual e os gêneros textuais comandarão a seleção dos recursos linguísticos mais adequados.

Por essa razão, entendo que, uma vez que sistematizarmos os diferentes usos desses recursos em uma perspectiva macrotextual (a perspectiva de *corpus*), os mesmos podem justamente servir como base para a automatização de algumas etapas da avaliação, ou podem revelar que de fato não deveriam estar separados nas grades de correção. Isso porque, de uma forma ou de outra, eles vão sempre fazer parte da avaliação de adequação contextual, o contexto de uso da linguagem não existe sem o contexto de uso dos recursos linguísticos.

Numa tentativa de sistematizar os diferentes usos dos recursos linguísticos, utilizamos textos que foram previamente considerados adequados ou inadequados adotando os pontos de vista dados nas grades de correção. Para buscar essa automatização, a LT aponta alguns princípios que garantem a textualidade de uma produção textual. Esses princípios deram origem a diversos tipos de análise e inclusive geraram ferramentas em PLN capazes de mensurar a inteligibilidade de um texto e de pontuar os diferentes recursos linguísticos que dele fazem parte. Na seção seguinte serão apresentados alguns dos princípios considerados relevantes e serão apontados os modos de contribuição desses elementos para o processo de semi-automatização da correção do exame.

3.2.1 PRINCÍPIOS DE TEXTUALIDADE

Considerado um evento comunicativo, o texto deve, idealmente, possuir princípios que lhe permitam estabelecer a comunicação. Beaugrande e Dressler (1981) propuseram, nos anos 80, sete princípios de textualidade presentes nesse evento comunicativo: coesão e coerência, intencionalidade, aceitabilidade, situacionalidade, informatividade e intertextualidade. Esses princípios estão separados em dois conjuntos: a coesão e a coerência fazem parte do conjunto de fatores semântico/formais; o restante dos princípios pertence ao conjunto de fatores pragmáticos. Neste trabalho, de todos esses princípios, o eixo de fatores semântico/formais, em que se encontram a coesão e a coerência, são os princípios de textualidade que interessam a este trabalho. Justamente por serem considerados fatores semânticos e formais, logo, mais formalizáveis, esses princípios são capazes de fornecer pistas para a semi-automatização do processo de correção.

Beaugrande e Dressler (1981) também apontam que há alguns princípios que regem a comunicação textual. São eles: a eficiência de um texto, que depende do seu uso em uma situação de comunicação; a efetividade de um texto, que depende de causar uma boa impressão e atingir seu objetivo; e a apropriação, que depende da união entre o contexto e as formas pelas quais a textualidade é mantida. O objetivo dos autores ao elencar todas essas características não era o de definir as unidades e os padrões de determinados textos, mas, sim, tentar mostrar as operações que governam e desencadeiam padrões de uso da linguagem, sendo o texto um resultado concreto dessas operações. Assim, propuseram que o texto fosse visto como um sistema, em que um conjunto de elementos funciona em função de outros e que esses elementos dependem, portanto, uns dos outros para a formação de um todo de sentido.

Além disso, existem características que permitem ou impedem a constituição de sentenças separadas em um texto. Essas características são definidas pelas relações coesivas, que são a referência, a substituição, a elipse, a conjunção e a coesão lexical. A coesão textual está presente quando a interpretação de um elemento dentro do texto depende de outro elemento, assim, ela funciona como a “costura” do texto. É a coesão que permite que uma ideia esteja ligada à próxima ou à anterior sem causar estranhamento no ato da leitura, uma vez que “[...] em um texto, com exceção da primeira sentença, toda sentença apresenta uma ligação com a anterior e com a próxima” (HALLIDAY e HASAN, 1989).

Dessa forma, Halliday e Hasan deixam muito claro que um texto é, também, a sua textura. É essa textura que permite que um texto seja um todo de sentido, e não um conjunto

de frases desconexas. Por exemplo, nas frases “Deixei meu casaco no carro. Você poderia pegá-lo, por favor?”, a *coesão* é o resultado de haver dois elementos com um só referente (“casaco” e o pronome “lo”) que fazem referência um ao outro, enquanto a *textura* é o resultado da relação coesiva entre esses dois elementos. A coesão e a textura permitem que a leitura de um texto flua, tornando-o tanto coeso como coerente. Um texto que não faz uso das relações coesivas corre o risco de poder parecer um aglomerado de frases desconexas, que farão pouco sentido durante a leitura.

Van Dijk (1984) também trouxe contribuições interessantes nesse sentido. Ele reestabelece as noções de macro e microestrutura textual, apontando que a macroestrutura está relacionada ao nível global do texto e a microestrutura diz respeito à organização interna desse texto. Ainda segundo o autor, as expressões utilizadas, a superfície sintática, a estrutura lexical das sentenças, todos esses fatores juntos, apontam a coerência de um texto. Exemplos desses fatores seriam a ordem das palavras, a ordem das sentenças, o uso de conectivos, o tempo verbal utilizado, os pronomes, entre outros. Existem, portanto, características identificáveis e estruturas mapeáveis que também permitem que um texto seja considerado um todo de sentido.

Sobre as expressões utilizadas, a superfície sintática e a estrutura lexical das sentenças, sabe-se que são parte de um sistema de linguagem. No entanto, esses elementos têm um potencial de significado que só se realiza no texto. O texto, como afirma Baker (1993), tem características de organização específicas da linguagem e da cultura que o distinguem de um não-texto, que seria uma coleção aleatória de sentenças e parágrafos. A autora também menciona que cada comunidade linguística tem preferências para organizar esses elementos e dar origem a diferentes textos. As conexões podem ser estabelecidas no arranjo de informações em cada frase e a forma como elas relacionam as informações com frases anteriores e posteriores; na superfície, em que estabelecem inter-relações entre pessoas e eventos; nas conexões semânticas subjacentes, que nos permitem dar sentido a um texto, caracterizando-o como uma unidade de significado.

Dito isso, entendo que os princípios de textualidade são de extrema relevância para esta pesquisa. Aqui, pretende-se encontrar pontos formalizáveis na correção do exame Celpe-Bras, a fim de alimentar um sistema semi-automatizado de correção. Além disso, acredito ser interessante estreitar os conceitos de coesão e coerência. Esses dois construtos, assim como as palavras, representam a materialidade de constituintes oracionais e ideias em um determinado texto. A coesão e a coerência estão conectadas a níveis específicos da linguagem, do discurso e do conhecimento de mundo. A coesão e coerência são descritores importantes dentro do

processo de avaliação da proficiência dos examinandos do exame Celpe-Bras, e são aferidos pelo eixo **Clareza e Coesão**, em que se observa o desenvolvimento do texto e verifica-se o uso de articuladores, a concordância e as referências.

3.2.2 COESÃO E COERÊNCIA

Coesão e coerência são descritores importantes dentro do processo de avaliação da proficiência dos examinandos do exame Celpe-Bras. Considerando que os examinandos devem produzir textos dotados de coesão, e que a coerência relaciona-se tanto com o sentido do texto quanto com os enunciados das tarefas, esses elementos estão diretamente relacionados à situação de comunicação que envolve a produção escrita do texto. O Manual do Examinando apresenta o que a Comissão Técnica do exame entende por coesão e coerência e de que forma esses itens estão sendo avaliados no contexto do exame:

Coesão e coerência são vistos como conceitos relacionados e complementares. A coerência textual é um processo de construção de sentidos que se estabelece na interação texto-usuário. Trata-se da possibilidade de se estabelecer no texto alguma forma de unidade, relação e continuidade de sentidos. Colaboram para a construção da coerência aspectos como: a manutenção de um tópico por meio de retomadas de conceitos e ideias; a progressão do texto, ou seja, a organização da estrutura informacional para guiar o leitor em sua compreensão; a articulação do texto, ou seja, as relações lógicas que se estabelecem entre fatos, ações ou eventos e conceitos no universo textual; a não-contradição, ou seja, a compatibilidade entre ideias e conceitos no mundo textual e o mundo real a que se referem. A coesão textual caracteriza-se pela presença de elementos linguísticos na estrutura de superfície do texto, que sinalizam conexões sintáticas e semânticas entre as sentenças e permitem a integração destas com o todo. Entre os mecanismos de coesão estão, por exemplo, paráfrase, operadores de junção (sinais que explicitam as relações entre eventos no texto), tempo e aspecto verbal e elipse. Inadequações no uso desses elementos de coerência e coesão, seja pela imprecisão ou ambiguidade, causando quebras tanto na continuidade quanto na progressão, podem comprometer a estruturação do texto e assim dificultar sua compreensão. Em lugar de uma aferição quantitativa de pontos isolados da língua, faz-se uma avaliação qualitativa do desempenho dentro do objetivo da tarefa. Muitas vezes, uma produção textual com pouca ou nenhuma inadequação linguística não necessariamente demonstra compreensão do propósito da tarefa. Proficiência implica efetivamente agir mediante o uso da linguagem. (BRASIL, 2013, p. 7-8).

Acredito que, de certa forma, por estarem separadas na grade de correção, a coesão e a coerência são, mesmo que isso não seja instruído aos corretores, quantificadas. As inadequações de uso dos recursos linguísticos no contexto do exame Celpe-Bras são traduzidas, também, em problemas de coesão e estrutura textual. Essas inadequações estão apontadas explicitamente no eixo da grade de correção **Clareza e Coesão**, em que se explicita a observação e avaliação do uso de articuladores (conjunção, advérbio), da concordância

verbal e nominal e da referência (pronomes). A falta desses fatores ou sua presença em demasia, no entendimento do exame, podem contribuir para a leitura/avaliação dos textos ou podem prejudicá-la, tornando a interlocução mais ou menos consistente. Sua ausência exigiria que o leitor contribuísse com inferências a fim de preencher as lacunas deixadas no texto. Portanto, esses elementos, apesar de parecerem ter um papel menos importante, são essenciais para a concretização e cumprimento das tarefas propostas.

No caso da coesão, as conexões ocorrem também através do uso de elementos linguísticos explícitos (por exemplo, palavras, sinais, pistas, constituintes) e suas combinações. Como no caso de todos os sistemas simbólicos e semióticos, esses elementos são interpretados dentro de um contexto sociocultural específico (a linguagem utilizada e a comunidade a que se destina) e, portanto não devem ser avaliados de forma separada das práticas interacionais.

Já a coerência, no entanto, é algo um pouco mais complexo de ser mensurado. A coerência é o resultado de uma interação entre a coesão textual e o leitor. O nível da coesão é capaz de levar a uma representação mental coerente por parte do leitor, mas pode resultar em uma representação mental incoerente por parte de outro leitor (MCNAMARA et al., 2002). Essas conexões entre as representações mentais dos leitores e os textos são construídas com base em elementos disponíveis no próprio texto, combinados às suas intenções e capacidades cognitivas. Essas conexões são chamadas de coerência (GRAESSER et al., 2004). Dessa forma, tem-se que a construção da coesão textual está relacionada à maneira como os elementos linguísticos do texto estão interligados entre si, de forma a produzir sentido e fazer o texto progredir. A coerência textual, por sua vez, está diretamente ligada à possibilidade de estabelecer um sentido para o texto com relação ao usuário. Por envolver o conhecimento de mundo e as intenções de leitura do usuário, a coerência está ligada à interpretabilidade do leitor (KOCH e TRAVAGLIA, 2009).

De acordo com Halliday e Hasan (1989), a coesão é um conjunto de itens linguísticos através dos quais os textos são constituídos e interconectados, possibilitando o estabelecimento do sentido por meio de uma sequenciação. A literatura distingue diversos tipos de coesão, e mais usual é a distinção entre local e global. Tanto a coesão como a coerência de um texto estruturam-se local e globalmente. O leitor encontra as relações de coesão local entre as sentenças adjacentes no texto, e a coesão global entre grupos de sentenças e grupos de parágrafos. A distinção é importante porque, tanto a coesão global quanto a local fornecem pistas para os leitores organizarem e processarem a leitura. As sentenças presentes nos parágrafos são marcas da coesão global. As anáforas, a repetição de termos e as relações

intersentenças são marcas da coesão local. Os textos que são localmente coesivos mas apresentam problemas de coesão global oferecem obstáculos à compreensão e à recuperação de informações ao leitor (MCNAMARA et al., 1996). Da mesma forma, os textos que são globalmente coesivos, mas apresentam problemas de coesão local, algumas vezes, impõem dificuldade à leitura e à compreensão (VAN DIJK, 1984; MCNAMARA et al., 2002).

A coesão e coerência são divididas, muitas vezes, em categorias que costumam responder quem, quando, onde, por que e como os eventos descritos em um texto estão facilitando o processamento de leitura. A coesão referencial, por exemplo, estabelece-se através do uso de anáforas, pela retomada de frases, pelo uso de artigos definidos e pela repetição de termos. A coesão pode também referir-se à noção de tempo, estabelecendo-se por conectivos (como *antes*, *depois* e *então*) ou frases preposicionais (*mais tarde*, *naquele dia*), pelo tempo e aspecto verbal (*crescera*, *comia*) ou pelo uso de marcadores de ordenação (*primeiro*, *segundo*, *décimo*).

Os marcadores de coesão são muitos e os *parsers* sintáticos já são capazes de identificar uma grande gama de informações, como a presença e a adequação de títulos e cabeçalhos, a consistência do assunto entre as frases de um texto, as categorias sintáticas e suas estruturas hierárquicas e a concordância verbal. O vocabulário pode ser identificado através de elementos gramaticais e lexicais, assim como através da relação semântica entre as palavras. Os mecanismos são infundáveis, mas esses exemplos sugerem onde fórmulas de inteligibilidade podem ser aplicadas com ferramentas de PLN.

Neste trabalho, os textos que estamos manipulando já foram avaliados de forma integrada, ou seja, foram julgados por avaliadores humanos que levaram em consideração o uso dos elementos coesivos na medida em que permitem a configuração da interlocução com o objetivo de cumprir determinado propósito interlocutivo de forma mais ou menos preferível em determinado contexto. De acordo com Schoffen (2009):

É somente dentro do contexto que se pode saber quais recursos são preferidos [historicamente], na medida em que eles são definidos situadamente pelos interlocutores. Nesse sentido, entendemos que a preferência por determinados recursos linguísticos utilizados pode se tornar mais (ou menos) flexível em diferentes tarefas. (p. 121)

Fazendo um adendo ao que diz a autora, entendo que a preferência por determinados recursos linguísticos nos textos já avaliados pode apontar diferentes tendências entre falantes mais e menos experientes da língua portuguesa (examinandos que obtiveram níveis mais avançados de proficiência no exame e examinandos que obtiveram níveis mais básicos). Indica, também, a relevância da presença do eixo **Clareza e Coesão** na grade de correção.

Retomando a ideia de que a coesão é marcada, também, pelo uso de elementos linguísticos explícitos, ela pode ser, portanto, extraída por programas computacionais. A coerência, no entanto, resulta de uma interação entre a coesão textual e um leitor. Um nível específico de coesão pode vir a levar a uma representação mental coerente para um leitor, mas isso pode não acontecer com outro (MCNAMARA et al., 1996). Essas conexões, na representação mental do leitor, são construídas com base em elementos presentes no texto combinados às capacidades cognitivas do leitor e suas intenções; a isso se dá o nome de coerência (GRAESSER et. al., 2004).

Há algumas décadas, não seria possível investigar sistematicamente a coesão e a coerência por causa da impossibilidade que havia de manipular medidas de conhecimento de mundo e por causa da falta de medidas de proficiência de linguagem em múltiplos níveis. Hoje, no entanto, graças aos avanços de diversas áreas dos estudos linguísticos e computacionais – especialmente da Psicolinguística, nos anos 90, do Processamento do Discurso, da Linguística de Corpus (BIBER, CONRAD, REPPEN, 1998) e da Linguística Computacional –, chegou-se ao desenvolvimento de ferramentas sofisticadas capazes de analisar textos em suas diversas dimensões. Existem, atualmente, programas com grandes léxicos, *parsers* sintáticos, analisadores semânticos, avaliadores de textos e *softwares* de apoio à escrita, como elaboradores automáticos de resenhas ou um dos projetos brasileiros do NILC (Núcleo Interinstitucional de Linguística Computacional), o Sci-Po-Farmácia, que auxilia pesquisadores a redigir textos acadêmicos de vários gêneros¹⁷. Há ainda outras ferramentas que também analisam textos em diversas dimensões, são elas alinhadores de sentenças, sumarizadores automáticos, tradutores automáticos, analisadores discursivos (que utilizam a RST, como o DiZer 2.0, disponível em <<http://www.nilc.icmc.usp.br/dizer2/>>), analisadores lexicométricos em *corpus* (como o Léxico 3, disponível em <www.tal.univ-paris3.fr/lexico/>) ou mesmo aquelas advindas do Projeto *Open Mind Common Sense* <http://www.sensocomum.ufscar.br:8080/omcs/login_pt_BR.jsp>.

Para Graesser et al. (2004), as fórmulas de inteligibilidade (explicadas a seguir) e de avaliação de complexidade ignoram componentes linguísticos e discursivos que influenciam na dificuldade de compreensão textual. Os autores apontam para o fato de que, apesar de os parâmetros de tamanho das sentenças e das palavras terem alguma validade, tais parâmetros não revelam, por si só, a complexidade de um texto. Assim, propõem uma análise da coesão e da coerência textual em múltiplos níveis. De acordo com Graesser et. al. (2004), coesão

¹⁷ Disponível em <<http://www.nilc.icmc.usp.br/scipo-farmacia/>>. Acesso em 03 mai 2012.

textual é uma propriedade objetiva do texto, e coerência é a representação mental do conteúdo do texto feita pelo leitor através das palavras, sentenças e frases que orientam a leitura e conectam as ideias umas às outras. O desafio, segundo os autores, é automatizar esses níveis mais profundos de análise textual. Essa foi a motivação da criação da ferramenta Coh-Metrix, descrita nas seções 5.2 e 5.3 desta dissertação.

3.3 MEDIDAS DE INTELIGIBILIDADE

No âmbito dos projetos que visam a automatização da avaliação textual, além de ferramentas capazes de mensurar a coesão e a coerência, existe uma série de medidas anteriores já utilizadas há bastante tempo, chamadas de *medidas de inteligibilidade*. Essas medidas se baseiam em fatores como número de palavras em sentenças e número de letras ou sílabas por palavra para dar um escore a um texto. Duas das medidas mais utilizadas são as fórmulas *Flesch Reading Ease* e *Flesch-Kincaid Grade Level*. O resultado da primeira fórmula é um número de 0 a 100, sendo que o escore mais alto indica a facilidade de leitura do texto. A conversão mais comum do resultado dessa fórmula é transformar os números de 0 a 100 em níveis ou séries escolares.

As fórmulas de inteligibilidade, ao menos nos Estados Unidos, guiam a produção de livros didáticos de modo que a escrita presente no livro seja adequada ao nível do público-alvo daquela obra. Existem, no entanto, alguns problemas nessas fórmulas, que acabam impossibilitando a previsão da dificuldade de compreensão dos textos ou sua avaliação em níveis de proficiência. Um desses problemas é que os escores baseiam-se em características da superfície do texto. Mensurar a compreensão depende, em grande parte, do processamento desses textos e das situações nas quais se encontram (MCNAMARA et al., 2002).

Avanços recentes em processamento do discurso e na Linguística Computacional permitiram a produção de medidas mais sofisticadas de inteligibilidade. Prever o nível de leitura, de compreensão que se terá do texto e do aprendizado ao qual o texto levará requer considerar o conhecimento do leitor, suas habilidades e outras questões cognitivas. Embora as características dos textos sejam capazes de prever com certo grau de acurácia a inteligibilidade de um texto, essa medida ainda precisa ser vista como uma interação entre um texto e as aptidões cognitivas do leitor. Ainda, as fórmulas de inteligibilidade não são capazes de capturar a coesão e a coerência dos textos, um item fundamental como mostraram

pesquisas envolvendo a avaliação das dificuldades de ler um texto mais coeso realizada com leitores (GRAESSER, GERNSBACHER e GOLDMAN, 2008).

Nessas pesquisas, os escores de inteligibilidade foram maiores para textos coesos do que para textos menos coesos. A coesão, nesses casos, é medida na presença de elementos explicitamente coesivos, como os conectivos, as conjunções e os pronomes. Porém, nem sempre o aumento da presença desses itens equivale a dizer que o texto é mais coeso e mais fácil de entender. Algumas passagens que possuem fortes efeitos coesivos sem a presença de marcadores de coesão existem em abundância no uso da linguagem, um exemplo disso é a frase “A chuva cai”. Um texto com muitos elementos de coesão sobre a Mitose (MCNAMARA et. al., 2002) resultou em melhor compreensão por parte dos leitores e teve um *Flesch-Kincaid* de 11,2, comparados a 9,3 de uma versão com poucos elementos de coesão. Muitos exemplos estão disponíveis, mas o aumento da presença de elementos coesivos muitas vezes acarreta no aumento do número de palavras de um texto, e sentenças mais longas resultam em números mais altos nas medidas de inteligibilidade, que apontam um texto de compreensão mais difícil.

Além do que já foi exposto, as fórmulas de inteligibilidade falham em muitos outros aspectos na avaliação da coesão. Por exemplo, o número menor de pronomes pode aumentar a coesão referencial. Por outro lado, os pronomes são palavras menores, e a extensão das palavras está diretamente ligada ao aumento ou diminuição do escore de inteligibilidade. Ainda, a coesão referencial também aumenta quando a repetição conceitual está mais presente, o que as fórmulas de inteligibilidade não são capazes de capturar.

Para Graesser et al. (2004), apesar de os parâmetros de tamanho das sentenças e das palavras terem alguma validade, tais parâmetros não revelam, por si só, a complexidade de um texto. Assim, propõem uma análise da coesão e da coerência textual em múltiplos níveis. De acordo com os autores, coesão textual é uma propriedade objetiva do texto, e coerência é a representação mental do conteúdo do texto feita pelo leitor através das palavras, sentenças e frases que orientam a leitura e conectam as ideias umas às outras. O desafio, segundo os autores, é automatizar esses níveis mais profundos de análise textual. Essa foi a motivação da criação da ferramenta Coh-Metrix.

Graesser et al. (2004) apontam para interações “intrigantes” entre a tessitura coesiva de um texto e o conhecimento de mundo do leitor ao construir e usar modelos mentais subjacentes a, por exemplo, textos científicos. Leitores com menos conhecimento prévio a respeito da área em questão beneficiam-se de textos com maior coesão, ao passo que leitores conhecedores do assunto tratado no texto beneficiam-se mais de textos menos coesos. Uma

menor coesão textual permite que o leitor que domina o assunto faça inferências e, conseqüentemente, estabeleça mais conexões entre as ideias do texto e o seu conhecimento sobre o assunto. Esse processo resulta em uma representação mental mais coerente e sugere que nem sempre um texto com coesão homogênea é o texto ideal para todos os tipos de leitores.

4 LINGUÍSTICA DE CORPUS

Para Halliday, Sinclair e Gries (1989, 1991 e 2006, respectivamente), expoentes da LC, a única forma segura para se descrever uma língua e seus usos é através da observação dessa língua em uso, por meio de registros autênticos. Complementando essa observação, Hoey (1991) bem salienta que a LC já não é mais apenas uma ramificação da linguística, mas sim um caminho para encontrar a linguística. Dessa forma, a LC é um ponto de partida neste trabalho, mas não necessariamente será seu ponto de chegada.

O objeto de estudo desta pesquisa é um *corpus*, composto por textos submetidos a um exame que avalia a proficiência de estudantes de português como língua adicional, o ponto de partida. Esse exame faz uso de tarefas comunicativas as quais tomam por base gêneros do discurso que são frequentemente acionados em situações reais de comunicação, como por exemplo ler uma reportagem de revista e escrever um e-mail sobre ela a um amigo que possa estar interessado no assunto; assistir a um programa sobre saúde no trabalho e escrever, enquanto funcionário de uma empresa, solicitações de melhoria no ambiente de trabalho remetendo ao que assistiu no programa; ler uma reportagem sobre turismo no Brasil e escrever um texto para o seu blog incentivando pessoas a conhecerem o Brasil. O exame avalia o desempenho dos examinados ao usarem a língua nas tarefas solicitadas, e esse uso está conceituado no Manual do Examinando (BRASIL, 2013, p. 7) da seguinte forma:

‘uso adequado da língua para desempenhar ações no mundo’. Esse conceito revela uma visão de linguagem como ‘ação conjunta de participantes com um propósito social’, que para ser implementada tem de levar em conta ‘o contexto, o propósito e o(s) interlocutor(es) envolvido(s) na interação com o texto.’

Entende-se, portanto, que o uso da língua, no contexto desse exame, é a forma como o examinado aciona seu conhecimento linguístico em diferentes tarefas propostas pelo instrumento avaliativo, onde

[...] ler, por exemplo, significa mais do que compreender as palavras do texto. Uma leitura proficiente e crítica envolve atribuir sentidos autorizados pelo texto, selecionar informações relevantes, relacioná-las e usá-las para propósitos específicos solicitados pela tarefa do Exame. [...] proficiência na escrita significa usar a informação relevante e adequar a linguagem ao propósito da escrita (reclamar, opinar, argumentar etc.) e ao interlocutor (amigo, chefe, leitores de um jornal etc.),

levando-se em conta os parâmetros de textualização de diferentes gêneros discursivos (mensagem eletrônica, cartas do leitor, texto publicitário etc.).

Nesse sentido, a LC é capaz de fornecer pistas para o esclarecimento do que pode ser considerado um “uso da língua” mais ou menos adequado. Por ser uma área dos estudos linguísticos que analisa os *padrões* de uso real da língua em grandes conjuntos de textos autênticos, faz muito sentido trazê-la para análise dos textos dos examinandos no contexto do exame Celpe-Bras. Essa teoria/metodologia de estudo faz isso através de observação empírica com a finalidade de apontar quais formas gramaticais e formas de uso de palavras e estruturas possíveis e prováveis são acionadas com maior frequência pelos falantes em condições reais de produção.

Apesar de muitas vezes poder ser confundida com uma forma racionalista de abordar a língua, a LC traz, na verdade, uma visão de língua como *probabilidade*, e não como *possibilidade*, contrapondo-se aos estudos formalistas chomskinianos. Enquanto a LC tem Halliday, de tradição empirista, ao seu lado, Chomsky está do lado oposto, como representante do racionalismo na Linguística. Por sua característica empírica, essa área dos estudos linguísticos já se fortaleceu no universo acadêmico brasileiro e possui um longo histórico de pesquisas sobre e na língua inglesa, especialmente voltada ao ensino de línguas e tradução. Biber (1998), um de seus principais propagadores, ressalta que linguistas de *corpus* “estudam a língua realmente utilizada em textos naturais” (BIBER, 1998, p. 1), fortalecendo o contraponto entre linguistas de *corpus* e linguistas de poltrona, metáfora provocativa colocada por Fillmore em 1991. Em virtude desse forte contraponto, infelizmente, o que se vê hoje é uma espécie de guerra entre campos de pesquisa que deveriam ser, por que não, complementares. Sobre isso, vale retomar o texto de Fillmore na íntegra, que, através de uma espécie de anedota, relata como linguistas dos dois lados se enxergam:

A Linguística de Poltrona não é bem vista em alguns círculos linguísticos. A caricatura que se faz do linguista de poltrona é mais ou menos esta: ele senta em uma poltrona extremamente confortável, com os olhos fechados e as mãos cruzadas atrás da cabeça; de vez em quando, ele abre os olhos e senta, gritando “Nossa, que interessante!”, pega um lápis e escreve alguma coisa. Depois, fica andando para lá e para cá animado por ter chegado um pouco mais perto de descobrir o que é a língua. (Não há ninguém que seja exatamente assim, mas existem pessoas parecidas.)

A Linguística de Corpus não é bem vista em alguns círculos linguísticos. A caricatura que se faz do linguista de corpus é mais ou menos esta: ele tem todos os fatos primários de que precisa, em forma de um *corpus* de aproximadamente um zilhão de palavras correntes, e seu trabalho é descrever fatos secundários a partir dos fatos primários. No momento ele está ocupado determinando as frequências relativas de onze categorias gramaticais da primeira palavra de uma frase *versus* as da segunda palavra de uma frase. (Não há ninguém que seja exatamente assim, mas existem pessoas parecidas.)

Esses dois linguistas não se falam muito, mas quando conversam, o linguista de *corpus* pergunta para o linguista de poltrona “Por que eu deveria acreditar no que você me diz?, enquanto o linguista de poltrona pergunta ao linguista de *corpus* “Por que eu deveria achar o que você faz interessante?”¹⁸

Hoje, a LC tem se popularizado por causa do fácil acesso a *corpora* digitalizados e com as ferramentas gratuitas disponíveis para análise de *corpus*. Parte do ferramental comum aos trabalhos de LC são a *suite* WordSmith Tools¹⁹ – *suite* de aplicativos paga, muito utilizada nos anos 90 e início dos anos 2000 para fazer análise lexical; fornece listas de palavras, de concordâncias, entre outras utilidades –; e, mais recentemente, o AntConc²⁰, *software* livre disponível para todas as plataformas e que possui extensa documentação on-line. Outros exemplos, ainda, de ferramentas disponíveis para análise de *corpus* seriam o Kitconc 3.0 (desenvolvido por Moreira Filho em 2008, ferramenta para Windows apenas), o Unitex²¹ e o Philologic²². O uso do computador para a extração de dados, portanto, já não é algo tão novo, embora ainda existam muitos linguistas que permaneçam resistentes à ideia de usar a ajuda de uma máquina para fazer análise textual e que não acreditem que esse possa ser um ferramental útil aos estudos linguísticos. Muito em função dessa resistência, alguns problemas são frequentemente levantados por linguistas em relação aos estudos baseados em *corpus*. Sinclair (1999) sintetiza os principais argumentos levantados pela academia contra o uso de *corpus* nos estudos linguísticos:

- um *corpus* não é um exemplo acurado da língua;
- por algo estar presente no *corpus*, não quer dizer que esteja “correto”;

¹⁸ Tradução da autora. No original:

Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, “Wow, what a neat fact!”, grabs his pencil, and writes something down. Then he places around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn’t anybody exactly like this, but there are some approximations.)

Corpus linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn’t anybody exactly like this, but there are some approximations.)

These two don’t speak to each other very often, but when they do, the corpus linguist says to the armchair linguist, “Why should I think that what you tell me is true?”, and the armchair linguist says to the corpus linguist, “Why should I think what you tell me is interesting?” (1991, p. 35-36)

¹⁹ Disponível em <<http://www.lexically.net/wordsmith/>>. Acesso em 9 abr 2012.

²⁰ Documentação e *software* disponíveis para *download* em <<http://www.antlab.sci.waseda.ac.jp/software.html>>. Acesso em 9 abr 2012.

²¹ Disponível em <<http://www-igm.univ-mlv.fr/~unitex/>>. Acesso em 28 mai 2013.

²² Disponível em <<https://sites.google.com/site/philologic3/>>. Acesso em 28 mai 2013.

- “frequência de ocorrência” é a mesma coisa que “importância de ocorrência”; e
- mesmo o maior *corpus* disponível possui lacunas e não representa o todo da língua.

Algumas dessas críticas são facilmente rebatidas, visto que demonstram a falta de conhecimento do que de fato ocorre nos estudos com *corpus*. Os estudos com *corpus*, por exemplo, não têm como objetivo mostrar um exemplo de uso perfeito ou um modelo de linguagem somente: eles têm como objetivo mostrar o que ocorre no uso real da língua, em diferentes recortes, mostrando suas diferentes nuances e situações de uso, sejam elas consideradas mais ou menos adequadas, formais ou informais. Aliás, é justamente para isso que se faz pesquisa com *corpus*: para mostrar as diferentes variações de uso da linguagem, para verificar se determinados padrões ou desvios do que se entende por “norma” de fato existem, em que quantidade e em quais contextos.

A LC funciona dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem como sistema probabilístico. Essa visão particular de como funciona a linguagem indica que, em certa medida, ela é padronizada e que, “embora muitos traços linguísticos sejam possíveis teoricamente, eles não ocorrem com a mesma frequência” (BERBER SARDINHA, 2004, p. 350). Isso significa acreditar que a frequência de ocorrência de determinados elementos linguísticos é maior do que a de outros, e que essas frequências estão atreladas a contextos de uso. Tomando conectivos causais como exemplo, a partir dessa pressuposição, é possível dizer, como já foi mostrado (EVERS, ALLE, MARCOLIN, 2008; FINATTO, EVERS, ALLE, 2009), que existe uma probabilidade maior de determinados conectivos causais ocorrerem em maior número em textos didáticos de uma área (por exemplo, Capítulos de Manuais Universitários de Química comparados a artigos científicos de Química). Para fazer esse tipo de constatação, na maioria das vezes impossível de ser observada a olho nu ou impensada de ser contada manualmente e em pequenos excertos de texto, são necessárias grandes quantidades de dados e, geralmente, ferramentas computacionais robustas para apoiar essa observação.

Dito isso, Biber (1998) afirma que a LC, por mostrar uma forma de observação tão diferente, acaba por romper com uma visão tradicional dos estudos linguísticos. Através da observação de textos autênticos em busca de padrões que fazem parte de determinados contextos, a LC vai de encontro àquelas típicas observações que avaliam a gramaticalidade ou que tentam mapear o que há de diferente nos textos daquilo que as gramáticas postulam. Para o autor, é através da LC que somos capazes de deixar de lado essa pré-visão treinada que

temos enquanto linguistas, buscando encontrar aquilo que nos é estranho. É através da observação de grandes quantidades de dados, considerando as diferentes variáveis contextuais, que é possível chegar ao que é uso padrão e recorrente.

Stubbs (1996) ainda reforça outra ideia: a de que a pesquisa linguística baseada em *corpus* deve sempre tratar dados e análise de maneira independente. Com isso, o autor quer dizer que o pesquisador que faz uso de *corpus* está comprometido em dar voz aos dados, mesmo que isso signifique dizer exatamente o contrário do que era esperado como resultado ao se dar início à pesquisa. Por se tratar de um estilo de pesquisa empírico e que toma por base *corpus*, é possível, enquanto pesquisador, deixar que os textos guiem a pesquisa. Isso significa que o *corpus* de estudo não é um lugar para encontrar respostas a perguntas pontuais ou confirmar hipóteses e buscar exemplos para essas confirmações; ele é uma fonte de novas perguntas e de reflexão. Dessa forma, é até desejável que se aborde um *corpus* sem ideias pré-formuladas ou categorias preestabelecidas.

Retomando o que foi dito nesta seção, é possível sistematizar alguns princípios da LC da seguinte forma:

- A Linguística é uma ciência social e aplicada;
- A linguagem precisa ser estudada em instâncias de uso autênticas, não através de sentenças inventadas;
- Forma e significado são inseparáveis;
- Léxico e gramática são interdependentes;
- A Linguística de Corpus é empírica, observa os padrões de uso em textos autênticos;
- Utiliza uma grande coleção de textos naturais;
- Faz uso extensivo de computador;
- Depende de técnicas de análise quantitativas e qualitativas;
- Tem como objetivo explorar a importância dos resultados quantitativos para aprender sobre os padrões de uso;
- Com ela é possível descrever com maior precisão um uso da língua em extensão, quantitativa e qualitativa.

Seguindo essa linha de raciocínio, é possível afirmar que os contextos de uso de palavras e as estruturas possuem uma certa estabilidade observável, ou seja, existem determinadas combinações que não ocorrem por acaso e que, por isso, estão atreladas a cada evento comunicativo. Além disso, os linguistas de *corpus* não negam o fato de que a linguagem

é inata, e o ser humano tem sua competência linguística específica. No entanto, devido a diferentes objetos e objetivos de pesquisa, diferentes metodologias devem ser utilizadas conforme as perspectivas de investigação. A Linguística de Corpus estuda o que os falantes fazem com a língua (desempenho), e não a capacidade de usar a língua, considerada em abstrato (competência). É com esse olhar, entendendo que a linguagem é um sistema probabilístico, o qual funciona através de correlações entre recursos linguísticos e situações de uso e no qual a variação não é livre de contexto (Biber, 1998), que se está observando o *corpus* de textos submetidos ao exame Celpe-Bras.

4.1 LINGUÍSTICA DE CORPUS E USO DA LINGUAGEM

O que a LC entende por “uso da linguagem” é bastante relevante para o contexto de análise do *corpus* do exame Celpe-Bras. Azeredo (2007) esclarece que “para observar o uso da linguagem é preciso saber em que estrutura o uso está inserido, e para observar a estrutura, é preciso ver em que contextos de uso tal estrutura é usada” (p. 71). Os textos que fazem parte do *corpus* desta pesquisa foram produzidos durante uma testagem, na qual produções textuais foram avaliadas com relação ao desempenho do examinando na habilidade de “usar a linguagem”. Nesse contexto, uso da linguagem significa fazer algo (HYMES, 1972) em língua portuguesa, ou seja, articular os conhecimentos linguísticos para produzir sentenças que tragam o resultado desejado que deu origem ao ato comunicativo.

O uso da linguagem, de acordo com os princípios da LC, é entendido como *ocorrência em diferentes combinações*. Isso equivale a dizer que a maioria das palavras que utilizamos no nosso dia a dia possuem diferentes usos e, portanto, diferentes significados, que dependem de combinações com outras palavras. A visão da LC sobre uso da linguagem é a de que *usos* são “os caminhos sistemáticos em que os aspectos linguísticos são usados em associação com outros aspectos linguísticos e não linguísticos” (BIBER, 1998, p. 5). Por associações linguísticas, Biber entende as associações lexicais e gramaticais; por não linguístico, entende a observação da distribuição desses aspectos linguísticos em diferentes gêneros, dialetos ou períodos.

Para Sinclair (1991), também teórico da LC, há um princípio idiomático que rege a linguagem, através do qual o “usuário [...] tem a seu dispor um grande número de sintagmas semi ou pré-construídos, que constituem em escolhas únicas, mesmo que pareçam ser analisáveis em segmentos” (p. 110). Esse princípio implica que a presença ou ausência de uma palavra em um determinado sintagma depende de palavras que foram selecionadas

anteriormente. Qualquer texto, na opinião do teórico, é o resultado do entrelaçamento desses dois princípios: ora recorremos a unidades compostas por dois ou mais itens já ouvidos/lidos e internalizados ou fazemos escolhas complexas de natureza léxico-gramatical.

Entendo que o que a LC está trazendo sobre o *uso da linguagem* vai ao encontro do que teóricos de Linguística Aplicada e até mesmo os Manuais do exame Celpe-Bras propõem. Se a linguagem for entendida como um sistema probabilístico, isso significa acreditar que probabilidades maiores ou menores de determinados usos ocorrerem existem, estando a seleção desses usos vinculada ao locutor, ao indivíduo, à situação comunicativa, aos exemplos de uso da linguagem. Essas afirmações são muito semelhantes ao que dizem Schlatter, Garcez e Scaramucci (2004, p. 356):

[...] com base no conceito de uso da linguagem como uma ação conjunta dos participantes com um propósito social, o conceito de proficiência linguística/sucesso muda de conhecimento metalinguístico e domínio do sistema para uso adequado da língua para desempenhar ações no mundo.

Assim, por exemplo, não é qualquer verbo que pode preencher a necessidade imposta por um sujeito, mas sim somente um determinado verbo ou grupo de verbos restringidos por ele. Entende-se, dessa forma, que um falante ou usuário da língua precisa, para ser considerado proficiente, não somente conhecer o sistema abstrato dessa língua, mas saber como usá-la adequadamente em determinadas situações. Da mesma forma como não é qualquer verbo ou grupo de verbos que *pode* preencher a necessidade imposta por um sujeito, também não é qualquer escolha de palavra, de estrutura coesiva ou de entradas e fechos de um texto que podem ser acionadas a qualquer evento comunicativo.

4.2 LINGUÍSTICA DE CORPUS E CORPORA DE APRENDIZES

Algumas das questões e hipóteses de pesquisa presentes nesta dissertação tiveram sua origem a partir da leitura de uma série de outros trabalhos realizados com *corpora* de aprendizes. Esses trabalhos são numerosos e descrevem uma série de características do tipo de produção textual aqui exposto (produções textuais submetidas a exames de proficiência) e do tipo de autor aqui tratado (aprendiz/examinando).

De acordo com Berber Sardinha (2004), em estudos realizados utilizando *corpora* de aprendizes e *corpora* de falantes nativos²³ de uma língua, os contrastes sempre ajudam a

²³ Neste parágrafo e em parte desta seção, está-se reproduzindo nomenclatura amplamente consagrada em linguística de *corpus*, que ainda faz uso da dicotomia “falante nativo” e “falante não-nativo” com a finalidade de

traçar um perfil léxico-gramatical da escrita desses dois grupos de falantes. Ao comparar um *corpus* considerado “ideal” com um *corpus* de aprendizes, são fornecidos bons dados sobre as produções dos aprendizes que podem ser utilizados com a finalidade de, talvez, melhorar suas proficiências de língua, a produção de tarefas para a sala de aula e a abordagem do professor com relação a inadequações mais pontuais apresentadas.

Muitos desses trabalhos comparam *lexical bundles* (Biber, 1998) – conhecidos também como pacotes lexicais ou *n-gramas*, estudos que se baseiam na combinação de palavras recorrentes no texto com base em *corpora* específicos. Segundo Biber (idem, p. 990), os “*lexical bundles* são sequências de palavras que co-ocorrem naturalmente no discurso”, ou seja, são formados por expressões recorrentes, independente de sua idiomaticidade ou de sua condição estrutural. Exemplos desses “pacotes lexicais” em inglês seriam: *the end of the, in addition to the* e *the point of view of*. A comparação de pacotes lexicais entre textos produzidos por falantes nativos de inglês e aprendizes já rendeu uma série de discussões sobre o porquê de haver diferenças entre as produções e o que fazer para melhorar o desempenho dos aprendizes com relação a isso.

Contrastes desse tipo também são realizados entre *corpora* de diferentes falantes de uma dada língua. Pesquisadores, tais como a pioneira Granger (1994), encontraram sequências pré-fabricadas – como as sugeridas na descrição do princípio idiomático de Sinclair (1999) – e pacotes lexicais que são utilizados sintática e pragmaticamente de maneira diferente quando a comparação é feita entre *corpora* produzidos por aprendizes de inglês que possuem diferentes primeiras línguas. Em termos de produção teórica sobre *corpora* de textos de aprendizes estão as pesquisas desenvolvidas sobre o inglês como segunda língua (o ICLE – *The International Corpus of Learner English* – de GRANGER, 1994; GRANGER et al., 2002). Exemplos de estudos desse tipo são os de Aijmer (2002), sobre a modalidade, sobre os advérbios intensificadores, sobre a forma causativa *make* e sobre marcadores discursivos. A tônica desses estudos, segundo Granger (2002), é fazer comparações entre a linguagem de “nativos” e “não nativos”, ou entre a “norma” e a “não norma”, para deixar em evidência tudo aquilo que confere estranheza ao texto produzido pelo “não nativo”, incluindo-se aí os “erros”, o uso em excesso ou econômico de palavras, expressões e estruturas.

marcar os diferentes *backgrounds* dos falantes de uma língua. É importante destacar que ao usar “falante nativo” não se está conferindo mais valor a um tipo ou outro de falante, nem se quer dizer que todos os falantes devam atingir um dado padrão de língua. Trago esta aproximação aqui a fim de ponderar e até de problematizar o que está sendo considerado, por exemplo, no contexto do exame Celpe-Bras, nível Iniciante e Avançado na amostra de 2006-1, e se há relação entre estar em um nível Avançado e estar mais próximo da fala de um falante nativo de português brasileiro, muito embora isso não apareça em qualquer manual ou guia do exame, nem nos trabalhos que fazem análise do construto teórico do exame.

No cenário brasileiro, os estudos e também a compilação de *corpora* de aprendizes ainda engatinham. Shepherd (2009) compilou e analisou um *corpus* oral no qual estudantes de graduação verbalizaram apreciação sobre textos literários. Também utilizaram a LC como metodologia para avaliar a atitude de professores de inglês com relação a um componente de curso de especialização recém-cursado por esses mesmos professores. Outro exemplo verificou como aprendizes de língua inglesa usam o modal *can* em textos escritos comparando suas preferências com as preferências de estudantes cuja língua materna é o inglês.

Com relação aos estudos realizados com *corpora* de aprendizes no Brasil, não se pode deixar de mencionar o Br-ICLE²⁴ (Brazilian International Corpus of Learner English), formado, atualmente, por 127 composições argumentativas escritas por universitários brasileiros, aprendizes de língua inglesa em nível avançado, cursando do quinto semestre em diante. Pesquisadores não vinculados ao grupo podem solicitar autorização para uso do *corpus* e realizar suas buscas de modo independente. Cada uma das composições coletadas está identificada com sexo, idade, tempo de contato com a língua inglesa, entre outras informações do aprendiz; o *corpus* possui 65.304 palavras.

No âmbito do PLA, infelizmente, ainda há pouco sobre o assunto. O trabalho de Yuqi (2011) foi o único encontrado que faz uso da metodologia e dos pressupostos teóricos da LC para analisar um *corpus* oral de aprendizes de PLA. Em seu trabalho, a autora compara o uso de *hedges* produzidos por falantes brasileiros de português e por estudantes chineses de português como língua adicional. Os *hedges* são notas cautelosas produzidas para mostrar de que forma uma palavra está sendo empregada, por exemplo, “até onde eu sei” é um *hedge*. Apesar de serem raros, há um esforço aparente nos últimos anos por parte dos linguistas de *corpus* em tentarem mostrar a gama de possibilidades que o trabalho com *corpora* pode oferecer para o planejamento de aulas, para a produção de tarefas e para o *feedback* dado aos estudantes sobre suas produções orais e textuais.

Exemplos disso são ainda outros trabalhos no Brasil que estão sendo atualmente desenvolvidos utilizando *corpora* para o ensino de línguas estrangeiras, conforme os relatados no livro *Corpora no ensino de línguas estrangeiras*, de Viana e Tagnin (2011). Uma compilação de trabalhos que abordam o ensino de alemão, francês, espanhol e inglês para falantes de português, mas que, infelizmente, ainda não traz nenhum relato de pesquisa envolvendo o ensino de português.

²⁴ Mais informações disponíveis em <<http://www2.lael.pucsp.br/corpora/bricle/>>. Acesso em 23 jan 2012.

Os estudos mencionados acima usam a abordagem dirigida pelo *corpus*, isto é, não lançam mão de categorias linguísticas preestabelecidas para confirmação de hipóteses. O estudo segue os preceitos de análise de *corpora* de aprendizes: a análise procura desenvolver meios para descrever as estratégias usadas ou não usadas pelos aprendizes com a finalidade de ajudá-los e de, no futuro, informar a prática pedagógica. Os autores ainda vão mais além, afirmando que um exame cuidadoso de uma lista de agrupamentos lexicais pode inclusive ajudar a entender como os textos de usuários experientes são formados e até que ponto os textos de aprendizes coincidem ou se diferenciam dos textos de usuários mais experientes.

Os *corpora* de aprendizes são usados na Europa já há quase duas décadas, mostrando que textos de falantes nativos e não nativos diferem em termos de frequência de palavras, de frases e de estruturas sintáticas. Os aprendizes frequentemente demonstram não reconhecer gêneros diferentes; aprendizes avançados de uma língua apresentam problemas semelhantes e enfrentam desafios parecidos com os de falantes nativos dessa língua conforme avançam na construção de suas proficiências; enfim, nota-se, no Brasil, que falta ainda um maior desenvolvimento nessas linhas de pesquisa, especialmente para descrever dificuldades de aprendizes de PLA com diferentes *backgrounds*.

Essa é, assim, uma das questões interessantes com relação à pesquisa com *corpus* de aprendizes/estudantes/examinandos em testes de proficiência: quais são os elementos linguísticos que ou os critérios capazes de diferenciar níveis de proficiência em LA? Alguns dos elementos e critérios mais fundamentais já foram encontrados através do uso do *Cambridge Learner Corpus*, no decorrer do projeto *English Profile Programme* (HAWKINS e BUTTERY, 2009). Embora alguns pesquisadores estejam apreensivos quanto a essa pesquisa (HULSTIJN, 2010), ela será capaz de mostrar possibilidades interessantes para o uso de grandes conjuntos de dados de aprendizes, oferecendo subsídios para a classificação semiautomática de proficiência do inglês.

4.3 PROJETOS E FERRAMENTAS COM CORPUS

Corpora contribuem para a descrição da língua, por exemplo, através da construção de recursos como dicionários e gramáticas. A seguir, apresento alguns dos principais projetos existentes e ferramentas existentes hoje (baseado em CANDIDO JR., 2007, p. 45-48).

4.3.1 CORPORA INTERNACIONAIS

- American National Corpus (ANC): *corpus* de inglês americano; textos posteriores a 1990, com 100 milhões de palavras no total; falado e escrito.
- British National Corpus (BNC): *corpus* de inglês britânico; 100 milhões de palavras; falado e escrito.
- Brown Corpus of Standard American English: projeto de *corpus* para o inglês criado em 1964 (o primeiro *corpus* desenvolvido).
- Czech National Corpus (CNC): *corpus* para o tcheco; mais de 20 milhões de palavras.
- FRANTEXT (Trésor de la Langue Française): *corpus* para do francês constituído por 2 mil textos, totalizando mais de 114 milhões de palavras; textos dos séculos XVIII, XIX e XX.
- International Corpus of English (ICE): projeto iniciado em 1990; tinha o objetivo de comparar o inglês falado em diferentes partes do mundo; *subcorpus* com 1 milhão de palavras para 15 variantes nacionais do inglês com textos produzidos após 1989.
- The Bank of English: falado e escrito; textos produzidos após 1990; possui mais de 450 milhões de palavras.

4.3.2 CORPORA DO PORTUGUÊS

- AC/DC (Acesso Corpora/Disponibilização Corpora): faz parte da Linguateca; agrega textos de 17 projetos de *corpus* diferentes, totalizando mais de 54 milhões de palavras.
- Corpus do NILC (Núcleo Interinstitucional de Linguística Computacional): *corpus* desenvolvido durante a criação da ferramenta ReGra (Revisor Gramatical), corretor gramatical construído para a língua portuguesa e integrado ao MS-Word; possui mais de 41 milhões de palavras; possui vários gêneros, com grande participação de textos jornalísticos.
- Corpus do Português: variantes brasileira e portuguesa do português; conta com 45 milhões de palavras.
- Lácio-Web: compilação de *corpus* do português do Brasil e implementação de ferramentas para análises linguísticas. Quatro *corpora* foram desenvolvidos durante o projeto: (a) *corpus* referência de português contemporâneo (Lácio-ref); (b) *corpus* fechado e manualmente anotado morfossintaticamente, composto de aproximadamente 1 milhão de palavras (Mac-morpho); (c) *corpus* de textos em inglês e português com originais e traduções

(Par-c); (d) *corpus* de textos originais em português e inglês com conteúdos similares (Comp-c).

- PLN-BR: recursos e ferramentas para a recuperação de informação em bases textuais em português do Brasil.

- Tycho-Brahe: *corpus* de português histórico composto por 40 textos dos séculos XVI e XIX; possui 2 milhões de palavras.

- Banco de Português: textos de revistas e jornais, literatura, textos acadêmicos e de negócios, conversas, reuniões, aulas, conversas telefônicas e entrevistas; possui 230 milhões de palavras.

4.3.3 FERRAMENTAS PARA COMPILAÇÃO DE TEXTOS

- Digitalizadores: digitalizam textos impressos, convertendo-os para o formato de imagem.

- Reconhedores óticos (OCR): convertem documentos em formato de imagem para o formato texto; ferramenta sujeita a erros de conversão.

- Buscadores Web: permitem a coleta de textos disponíveis via Web. Os textos são recuperados por palavras chaves. Exemplos de buscadores são o Google e o Yahoo.

- Navegadores off-line: armazenam *sites* no computador do usuário. O usuário pode analisá-los posteriormente. Um exemplo de ferramenta deste tipo é o HTTrack.

- Mineradores Web (Web crawlers): ferramentas que varrem a Web, acessando diversos *sites*; podem armazenar textos lidos na Web a partir de critérios de armazenamento definidos pelo usuário. Um exemplo de minerador é o HTDig.

- Conversores de formato: são úteis para converter textos em diversos formatos coletados via Web (por exemplo, HTML, DOC, PDF, PS) para texto sem formatação (TXT). Um exemplo de conversor é o XPDF, capaz de converter o formato PDF para texto.

4.3.4 ANOTADORES MANUAIS

- Editores de texto comuns: permitem a edição e inserção nos textos e etiquetas. Entretanto, oferecem poucos recursos para o tratamento de etiquetas. Um exemplo é o editor Emacs, que diferencia as etiquetas do restante do texto através de mudanças nas cores do elementos (destaque de sintaxe ou *syntax highlighting*).

- Editores para anotação XML/SGML: possuem tratamento próprio para etiquetas XML. O editor Xanthipe é um exemplo de editor avançado e chegou a ser usado para editar etiquetação sintática no BNC. Esse editor é capaz de realizar alterações em série nos textos. O editor Open XML Editor é capaz de checar erros de sintaxe em documentos XML e analisar a integridade do documento a partir de DTDs.

4.3.5 ANOTADORES AUTOMÁTICOS

- Segmentadores: dividem o texto em sentenças e parágrafos de forma automatizada, inserindo as etiquetas adequadas. Um exemplo desta ferramenta é o Senter, desenvolvido no NILC.

- Etiquetadores morfossintáticos e sintáticos: utilizados durante a etapa de compilação do *corpus*, os etiquetadores inserem automaticamente informações morfossintáticas e sintáticas com um alto grau de precisão. A precisão dos etiquetadores morfossintáticos chega a 98%, enquanto que a precisão de etiquetadores sintáticos chega a 91% (para o inglês). Um exemplo de software pertencente a esta categoria é o etiquetador *Palavras*, baseado em regras construídas manualmente. O *Palavras* etiqueta textos em português e sua precisão é superior a 97% (nível morfossintático). Há três etiquetadores morfossintáticos treinados por regras automáticas disponibilizados no NILC: Tree Tagger, MXPOST e TBL Tagger. Os etiquetadores foram treinados a partir do *corpus* MacMorpho e possuem uma precisão superior a 96%. O resultado do treinamento está disponibilizado publicamente.

- Lematizadores: encontram e anotam o lema (ou forma canônica) das palavras presentes no *corpus*. A lematização é um processo útil para estudo de palavras com muitas flexões, como os verbos. Ferramentas desta categoria também devem tratar ambiguidade. Por exemplo, a forma flexionada “foi” pode ter como lema o verbo “ser” ou o verbo “ir”. Essas ferramentas são particularmente úteis para tarefas lexicográficas. Um exemplo é o lematizador de verbos do português LX Lemmatizer.

- Anotadores de co-referência: são ferramentas capazes de identificar expressões que se referem a um mesmo elemento dentro de um texto. Como exemplo, é possível definir o elemento a que um pronome se refere. Anotadores de co-referência são úteis na análise do discurso e para sumarização. O sumarizador automático RHeSumaRST identifica co-referências automaticamente a partir de heurísticas.

- Anotadores diversos: aplicados a outros níveis linguísticos não descritos acima. Um exemplo é o RSTool, que atua no nível retórico e o segmentador de estruturas esquemáticas de resumos em português, também desenvolvido no NILC e disponível no ambiente SciPo.

4.3.6 FERRAMENTAS PARA ACESSO A CORPUS

- Concordanceadores: são os principais tipos de ferramentas, pois permitem o estudo do conteúdo do *corpus* através de buscas sofisticadas. Além disto, estas ferramentas são capazes de alinhar diversos resultados e exibi-los simultaneamente. Os concordanceadores podem ser agrupados em três tipos: (a) KWIC (KeyWord In Context) no qual as ocorrências buscadas são exibidas juntamente com seu contexto (blocos de texto a direita e a esquerda de cada ocorrência buscada), (b) KWAC (KeyWord And Context) no qual as ocorrências e seus contextos são exibidos separadamente (de forma a destacar as ocorrências) e (c) KWOC (KeyWord Out of Context) no qual as ocorrências são exibidas separadamente de seus contextos e exibidas novamente dentro dos contextos. A maior parte dos processadores de *corpus* possuem concordanceadores, entre eles, o Unitex. Um concordanceador via Web de resultados da Web é disponibilizado pelo processador WebCorp.

- Buscadores textuais: realizam buscas no *corpus* de forma semelhante ao concordanceador, mas exibem apenas um resultado por vez. Alguns processadores de *corpus* não possuem buscadores textuais, pois estes são substituídos pelos concordanceadores. Exemplo de buscadores textuais podem ser encontrados em editores de texto simples como o Notepad, distribuído juntamente com MS-Windows.

- Buscadores de dados de cabeçalho: realizam buscas nos metadados que descrevem um *corpus*. São úteis tanto para a obtenção de informações de bibliografia utilizadas para análise de balanceamento e representatividade do *corpus* quanto para a formação de *subcorpus* para pesquisas específicas. Por exemplo, é possível formar um *subcorpus* com todas as obras de um autor ou todas as obras de um período. O processador de *corpus* Philologic possui um buscador de dados de cabeçalho. O projeto Lácio-Web também possui e este permite 3 tipos de buscas desde simples a sofisticadas.

- Contadores de frequência: contam o número de palavras total em um *corpus* ou *subcorpus*, além de mostrar as palavras presentes no *corpus* e suas frequências. As ferramentas devem ignorar pontuação, números e outros tipos de símbolos, pois estes não representam palavras. Da mesma forma, as etiquetas, caso presentes, devem ser ignoradas. Os

contadores de palavras mais utilizadas contam formas (*tokens*) como é o caso do contador do MS-Word que conta palavras ortográficas (separadas por brancos). O contador do projeto Lácio-Web, em especial, conta alguns padrões de multipalavras.

- Geradores de n-gramas: são ferramentas capazes de reconhecer e levantar n-gramas um texto. N-gramas consistem em sequências de n palavras pertencentes a um texto e são usados em PLN para extração de estatísticas sobre o texto e em Recuperação de Informações (RE) para a execução de buscas. Um exemplo é a ferramenta NSP (Ngram Statistics Package).

- Buscadores de colocações: colocações de uma dada palavra são enunciados de uso comum na língua nos quais a palavra é empregada. Um exemplo para a palavra “arma” é o enunciado “armas de destruição em massa”. Buscadores de colocações são capazes de encontrar e exibir as colocações através de técnicas de análise estatística. O Philologic é capaz de gerar colocações.

- Buscadores léxicos: permitem a pesquisa das palavras de um ou mais léxicos (dicionários computacionais) em um *corpus*. O concordanceador do Unitex é capaz de realizar buscas léxicas através léxicos.

4.3.7 FERRAMENTAS DE EXTRAÇÃO DE CONHECIMENTO

- Sumarizadores: realizam resumo automático de um texto. O uso do *corpus* é útil tanto para a construção do sumarizador, como para sua avaliação. Exemplos de sumarizadores incluem as ferramentas EXPLORA (Exploração de Métodos Diversos para a Sumarização Automática) e GistSumm (Gist Summarizer), desenvolvidos no NILC.

- Tradutores automáticos: este tipo de ferramenta pode ser usado em *corpora* paralelos para aprendizado de regras de tradução automática e para avaliação automática de qualidade da tradução. A avaliação também pode ser feita (manualmente) com base em *corpora* monolíngues. No NILC é desenvolvido um tradutor automático entre português e UNL (Universal Networking Language).

- Ferramentas gerais de recuperação de informação: são ferramentas para recuperação de documentos com base em buscas por palavras chaves. É possível avaliar a qualidade de uma busca a partir das medidas precisão (a proporção de entre documentos relevantes e irrelevantes recuperados) e *recall* (a proporção entre documentos relevantes recuperados e o número total de documentos relevantes).

4.4 LIMITES DA LINGUÍSTICA DE CORPUS: POSSIBILIDADES NO PLN

Conforme o que foi exposto na seção anterior, a LC oferece caminhos para uma observação em que é possível entender e descobrir características e propriedades linguísticas mais típicas e específicas de determinados grupos de falantes, gêneros textuais e níveis de proficiência. O desafio está em descobrir que características e propriedades são essas. Os trabalhos que foram citados anteriormente lidam com *lexical bundles*, com a observação de uso de tempos verbais inadequados, com a observação de ocorrências de problemas de concordância, com o uso da ordem de palavras, sempre privilegiando o contraste entre a produção de falantes nativos (que são os autores/produtores dos *corpora* “ideais”) e não nativos (autores/produtores dos *corpora* de estudo). Existe, no contexto desses trabalhos, via de regra, a descrição da dependência de determinados “erros” a diferentes *backgrounds* dos estudantes que os cometem e a descrição da dependência de determinados usos a contextos correlacionados.

O que foi possível verificar é que a LC é, *grosso modo*, uma linguística que faz contas e trabalha com a quantificação da linguagem. Os perfis de trabalhos citados nas seções deste capítulo fazem uso de ferramentas bastante simples de análise lexical, que oferecem dados interessantes com a finalidade de problematizar questões de linguagem. A LC se presta muito a isso: para mostrar, na observação em larga escala, os padrões de uso da linguagem e permitir que o linguista apresente fatos concretos/primários e dados para problematizar esses usos e criar fatos secundários, as suas descrições.

No entanto, chega-se ao ponto em que a LC é, realmente, um ponto de partida, mas não é o ponto de chegada. Sua metodologia e ferramental, quando postos ao lado de ferramentas computacionais estatísticas mais robustas, parece já não ser mais suficiente. Até aqui, seu arcabouço teórico se prestou bem para a compreensão da visão de língua e linguagem que se tem ao adotar o *corpus* como ponto de partida. Mas o objetivo deste trabalho é fornecer uma solução, e não apenas criar um novo problema. Existem, no cenário acadêmico, trabalhos de linguistas que fazem uso de *corpora* em suas análises, mas que, no entanto, já não fazem uso da metodologia da LC para chegarem às suas conclusões. Os grandes problemas encontrados atualmente por linguistas que fazem uso de *corpora* em suas pesquisas são encontrar a facilidade ao lidar com estatística e a viabilizar estudos texto a texto, e não em grandes pacotes textuais, tratando o *corpus* como uma massa de palavras.

No contexto brasileiro, exemplos recentes de trabalhos desse tipo são os de Souza (2011) e Pasqualini (2012). Souza (2011), por exemplo, descreveu traços linguísticos

característicos de textos históricos, correlacionando-os a seus respectivos gêneros, propondo uma tipologia de traços para identificar o gênero de cada texto automaticamente. Utilizou postulados metodológicos da LC e ferramentas tipicamente presentes nesse tipo de estudo, como o Philologic²⁵ e o Unitex²⁶. No entanto, para realizar a classificação automática, fez uso de algoritmos comuns a outra área, o Processamento de Língua Natural (PLN).

Pasqualini (2012), por sua vez, abordou o tema da complexidade textual em traduções de literatura em língua inglesa produzidas no Brasil, também fazendo uso dos pressupostos da LC. Ao buscar comprovar a hipótese de que textos traduzidos são mais complexos do que seus originais, fez uso das ferramentas de PLN como o Coh-Metrix e o Coh-Metrix-Port. Sendo frutos de parceria entre informatas e linguistas, o estudo foi capaz de mostrar que as traduções para o português produziram textos mais complexos do que seus textos-fonte, considerando algumas das métricas devolvidas pela ferramenta.

Mais importante que isso, para este trabalho, foi a constatação da existência de pesquisas, no âmbito da língua inglesa, que tinham – e ainda têm – como finalidade mensurar ou aferir proficiência escrita em língua adicional de forma automática. Essas pesquisas fazem uso de técnicas que não pertencem a LC, muito embora trabalhem com *corpora*, mas utilizam técnicas de PLN para obter soluções mais apuradas e confiáveis. Jarvis et al. (2003) e Ferris (2002), por exemplo, utilizam medidas consideradas impressionistas – índices superficiais tais como diversidade lexical, repetição de palavras e tamanho do texto – para aferir proficiência escrita em língua adicional.

Além das medidas lexicais, outras medidas já existentes, como *sentido e intenção*, *níveis de linguagem* e *análise conceitual* ofereciam valores interessantes para a finalidade de prever níveis de proficiência escrita. Porém, essas medidas dificilmente são apreendidas automaticamente pelas ferramentas computacionais disponíveis. Das ferramentas existentes, os pesquisadores de psicolinguística da Universidade de Memphis fizeram um interessante progresso nesse sentido ao construírem a ferramenta Coh-Metrix (GRAESSER et al. 2004; MCNAMARA et al. 2002; CROSSLEY et al. 2007).

Como o nome da ferramenta sugere – *Cohesion Metrics* –, a finalidade do sistema é a de mensurar a coesão, a coerência e a inteligibilidade de um dado texto em inglês, explorando três grandes níveis de análise linguística: lexical, sintático e semântico. De acordo com Graesser et al. (2004), a coesão textual é uma propriedade objetiva do texto, e a coerência é a representação mental do conteúdo do texto feita pelo leitor através das palavras, sentenças e

²⁵ Disponível em <<https://sites.google.com/site/philologic3/>>. Acesso em 28 mai 2012.

²⁶ Disponível em <<http://www-igm.univ-mlv.fr/~unitex/>>. Acesso em 28 mai 2012.

frases que orientam a leitura e conectam as ideias umas às outras. Sendo a coesão a parte objetiva do texto, a ferramenta é capaz de calcular e conferir valores a 600 métricas, disponíveis em sua versão fechada (108 na versão aberta).

De acordo com um estudo recente de Crossley e McNamara (2012), as métricas calculadas pelo Coh-Metrix foram consideradas capazes de prever a proficiência escrita em inglês como LA. Dentre as métricas avaliadas pela ferramenta, as métricas diversidade lexical, frequência de palavras, significado atribuído pelo leitor, repetição de aspecto verbal e familiarização com a palavra foram as consideradas mais distintivas entre os níveis de proficiência, portanto, mais produtivas para realizar a classificação entre textos mais e menos avançados escritos em língua inglesa. No estudo, essas foram as métricas que se mostraram distintivas em um conjunto de textos escritos por estudantes de uma escola de Hong Kong no Hong Kong Advanced Level Examination (HKALE²⁷), um exame de proficiência de inglês anual obrigatório aos estudantes chineses.

Os trabalhos mencionados apontam que, por exemplo, quanto maior o nível de proficiência, maior será a variedade/diversidade de vocabulário utilizado, naquele contexto de produção de textos sob estudo. É evidente que muitos desses apontamentos podem ser discutidos; essa afirmação, por exemplo, pode ser contestada em outros contextos, como já foi verificado em redações de vestibular (FINATTO et al., 2008), em que redações que receberam nota maior possuíam menor variedade lexical, indicando que se detinham mais ao tema proposto e se mostravam mais coesas através da repetição de palavras. Para aquele gênero e tipo de exame, essas medidas seriam relativizadas e entendidas de outra forma.

No cenário do PLA, há poucos trabalhos com corpora e que usem os pressupostos teóricos da LC. É evidente, por causa disso, que existem ainda menos trabalhos que envolvam técnicas de PLN. Em virtude disso, não foi possível encontrar qualquer pesquisa relacionada a esse tipo de desenvolvimento. Então, como dito anteriormente, o ponto de partida desta pesquisa é o estado da arte existente no inglês. Porém, há que se lembrar que o exame Celpe-Bras, ao contrário de exames como Cambridge, TOEFL ou mesmo o HKALE, mencionado anteriormente, não têm como objetivo aferir objetiva e exclusivamente ou de forma separada os conhecimentos gramaticais do examinando, a fim de descobrir, por exemplo, por quantos anos ele estudou a língua, quais são as estruturas mais complexas que consegue produzir ou o vocabulário o qual consegue fazer uso. São exames diferentes e essas diferenças serão aqui relativizadas.

²⁷ Maiores informações sobre este exame podem ser encontradas em <<http://www.hkeaa.edu.hk/en/hkale/>>. Acesso em 22 jul 2012.

Em se tratando da língua portuguesa, as medidas apontadas por Crossley e McNamara (2012), bem como as presentes em outros trabalhos, servem como um ponto de partida. No cenário de PLA, conforme já foi reiterado algumas vezes neste texto, inexistem estudos que tenham verificado se essas medidas, no contexto em que estão sendo colocadas, seriam de alguma relevância para mensurar proficiência. Levando em conta que esse tipo de indicativo já se mostrou capaz de avaliar diferentes níveis de proficiência em outras ocasiões, mesmo que em outra língua, acredito que este caminho é produtivo para descobrir algum ponto de semi-automatização.

5 PROCESSAMENTO DE LÍNGUA NATURAL

Em virtude dos limites atuais da LC – especialmente no que diz respeito à manipulação de dados de textos em isolado, de textos reunidos em *corpora* e de acurácia estatística –, este trabalho foi buscar na Linguística Computacional ou Processamento de Língua Natural (PLN) metodologias e técnicas de pesquisa que viabilizassem uma observação quantitativa mais aprimorada. Isso se deu principalmente em razão da dimensão do *corpus* reunido, sabendo de antemão que os recursos de PLN poderiam oferecer maior acuidade para uma análise multifatorial com muitos elementos em correlação, mesmo em um número relativamente pequeno de textos.

A fim de clarificar o que é e de onde vem o PLN, é importante trazer aqui algumas informações mais gerais da área, visto que o linguista que estiver lendo este trabalho pode estar entrando em contato com o PLN pela primeira vez. O PLN é uma subárea da Inteligência Artificial, ramo da Ciência da Computação. Acredito que esteja já claro para todos que muitos computadores e equipamentos eletrônicos, hoje, obrigam seus usuários a aprenderem formas não intuitivas de comunicação com a finalidade de darem comandos precisos a esses aparelhos. Isso se dá através de linguagens de programação específicas, que estão por trás de qualquer menu, *link* ou botão que tocamos/selecionamos em um aparelho. Assim são as interfaces existentes entre máquinas e seres humanos, que estão ficando, felizmente, cada vez mais sofisticadas e caminhando, aos poucos, em direção a formas mais humanas, naturais e intuitivas de comunicação.

O PLN, dessa forma, agrupa métodos formais utilizados para analisar textos e gerar frases escritas em língua natural, ou seja, uma língua falada por pessoas. Os computadores normalmente estão aptos a compreenderem instruções escritas em linguagens computacionais – tais como Java, C++, Python, Perl –, mas possuem muita dificuldade para entender comandos simples escritos em uma língua humana. Isso se deve ao fato de as linguagens computacionais serem extremamente precisas, contendo regras fixas e estruturas lógicas bem definidas que permitem ao computador saber exatamente como proceder a cada comando. Já em uma língua, uma simples frase pode conter ambiguidades e interpretações que dependem

de contexto, do conhecimento de mundo, de regras gramaticais e culturais e de conceitos abstratos.

O objetivo final do PLN, então, é capacitar computadores a entender e compor textos em língua natural. Por "entender" queremos dizer serem capazes de reconhecer o contexto, fazer a análise sintática, semântica, léxica e morfológica, criar resumos, extrair informação, interpretar os sentidos e até aprender conceitos através de textos processados. Daí a área fazer parte da Inteligência Artificial. Não se sabe se um dia os computadores poderão igualar a capacidade humana de entender e compor textos e atualmente essas capacidades ainda são bastante limitadas, mas muitos resultados práticos já existem e são utilizados por diversos tipos de programas e para muitas finalidades.

Em resumo, o PLN pode ser definido como uma disciplina que estuda e desenvolve a habilidade de um computador processar a mesma língua que os seres humanos usam no seu dia a dia (ROSA, 2011, p. 137). Para demonstrar algo que o PLN possibilitou e que usamos cotidianamente ao redigir e-mails e outros tipos de texto, podemos tomar um exemplo corriqueiro: qualquer pessoa que tenha utilizado uma ferramenta de processamento de texto, como o MS Word, OpenOffice ou Pages, sabe que ela contém um corretor ortográfico, uma ferramenta que destaca possíveis erros ortográficos e gramaticais e propõe correções. Esse corretor ortográfico é um clássico exemplo de uma aplicação do PLN. Os primeiros corretores ortográficos funcionavam através da comparação de uma lista de palavras extraídas do texto com uma lista de palavras (dicionário de palavras) corretamente grafadas, uma tarefa extremamente simples e que não exige processamento complexo. Com o passar dos anos, essas ferramentas, tão úteis, tornaram-se bem mais sofisticadas. São capazes de detectar erros relacionados não só à ortografia, como à morfologia (formação de plurais) e à sintaxe (ausência de um verbo ou falta de concordância), apontar problemas de pontuação e até sugerir itens lexicais que sejam mais adequados ao tipo de texto que se está produzindo (acadêmico, jornalístico, entre outros).

De acordo com Vieira e Strube de Lima (2001), as pesquisas em PLN, para além de simples corretores ortográficos, incluem tarefas muito mais complexas e ainda não totalmente resolvidas, como o reconhecimento, a interpretação, a tradução e a geração de linguagem. Essas tarefas de PLN, por serem diversificadas, conforme bem apontam as autoras, demandam o diálogo intenso entre Cientistas da Computação e profissionais e pesquisadores de diversas outras áreas, tais como linguistas e psicólogos. Estes são apenas alguns exemplos da interação que a Computação precisa estabelecer de modo a desenvolver trabalhos de PLN. Para Martins (2011),

embora seja evidentemente profícuo na produção de aplicativos de utilidade incontestável, [o PLN] constitui principalmente uma dispersão, sem que possa ser observada, nitidamente, a hegemonia de um corpo teórico sobre os demais. Trata-se, na verdade, de uma coleção de posturas difusas e fragmentárias que orbitam um objetivo comum: ensinar a máquina a falar. (2011, p. 291)

Assumindo aqui o ponto de vista de um linguista, posso dizer que a interação com cientistas da computação é, algumas vezes, fonte de mal-entendidos. Isso ocorre porque cientistas da computação, via de regra, esperam dos linguistas uma concepção pronta de língua ou dos problemas linguísticos que estamos, justamente, tentando compreender; muitas vezes, esperam de nós uma teoria linguística capaz de ser computada, formalizada e processada. Essa abordagem acontece porque o PLN é aplicado, ou seja, busca soluções para problemas, algo que, na maioria das vezes, a Linguística não é.

Da parte dos linguistas, geralmente esperam ansiosos que os cientistas da computação os sirvam, tragam soluções imediatas aos seus problemas de processamento de *corpus* e que “façam as contas” necessárias, resolvendo estatisticamente todas as questões. Essa postura demonstra um verdadeiro desconhecimento sobre procedimentos e interesses envolvidos nas tarefas de PLN. Evidente que muitos mal-entendidos entre linguistas e pesquisadores de PLN ocorrem justamente por se tratarem – a Linguística e o PLN – de áreas de trajetórias históricas e epistemológicas diferentes.

Por ser essencialmente aplicado, o PLN prioriza a construção de sistemas que têm a capacidade de reconhecer e de produzir informação apresentada em língua natural. Assim, trabalhos em PLN buscam uma solução prática para um problema ou necessidade definidos. Até hoje, já muito se avançou na área, mas compreender e reproduzir plenamente a língua e a linguagem via métodos computacionais é ainda um grande problema a ser resolvido.

O tratamento computacional de línguas naturais tem como tarefa investigar e criar modelos formais de língua que possam ser operados pelo computador (DIAS DA SILVA, 2006). Quando se fala em modelos de língua, porém, há que se ressaltar que a manipulação da fala, por apresentar problemas teóricos e tecnológicos específicos, acaba sendo desenvolvida de forma independente, principalmente no âmbito da Engenharia Elétrica ou de Processamento de Sinais em Computação. Assim, quando se refere ao PLN, atualmente está-se, na verdade, falando do processamento da língua natural escrita (VOLPE NUNES, 2008).

5.1 PLN: BREVE HISTÓRICO E MODO DE TRABALHO

Grandes *corpora* contribuem para a descrição e compreensão da língua, sendo um auxílio importante para se progredir de maneira rápida e confiável nessa tarefa. Seja por meio da construção de recursos, como dicionários e gramáticas, ou de ferramentas, como lematizadores e etiquetadores morfossintáticos, largamente utilizados para o processamento desses *corpora*, o PLN contribui para essa descrição oferecendo técnicas confiáveis e flexíveis para o processamento de dados. Atualmente, o número de dados linguísticos ou *corpora* existentes é enorme²⁸, e muitos outros *corpora* estão sendo compilados ano a ano, com o objetivo de serem cada vez maiores, mais balanceados, mais flexíveis e mais representativos (CANDIDO JR, 2007).

Acompanhando o desenvolvimento e ampliação da capacidade dos computadores, desde a década de 50 o PLN vem sofisticando suas ferramentas, o que permite, hoje, a realização de tarefas cada vez mais complexas de forma mais eficiente. Sua principal tarefa, naquela época e durante muito tempo, foi – e ainda é – a tradução automática (TA), na expectativa de que algum dia o computador se torne capaz de traduzir perfeitamente um texto da língua-fonte para a língua-alvo. Esse objetivo, ainda não concluído, perdeu um pouco sua força na área, mais ainda é bastante presente em pesquisas e eventos de PLN.

A ideia de usar computadores para a tradução das línguas naturais surgiu em 1946 e recebeu financiamentos substanciais nos anos 50 e, posteriormente, nos anos 80. A tradução automática encontra-se hoje ainda longe de corresponder às expectativas geradas em seus primeiros anos de investigação. Em um nível mais básico, pode ser realizada através da substituição das palavras de uma língua pelas de outra. Mas uma boa tradução requer a correspondência precisa entre unidades de texto/de significados maiores (sintagmas, frases ou mesmo textos completos). No caso do português em particular, a falta de recursos para a desambiguação do significado das palavras – ontologias lexicais e *corpora* anotados – bem como de *softwares* desenvolvidos a partir desses dados é um dos motivos para os resultados ainda serem insatisfatórios.

²⁸ A título de exemplo, podemos citar novamente o Banco de Português, um *corpus* do português brasileiro escrito e falado, com 233 milhões de palavras, disponível em <<http://www.pucsp.br/pos/lael/>>. Há também o mais recente *corpus* do português brasileiro, o Corpus Brasileiro, disponível em <<http://corpusbrasileiro.pucsp.br/cb/Inicial.html>>, contendo um bilhão de palavras. E o já consagrado WEBCORP: a Web como um *corpus*, disponível em <<http://www.webcorp.org.uk/>>, utilizando toda a *internet* como um grande *corpus*.

Por outro lado, embora reconheça-se que ainda há falta de conhecimento linguístico para a melhora da TA, uma das frentes do PLN é justamente pôr em cheque a utilidade de conhecimentos linguísticos para a melhora da execução desta tarefa em especial. O tradutor do Google, por exemplo, possui um desempenho cada vez melhor e tem como base apenas modelos estatísticos e não linguísticos. O projeto *Arcalabouço*, do professor Ronaldo Martins, é uma boa fonte para mais informações sobre o assunto (disponível em <www.ronaldomartins.pro.br/research.htm>).

Renata Vieira (2004) afirma que a Linguística Computacional preocupa-se com a compreensão da língua e de técnicas computacionais adequadas para o tratamento da língua escrita e falada, tanto para sua interpretação quanto sua geração, e que o PLN tem o objetivo de reproduzir comportamentos inteligentes em sistemas computacionais, como a solução de problemas e automatização do raciocínio. A natureza interdisciplinar do PLN, em que participam matemáticos, cientistas da computação e linguistas, é, por si só, uma fonte de mal-entendidos entre os pesquisadores, sobretudo por agrupar pesquisadores de áreas com tradições tão diversas. Por um lado, os cientistas da computação esperam dos linguistas uma concepção pronta de língua, matematizável, formalizável e processável; os linguistas, por sua vez, esperam dos cientistas da computação soluções instantâneas para problemas encontrados em seus projetos.

Volpe Nunes (2008), a esse respeito, afirma que as expectativas do usuário, após um convívio mais intenso com o computador, são mais realistas – como também são mais realistas as expectativas dos envolvidos na execução das tarefas de PLN. A autora, ao contrário de Martins, acredita que a complexidade da tarefa foi inicialmente subestimada e que, por isso, as abordagens ao problema da tradução automática foram, de certa forma, ingênuas – tanto da parte dos cientistas da computação quanto dos linguistas. Desse embate, fica claro que a expectativa irrealista de desempenho dos sistemas de tradução automática são, antes de mais nada, consequência das expectativas irreais em relação ao que cada uma das disciplinas contribuiria para a concretização da tarefa.

No que diz respeito à tradução automática, é fundamental que se tenha claro a quem servirá um sistema automatizado de tradução. O que me parece irônico é o fato de que quem mais se beneficiaria seriam os tradutores profissionais – irônico porque, em vez de tentar auxiliar operações humanas de tradução, a tradução automática parece se colocar numa posição de competição com os tradutores. A tradução automática se beneficiaria imensamente se ampliasse os seus objetivos também às necessidades reais da atividade profissional de tradutores reais. Para o “público geral”, os sistemas atuais de tradução automática são

razoavelmente aceitáveis. É quando um tratamento mais sofisticado do texto é necessário que os sistemas apresentam desempenho insatisfatório. Justamente nesse ponto é que a contribuição de tradutores humanos e de profissionais do texto é fundamental.

É também nesse espírito que esta dissertação, uma pesquisa circunscrita à Linguística, faz uso de ferramentas de PLN: com o intuito de extrapolar a ênfase nos resultados obtidos a fim de, futuramente, usar esses dados para projetar ferramentas mais adequadas às necessidades específicas de usuários específicos – e aqui me refiro, sobretudo, a tradutores e a revisores. Assim, como mencionei anteriormente, a motivação desta pesquisa foi também a de dialogar produtivamente com os pesquisadores de PLN, apontando problemas surgidos na prática profissional concreta de tradutores/revisores – neste caso, o fenômeno da complexidade textual em traduções literárias – e, por meio do uso de ferramentas criadas em PLN, contribuir para o aperfeiçoamento desses recursos.

Já a tradução (se utilizar conhecimento linguístico) requer um grau de 4 processamento de língua muito mais complexo, dado que trabalha com vários níveis de descrição, tais como ortográfico, morfossintático, lexical, semântico e às vezes retórico. Enfim, sugiro que toda essa parte seja refeita.

Para tarefas como a de correção ortográfica e de tradução automática, há um passo a passo para realizar o processamento da linguagem que é básico. Esse passo a passo é constituído por três módulos, essenciais a todos os trabalhos de PLN:

- pré-processamento: limpeza dos dados, análise ou remoção da formatação, e detecção do idioma;
- análise gramatical: detecção do verbo e dos seus complementos e modificadores, detecção de elementos de outras categorias, identificação da estrutura das frases;
- análise semântica: desambiguação (por exemplo, qual dos significados de “bateria” é usado em determinado contexto?), resolução de anáforas (por exemplo, que pronome recupera a referência de outra expressão na frase?), e representação do significado da frase num modelo interpretável pela máquina.

Para um linguista, é importante compreender como se desenvolve um trabalho em conjunto com um informata ou como pode ser desenvolvido um trabalho em PLN. Dias da Silva (1996), que transitou com maestria entre ambas as áreas, foi capaz de propor uma estratégia, dividida em três etapas, para o desenvolvimento de projetos em PLN, representadas na Figura 5 a seguir:

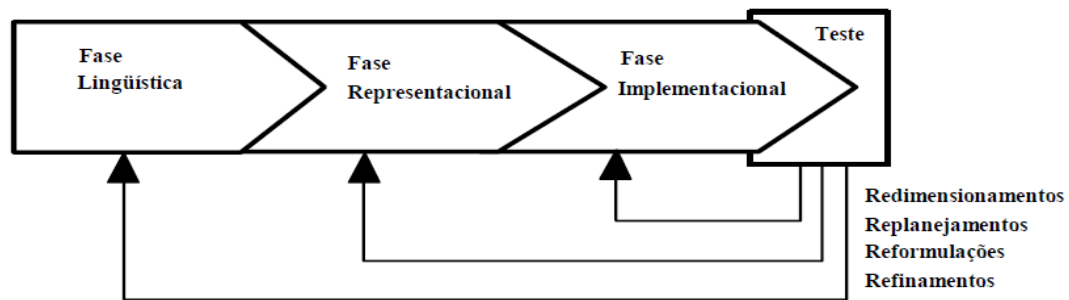


Figura 5 Etapas estratégicas para o desenvolvimento de um trabalho em PLN.²⁹

Na Fase Linguística, cria-se o modelo linguístico; na Representacional, procura-se traduzir o modelo de modo que o computador o entenda; na Implementacional, objetiva-se pôr o modelo em prática. Para o autor (1996, p. 178), o equacionamento do domínio representacional do PLN envolve a discussão de questões em três níveis:

- **morfo sintático**, que trata da representação das gramáticas e dos analisadores gramaticais, incluindo a representação das regras e das estruturas morfo sintáticas e de léxicos enriquecidos com informações pragmático-discursivas;
- **semântico**, que trata da representação de estruturas semânticas, de domínios conceituais e de estratégias computacionais de interpretação dessas representações;
- **pragmático-discursivo**, que trata da representação da estrutura do discurso e dos contextos pragmático-discursivo e situacional.

Bento Dias da Silva (1996) já afirmava que essas representações precisam ser explícitas, consistentes e não ambíguas para serem transformadas em programas de computador. Além disso, não necessariamente precisam ocorrer de forma sucessiva, podem ser realizadas simultaneamente a partir da etapa linguística, que é a base para as outras. Nessa etapa linguística, como é natural, esperar-se-ia a atuação de linguistas em diálogo com os cientista de computação.

Hoje, apesar de lidar com muitos problemas, o PLN constituiu uma comunidade científica e acadêmica em crescimento ao redor do mundo bastante forte, e são muitos os trabalhos realizados, principalmente para o inglês, o espanhol, o alemão, o francês e o japonês. Há, no entanto, ainda, uma enorme carência de interlocutores linguistas formados para o diálogo interdisciplinar respeitoso e produtivo.

Existem pólos, porém, fazendo grandes esforços no sentido de impulsionar os estudos do português, como é a iniciativa das Faculdades da Universidade de Lisboa, com cursos e unidades de investigação ativas no campo da tecnologia da linguagem – os Centros de

²⁹ Fonte: Dias da Silva (1996, p. 78).

Investigação em Tecnologias de Informação (CITI) e de Linguística (CLUNL). Ainda em Lisboa, existe o Instituto de Linguística Teórica e Computacional (ILTEC), que foi criado para receber o projeto EUROTRA; e na Universidade do Porto há dois centros, o Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC) e o Centro de Linguística (CLUP). No Brasil, temos centros no Ceará – o CompLin (Computação e Linguagem Natural)³⁰ –, no Rio Grande do Sul – o Grupo de pesquisa em PLN³¹ da PUCRS e no Instituto de Informática da UFRGS, em parceria com o Instituto de Letras da mesma universidade – e em São Carlos - SP – o Laboratório de Interação Avançada da UFSCAR e o NILC (Núcleo Interinstitucional de Linguística Computacional)³² da USP. O NILC é o grupo pioneiro do Brasil e uma referência para o tratamento computacional do português brasileiro em diferentes níveis e frentes.

Apesar dos esforços, carecemos da construção de recursos de base para língua portuguesa, como léxicos e bases de dados lexicais, ontologias e vocabulários, gramáticas e *parsers*, e grandes coleções anotadas para processamento, não apenas compilação de corpora. Neste contexto, para o português (europeu ou brasileiro), serão úteis as ontologias lexicais MultiWordnet.PT e WordNet.PT e o Thesaurus Eletrônico para o Português (TEP), em desenvolvimento como parte do projeto da WordNet.BR.

Comparando o português com o nível de financiamento para a tecnologia da linguagem não só para o inglês, mas para idiomas de menor projeção global que a língua portuguesa, o apoio para a investigação computacional da língua portuguesa é ainda muito baixo. Apesar de haver subáreas muito ativas neste campo, em termos de tecnologia da linguagem, o português é um idioma bem menos equipado, especialmente quando comparado a países como os de língua inglesa, alemã ou holandesa (REHM e USZKOREIT, 2012).

³⁰ Informações em <<http://www.leonel.profusehost.net/complin.htm>>.

³¹ Informações em <<http://www.inf.pucrs.br/~linatural/>>.

³² Informações em <<http://www.nilc.icmc.usp.br/>>.

	Quantidade	Disponibilidade	Qualidade	Cobertura	Maturidade	Sustentabilidade	Adaptabilidade
Tecnologia da Linguagem: Ferramentas de Processamento e Aplicações							
Reconhecimento da Fala	2	3	4	2	2	2	4
Síntese da Fala	3	3	4	4	4	3	4
Análise Gramatical	3	3	4	4	4.5	2.5	4.5
Análise Semântica	1.5	2	3	2	2.5	2.5	2.5
Geração de Linguagem	0	0	0	0	0	0	0
Tradução Automática	3	2	2	2	4	2	2
Recursos Linguísticos: Conjuntos de Dados e Bases de Conhecimento Linguístico							
Corpora Escritos	3	3	4	4.5	4	4.5	4.5
Corpora de Fala	4	2	4	4	4	3	3
Corpora Paralelos	2	4	2	2	2	3	3
Recursos Lexicais	3.5	3	4.5	3	4	3	3
Gramáticas	1	4	5	2	2	2	2

Figura 6 Ferramentas e recursos linguísticos disponíveis para o português: de 0 (muito baixo) a 6 (muito alto).³³

De acordo com Rehm e Uszkoreit (2012), os recursos disponíveis para o português, atualmente, são poucos e há uma necessidade premente para que esforços sejam concentrados na criação de recursos linguísticos para a investigação e desenvolvimento de ferramentas e aplicações para o processamento computacional do português:

- Existem dois grandes *corpora* de texto, mas não são anotados (um jornalístico e outro parcialmente disponível, devido a restrições de direitos autorais);
- Há um *corpus* anotado, de 1 milhão de palavras, juntamente com o respetivo etiquetador morfossintático e outras ferramentas de processamento de base morfológica;
- Para a fala, há um conjunto de sistemas comerciais para as variedades europeia e brasileira do português (reconhecimento da fala, síntese da fala e gestão de diálogo). Reservados ao uso dos laboratórios de pesquisa;

³³ Fonte: REHM e USZKOREIT, 2012, p. 32.

- Não existem ainda *corpora* anotados com informação sobre semântica lexical, o que origina um preocupante entrave à investigação sobre desambiguação de palavras em português e desenvolvimento de ferramentas associadas;
- Enquanto alguns *corpora* têm anotação morfossintática, os com anotação sintática (*treebanks*) são mais raros e de tamanho muito reduzido;
- Quanto mais conhecimento linguístico e semântico uma ferramenta tomar em consideração, mais lacunas existem (ver, por exemplo, recuperação de informação vs. semântica do texto): é preciso aplicar mais esforço de Investigação e Desenvolvimento no processamento linguístico profundo, incluindo a construção de gramáticas computacionais para o português.

Em meio a todas essas dificuldades de processamento do português, este estudo conseguiu encontrar uma forma de desenvolver uma pesquisa utilizando uma ferramenta de acesso livre e gratuita especialmente criada para o processamento do português brasileiro. A ferramenta Coh-Metrix-Port, desenvolvida junto ao NILC, é uma adaptação da ferramenta Coh-Metrix para o inglês. Sobre ela trato a seguir.

5.2 COH-METRIX

Nascido do PLN, o sistema Coh-Metrix³⁴ para o inglês, que significa *cohesion metrics*, é uma ferramenta para análise unitária de textos, disponível gratuitamente on-line. Elaborada por pesquisadores da Universidade de Memphis, nos Estados Unidos (GRAESSER et al., 2004), tem como propósito calcular índices de coesão e de coerência textual num amplo espectro de medidas lexicais, sintáticas, semânticas e referenciais com o objetivo de indicar a adequação de um texto a seu público-alvo (a “demanda cognitiva” e a inteligibilidade do texto).

Para o cálculo dessas métricas coesivas e de coerência, vários recursos e ferramentas de PLN são utilizados (a *Wordnet*, por exemplo, que fornece repertórios de sinônimos). A ferramenta Coh-Metrix também tem a função de apontar dados para identificar problemas textuais de ordem estrutural. Na Figura 7 a seguir, é possível visualizar a tela inicial do Coh-Metrix:

³⁴ Disponível em <<http://cohmetrix.memphis.edu/cohmetrixpr/index.html>>. Acesso em 10 dez 2012.

Created: September 1, 2012 **Coh-Metrix 3.0** Last updated: September 6, 2012

For the best effect, use IE 5.0 or above.

Title:

Genre:

Source:

Job Code:

LSA Space:

[DataViewer](#)

Headers

1. Enter the "Title" you wish to give to your study.
2. Select the genre you feel most closely describes your work.
3. Enter the source of the document. Where did you get this text?
4. Enter a "Jobcode". You may make up your own job code. You need to remember this job code to later retrieve your results.
5. Coh-Metrix uses Latent Semantic Analysis (LSA) in some of its indices. Your text will be analyzed slightly differently depending on the space (discourse type) that you choose. Please select a LSA Space you feel most closely describes your work. If you are not sure which space to use, we recommend you select "College Level".

Entering your Text

1. You may write OR cut and paste text.
2. Please try to limit text to a maximum of 15,000 characters and remove irregular characters.
3. Paragraphs are marked by hard returns.
4. Press "Submit" and Coh-Metrix will analyze your text.

Viewing and Understanding your Results

1. When Coh-Metrix has analyzed your text the results will appear on the right side of the screen.
2. You may continue to enter and submit text on this screen. The results will continue to appear on the right side of the screen.

Viewing Past Results

To view past results, Click the "Data Viewer" link at the bottom left of the screen (next to the Submit button). You will then be directed to a new page where you can retrieve your past data.

Figura 7 Tela inicial do Coh-Metrix 3.0.

A ferramenta é bastante amigável e simples de manipular. À esquerda estão informações que o usuário deve dar sobre o texto que deseja analisar (título, gênero – escolha entre científico, narrativo ou informativo –, fonte, código que deseja dar texto no sistema e dimensão para a análise de semântica latente³⁵ – no caso, há apenas o nível universitário para escolha).

Com as mais de 600 métricas disponíveis em sua versão restrita – paga -, o Coh-Metrix dá resultados que auxiliam a tarefa do leitor, professor, corretor ou estudante de encontrar problemas textuais de ordem estrutural nos textos a ela submetidos. Infelizmente, das 600 métricas noticiadas, apenas 108 estão disponíveis na versão gratuita on-line, no *site* do projeto.

Na versão livre e simplificada do Coh-Metrix 3.0, atualizado em 2012, índices que vão desde métricas simples (como a contagem de palavras) até medidas mais complexas (como algoritmos de resolução anafórica) estão entre os resultados. As 108 métricas estão divididas em seis blocos que avaliam a complexidade de um texto a partir da mensuração dos seguintes elementos:

- Identificação geral e informação de referência, índices de inteligibilidade, palavras gerais e informação do texto, índices sintáticos, índices referenciais e semânticos e dimensões do modelo de situações. Essa primeira classe

³⁵ A análise semântica latente é uma medida bastante complexa de semelhança de conteúdo entre textos feita via medidas de semelhança entre palavras. *Semântica* aqui é um termo cujo conceito é bastante diferente do que é para um linguista.

corresponde às informações que referenciam o texto, como título, gênero entre outros.

- Índices de inteligibilidade calculados com as fórmulas *Flesch Reading Ease* e *Flesch Kincaid Grade Level*. Essas fórmulas consideram tamanho de sentença, número de palavras por sentença e número de palavras diferentes por sentença.
- Verificação de quatro subclasses: contagens básicas, frequências, concretude, hiperônimos.
- Verificação de cinco subclasses: constituintes, pronomes, tipos e *tokens*, conectivos, operadores lógicos e similaridade sintática de sentenças.
- Verificação de três subclasses: anáfora, correferência e análise de semântica latente.
- Verificação de quatro subclasses: dimensão causal, dimensão intencional, dimensão temporal e dimensão espacial.

Como já mencionado, a ferramenta Coh-Metrix foi desenhada apenas para processar o inglês. No Brasil, temos uma iniciativa para adaptar os seus índices disponíveis on-line (na época, apenas 48) da ferramenta para o português, a ferramenta Coh-Metrix-Port, sobre a qual trato a seguir.

5.3 COH-METRIX-PORT

No âmbito do Projeto PorSimples³⁶ surgiu uma iniciativa de adaptação para o português brasileiro das sessenta métricas oferecidas gratuitamente pelo Coh-Metrix na época. Essa iniciativa tinha como objetivo identificar índices oferecidos pela ferramenta que fossem capazes a complexidade textual. Com essa finalidade, o Coh-Metrix-Port foi desenvolvido para auxiliar a tarefa de simplificação de textos e facilitação do acesso à informação para analfabetos funcionais e para pessoas com deficiências cognitivas. O Coh-Metrix-Port, versão adaptada para o português, opera com 48 métricas disponíveis gratuitamente³⁷.

A adaptação do sistema para o português foi finalizada em 2010 (SCARTON e ALUÍSIO, 2010). Entretanto, nem todas as métricas entre o sistema para o inglês e o para o português brasileiro podem ser pareadas, pois há medidas próprias da gramática de cada

³⁶ Projeto que contou com a elaboração de ferramentas e pesquisas para a simplificação do português. Disponível em <<http://www.nilc.icmc.usp.br/porsimples/>>. Acesso em 03 nov 2012.

³⁷ Disponível em: <<http://www.nilc.icmc.usp.br/cohmetrixport/>>. Acesso em 10 out 2012.

língua e medidas incompatíveis devido aos recursos utilizados como referência (*wordnet do inglês*, por exemplo). As principais medidas utilizadas na ferramenta brasileira são:

- Contagens básicas: número de palavras, número de sentenças, número de parágrafos, sentenças por parágrafos, palavras por sentenças e sílabas por palavras.
- Índice Flesch.
- Constituintes: incidência de sintagmas nominais, modificadores por sintagmas nominais e palavras antes de verbos principais.
- Conectivos: incidência de todos os conectivos, incidência de conectivos aditivos positivos, incidência de conectivos aditivos negativos, incidência de conectivos temporais positivos, incidência de conectivos temporais negativos, incidência de conectivos causais positivos, incidência de conectivos causais negativos, incidência de conectivos lógicos positivos e incidência de conectivos lógicos negativos.
- Operadores lógicos: incidência de operadores lógicos e número de negações.
Pronomes, tipos e tokens: incidência de pronomes pessoais, pronomes por sintagmas nominais e relação tipo/token.
- Correferências: sobreposição do argumento em sentenças adjacentes, sobreposição de argumento, sobreposição do radical de palavras em sentenças adjacentes, sobreposição do radical de palavras, sobreposição de palavras de conteúdo em sentenças adjacentes.
- Anáforas: referência anafórica em sentenças adjacentes e referência anafórica.

PorSimples

Simplificando o Português

Coh-Metrix-Port

Aline Evers | [Sair](#)
[Página inicial](#) > [Pesquisa](#) > Resultados

Resultados

[Imprimir](#) [Visualizar Texto](#)

- **Texto**

- | | | |
|--------------------|---------------------|--------------------|
| Título | T01E05 | Título |
| Autor | Examinando05 | Autor |
| Fonte | Celpe-Brasil
T01 | Fonte |
| Data de Publicação | Avançado Superior | Data de Publicação |
| Gênero | | Gênero |

Figura 8 Cabeçalho de arquivo produzido pelo sistema Coh-Metrix-Port.

- **Operadores Lógicos**

- | | | |
|--|---------|---|
| Incidência de Operadores Lógicos | 55.5556 | Incidência de operadores lógicos em um texto. Consideramos como operadores lógicos: e, ou, se, negações e um número de condições. |
| Incidência de E | 43.2099 | Incidência do operador lógico <i>e</i> em um texto. |
| Incidência de OU | 12.3457 | Incidência do operador lógico <i>ou</i> em um texto. |
| Incidência de SE | 0 | Incidência do operador lógico <i>se</i> em um texto. |
| Incidência de Negações | 0 | Incidência de Negações. Consideramos como negações: <i>não, nem, nenhum, nenhuma, nada, nunca e jamais</i> . |

- **Frequências**

- | | | |
|------------------------------------|---------|---|
| Frequências | 294573 | Média de todas as frequências das palavras de conteúdo encontradas no texto. O valor da frequência das palavras é retirado da lista de frequências do corpus Banco do Português. |
| Mínimo Frequências | 4440.75 | Identifica-se a menor frequência dentre todas as palavras de conteúdo em cada sentença. Depois, calcula-se uma média de todas as frequências mínimas. A palavra com a menor frequência é a mais rara da sentença. |

Figura 9 Parte do arquivo contendo métricas e contagens.

Ambas as ferramentas, ainda que não tenham sido criadas com o intuito de serem usadas na análise de textos submetidos a exames de proficiência ou comparações de textos de estudantes, abrem um universo de possibilidades para os pesquisadores de Linguística Aplicada. Portanto, friso que as métricas calculadas automaticamente pela ferramenta Coh-Metrix-Port, por si só, não conseguem indicar o nível de proficiência a que corresponda um dado texto.

É na interrelação entre as métricas trazidas pelo sistema que se poderia encontrar um caminho para uma separação dos textos por níveis de proficiência. É preciso salientar que, nesta dissertação, viso *comparar* os resultados gerados pelo sistema Coh-Metrix-Port para textos mais avançados com os de textos menos avançados e analisar as *diferenças* postas entre os resultados obtidos. É também na relação estatística entre os resultados obtidos que se pode caracterizar cada conjunto de textos de acordo com as métricas que mais os discriminariam.

5.3.1 ÍNDICE FLESCH

Um item de destaque nesse sistema de medidas de complexidade (também associado a inteligibilidade ou *readability*) é o índice Flesch. No Brasil, só mais recentemente pesquisadores de Estatística e de PLN se interessaram por fórmulas e medidas de complexidade textual, adaptando-as ao português. Adaptado para o português brasileiro por pesquisadores do Instituto de Ciências Matemáticas e da Computação (ICMC) da Universidade de São Paulo (MARTINS et al., 1996), o Índice Flesch brasileiro é uma das diferentes medidas de complexidade do texto associada à sua inteligibilidade para diferentes tipos de leitores. O resultado é um número de 0 a 100 que é assim mensurado com a devida adaptação para o sistema escolar brasileiro feita pela equipe do ICMC e NILC:

- **Muito fáceis:** índice entre 90 a 100, textos adequados para leitores com nível de escolaridade até a 4ª série do Ensino Fundamental.
- **Fáceis:** índice entre 80 a 89, textos adequados a estudantes com escolaridade até a 8ª série do ensino fundamental.
- **Razoavelmente fáceis:** índice entre 70 a 79, textos adequados a estudantes com escolaridade até a 8ª série do Ensino Fundamental.
- **Padrão:** índice entre 60 e 69, textos adequados a estudantes com escolaridade até a 8ª série do Ensino Fundamental.

- **Razoavelmente difíceis:** índice entre 50 a 59, textos adequados para estudantes cursando o Ensino Médio ou universitário.
- **Difíceis:** índice entre 30 a 49, textos adequados para leitores com Ensino Médio ou universitário.
- **Muitos difíceis:** índice entre 0 a 29, textos adequados apenas para áreas acadêmicas específicas.

Um elemento importante indicado pelas pesquisas que se ocupam do Índice Flesch e do comportamento de leitores foi a perda de interesse do leitor quando o texto se mostra muito complexo. Esse aspecto também foi corroborado por pesquisas posteriores (por exemplo, MCNAMARA et al., 2002; GRAESSER et al., 2004).

5.3.2 TYPE-TOKEN RATIO (TTR) OU MEDIDA DE RIQUEZA LEXICAL

Medidas de riqueza lexical – aquelas que, em geral, buscam apontar se textos orais ou escritos possuem vocabulário mais diversificado ou repetitivo – têm sido utilizadas para avaliar o nível de proficiência lexical de crianças e adultos já há bastante tempo, com estudos que comparam trechos de fala ou textos recentes com trechos e textos de referência externos ou mais antigos. Essas medidas dividem-se entre aquelas que mensuraram a densidade lexical (ou seja, a quantidade de palavras de conteúdo em um dado texto dividida pelo total de palavras desse mesmo texto) e que mensuram a sofisticação lexical (medidas que apontam a proporção de itens lexicais presentes e ausentes em uma lista de frequência, e que acabam por indicar, através da análise da frequência, que quanto mais difícil é uma palavra, menos frequente ela será em um dado *corpus*).

A medida de riqueza lexical mais utilizada por pesquisadores, em LC e no PLN, é a já bastante conhecida e não menos controversa *type-token ratio* (TTR). Essa medida é calculada de forma bastante simples: divide-se o número de palavras diferentes de um texto (os *types* ou tipos) pelo número total de palavras desse mesmo texto (os *tokens*). Assim, se temos um texto com, por exemplo, 87 palavras no total, ou seja, 87 *tokens*, e descobrimos, através de uma análise muito rápida usando ferramentas como o AntConc, que 62 palavras nele se repetem (os *types*), então, para calcular a TTR, devemos dividir *types* por *tokens*, ou seja, 62/87. Ao multiplicarmos o resultado por 100, para trabalharmos com uma porcentagem e compreendermos melhor esse número, chegaremos ao resultado de 71,3%, podendo concluir,

portanto, que esse texto, de apenas 87 palavras, apresenta uma grande variação vocabular, visto que mais de 70% dele é composto por palavras que não se repetem.

É evidente, no entanto, que essa porcentagem varia de acordo com o tamanho do texto, como já foi levantado por uma série de autores, Linguistas e Matemáticos. Dessa forma, é sabido que quanto mais longo for um texto, menor será a porcentagem resultante do cálculo de TTR. Esse problema é bem conhecido, e algumas alternativas já foram propostas tanto no campo da Linguística Aplicada quanto no da Linguística Matemática. Na Matemática, existe uma área de pesquisa especialmente dedicada a esse assunto, que estuda modelos de frequência de distribuição de palavras e da qual podemos nos beneficiar para futuras análises e pesquisas. Em torno desse tipo de discussão destaca-se Baayen (2001), que já confirmou que qualquer transformação na mensuração de TTR depende e vai sempre depender do tamanho dos textos.

Baayen, ao saber desse problema de distribuição, propõe que comecemos a análise sempre a partir de um *espectro de frequência* lexical, procurando, primeiro, classificar as palavras de um texto de acordo com sua frequência de ocorrência (por exemplo, verificar quais palavras ocorrem uma, duas, quatro vezes naquele texto, e assim por diante). Tendo o que ele chama de *espectro de frequência* em mãos, um modelo de distribuição de palavras pode ser criado, utilizando um ou mais parâmetros para descrever a forma de distribuição.

5.4 APRENDIZADO DE MÁQUINA

Dentro da Inteligência Artificial está o Aprendizado de Máquina (AM), que em PLN é geralmente utilizado em tarefas de mineração de informação. O AM dedica-se ao desenvolvimento de algoritmos e técnicas que permitam que o computador “aprenda” padrões, ou seja, que permitam que o computador melhore seu desempenho ao realizar uma dada tarefa à medida que a realiza. Algumas partes do AM estão intimamente ligadas à mineração de dados e à estatística.

Uma tarefa de mineração de dados recorrente em AM, dado que envolve identificação de padrões de atributos de dados associados entre si, é a Classificação. Observe a Figura 10 a seguir:

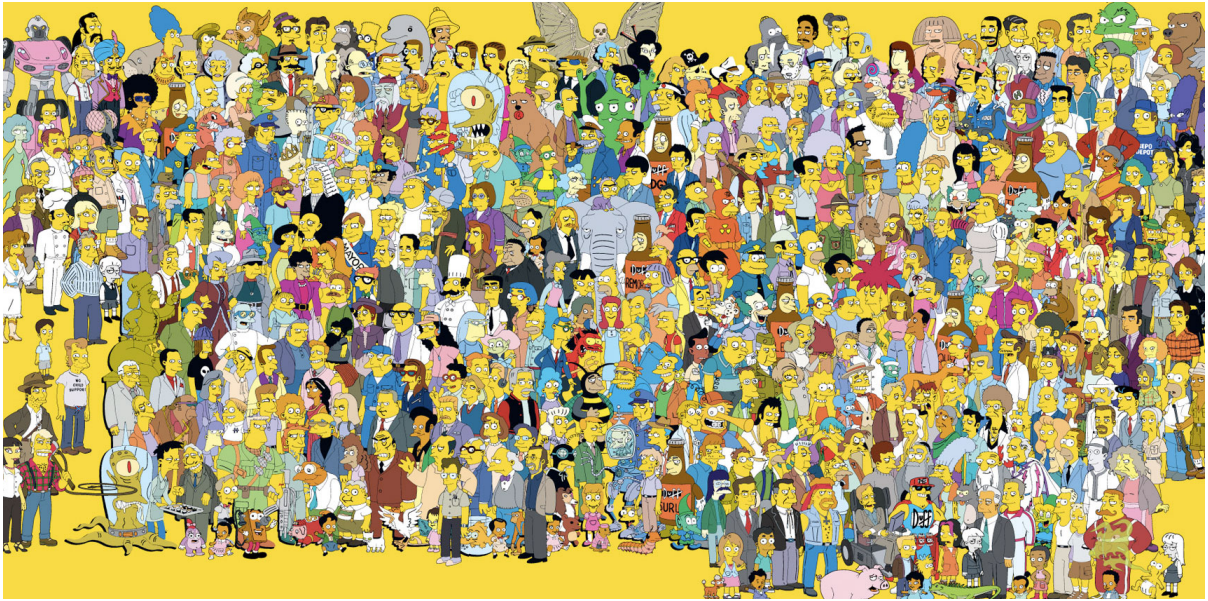


Figura 10 Todos os personagens do desenho animado Os Simpsons.

A tarefa de classificação de objetos, entidades, textos, pode ser feita de diferentes modos e a partir de diferentes critérios. Por exemplo, os textos de um jornal podem ser classificados entre reportagens e notícias ou, simplesmente, entre textos longos e textos curtos considerando-se uma medida X em número de palavras, números de caracteres, número de frases, entre outras medidas. Conforme se vê na Figura 11, é possível separar um conjunto de indivíduos utilizando diferentes critérios. As possibilidades podem ser infinitas, chegando, por exemplo, ao tipo de calçado que uma pessoa usa. Seria possível separar os personagens de *Os Simpsons* de diversas maneiras, apenas escolhendo algum critério ou fazendo uma pergunta. Por exemplo:

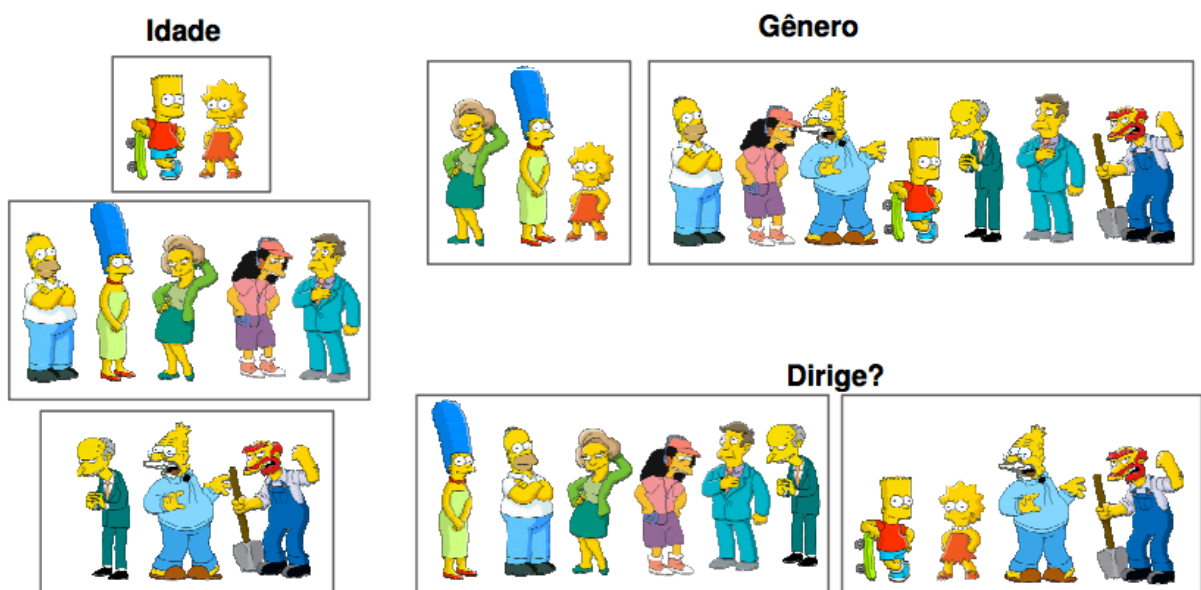


Figura 11 Personagens de Os Simpsons classificados em diferentes grupos.

Separar os elementos de um conjunto é uma tarefa de AM, assim como também pode ser a predição de níveis de proficiência em LA de um conjunto de redações X.

O AM existe há mais de 50 anos e Witten e Frank (2005) o definem como a área de estudos de sistemas automáticos de aquisição e de integração de conhecimento. Dito de modo muito simplificado, o AM ocorre quando um determinado programa aprende a executar uma tarefa de classificação de um dado X (ou grupo de dados Y) a partir de um modelo de classificação previamente sistematizado. Assim, podemos imaginar que um sistema desse tipo seria capaz de prever que há uma forte correlação entre os atributos “ser careca” e “ser personagem do sexo masculino” no conjunto dos personagens de *Os Simpsons*, e indicar a mesma correlação para um outro conjunto de dados. O seu desempenho ao executar tal tarefa aumenta com a experiência, ou seja, o aprendizado ocorre quando é possível verificar que o programa toma decisões melhoradas e acerta mais nas categorizações de dados tendo por base problemas que foram resolvidos previamente. Com essa lógica, por exemplo, pré-diagnósticos podem ser oferecidos ao médico com base em uma categorização de sintomas que ele arrole como verificados em um dado paciente. Para os linguistas, a principal contribuição do AM está na categorização automática de dados que são palavras, sintagmas, frases ou textos ou *corpora*.

5.4.1 WEKA

O método automático de classificação usado nesta pesquisa baseia-se no modelo estatístico supervisionado de AM, e a ferramenta usada é o Weka (Waikato Environment for Knowledge Analysis³⁸). O Weka é uma coleção de algoritmos de AM que contém ferramentas para pré-processamento, classificação, regressão, agrupamento e associação de dados. O Weka opera a partir de arquivos em extensão ARFF (Attribute-Relation File Format), um arquivo de texto em ASCII (American Standard Code for Information Interchange) que descreve uma lista de instâncias que compartilham um conjunto de atributos.

O processo de aprendizado supervisionado é caracterizado pela apresentação de dados de treinamento a um algoritmo de aprendizado, o indutor. Cada exemplo possui uma classe associada. Há também o conceito de atributo, que é uma característica ou uma informação que visa descrever o exemplo, o qual pode ter ou não um rótulo associado. Esse rótulo é a classe do exemplo e representa um atributo especial que descreve uma instância do fenômeno de

³⁸ O *software* é gratuito e está disponível, com toda a documentação, no *site* <<http://www.cs.waikato.ac.nz/ml/weka/>>.

interesse, que é o conceito que se deseja induzir em tarefas de classificação (MARTINS, 2003).

Arquivos com extensão ARFF têm duas seções distintas. A primeira é o cabeçalho (*header*), que contém o nome da relação a ser analisada, a lista de atributos e o tipo de atributo (se é numérico, nominal ou sequência de caracteres [*string*]); a segunda seção é composta pelos dados (*data*), com os valores de cada atributo listado.

O algoritmo escolhido foi o J48, que é uma implementação Java (C4.5) em C++. para construção de Árvores de Decisão. A Árvore de Decisão mostra quais são as relações discriminativas entre os atributos (no caso, as métricas do Coh-Matrix-Port) do total das instâncias (cada um dos textos analisados) em cada classe (ou seja, classes Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior). A estrutura em árvore de decisão mostra visualmente quais são as métricas estatisticamente mais características e distintivas em cada bloco de textos estudado, de acordo com a natureza do texto (a classe, na terminologia de AM). A Figura 12 apresenta um exemplo de arquivo com extensão ARFF (trecho), e a Figura 13 mostra a interface principal do Weka.

```
@relation "
@attribute texto STRING
@attribute numero_palavras numeric
@attribute numero_sentencas numeric
@attribute numero_paragrafos numeric
@attribute numero_verbos numeric
@attribute numero_substantivos numeric
@attribute numero_adjetivos numeric
@attribute numero_adverbios numeric
@attribute numero_pronomes numeric
@attribute palavras_por_sentencas numeric
@attribute sentencas_por_paragrafos numeric
@attribute silabas_por_palavras numeric
@attribute flesch numeric
[...]
@attribute anafora_refanaadj numeric
@attribute anafora_refana numeric
@attribute
class
{avancado,intermediario,basico,avancado_superior,intermediario_superior,iniciante}

@data
T01E09.txt,136.0,9.0,3.0,161.765,330.882,58.8235,66.1765,66.1765,15.1111,3.0,2.71429,48.
7458,617.647,352.941,212231.0,5798.0,0.4,257.353,0.628571,4.33333,0.0,0.666667,1
.89076,14.7059,14.7059,7.35294,7.35294,44.1176,66.1765,29.4118,14.7059,0.0,0.0,2
9.4118,0.0,29.4118,0.0,4.18182,1.2,0.0,0.625,0.75,0.916667,0.875,0.861111,0.875,0.2
22222,0.222222,avancado
```

Figura 12 Exemplo de arquivo ARFF.

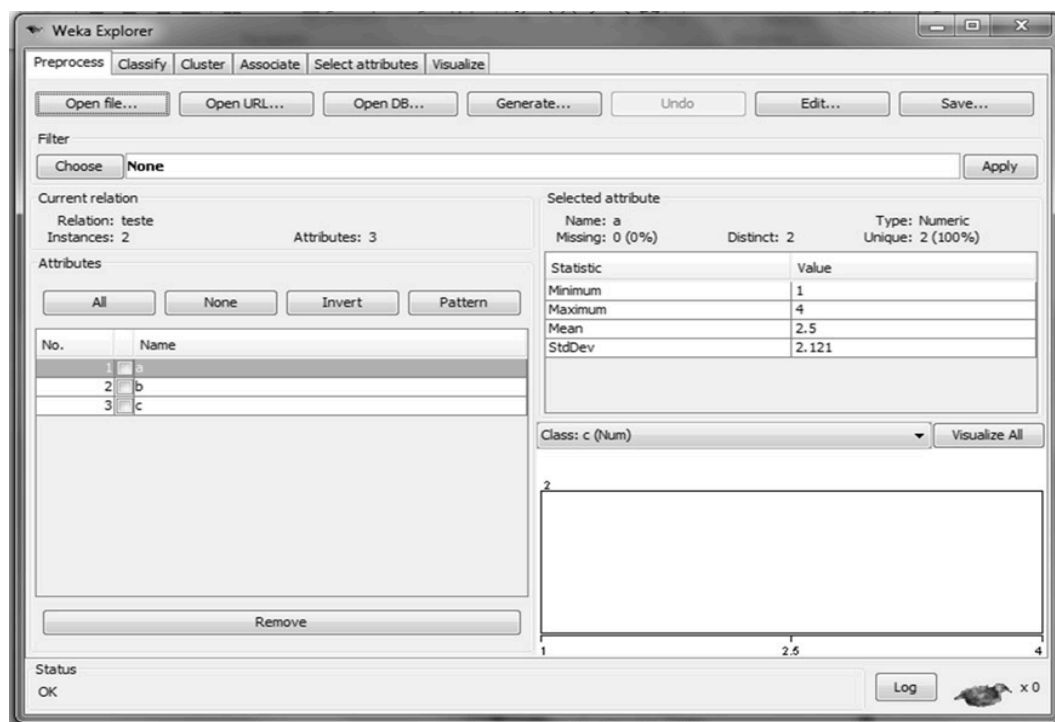


Figura 13 Interface do Weka.

5.4.2 ÁRVORE DE DECISÃO, MATRIZ DE CONFUSÃO E AVALIAÇÃO DE AM

Existem diversos modelos de classificação, de modo que a associação entre as classes e o conjunto de atributos que caracterizam um conjunto de textos a serem classificados pode se dar de formas variadas, empregando processamento simbólico e/ou numérico. A construção dos modelos computacionais de classificação geralmente emprega um dentre dois paradigmas alternativos:

- Top-down: obtenção do modelo de classificação a partir de informações fornecidas por especialistas;
- Bottom-up: obtenção do modelo de classificação pela identificação de relacionamentos entre variáveis dependentes e independentes em bases de dados rotuladas – aqui a classificação parte dos próprios dados.

O classificador é induzido por mecanismos de generalização fundamentados em exemplos específicos (conjunto finito de objetos rotulados). As Árvores de Decisão, utilizadas neste trabalho, estão fundamentadas no paradigma *bottom-up* e sua aplicação requer as seguintes condições:

- Toda informação sobre cada objeto (texto) a ser classificado deve poder ser expressa em termos de uma coleção fixa de propriedades ou atributos.

- O número de classes pode ser definido *a priori*, o que transforma a modelagem num processo de treinamento supervisionado.

O aprendizado indutivo de Árvores de Decisão é geralmente dividido em aprendizado supervisionado e não-supervisionado. Uma tarefa de classificação bastante conhecida é o diagnóstico médico, em que para cada paciente são definidos atributos contínuos ou categóricos ordinais (Exemplo: idade, altura, peso, temperatura do corpo, batimento cardíaco, pressão, etc.) e atributos categóricos não-ordinais (Exemplo: sexo, cor da pele, local da dor, etc.). A tarefa do classificador é realizar um mapeamento dos atributos para um diagnóstico (Exemplo: saudável, pneumonia, Influenza A, etc.).

Na Figura 14, é ilustrada uma Árvore de Decisão:

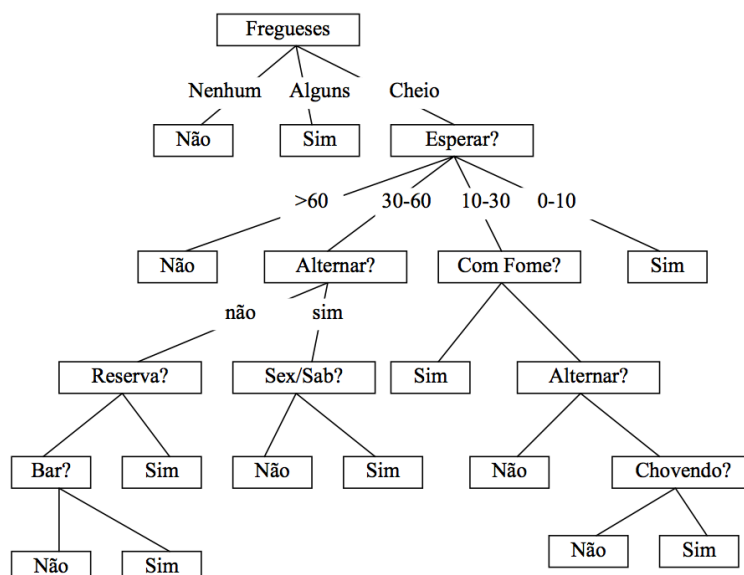


Figura 14 Árvore de Decisão para o problema de espera para jantar em um restaurante.³⁹

Muitas vezes, as Árvores de Decisão não apresentam o resultado esperado, ou seja, classificam de forma incorreta os elementos em questão. No campo do AM, para corrigir isso, é utilizada uma Matriz de Confusão, que é uma tabela específica cuja disposição de informações permite a visualização do desempenho de um determinado algoritmo na tarefa de classificação. Cada coluna dessa matriz representa uma instância de uma dada classe, enquanto cada linha dessa matriz representa as instâncias de uma classe verdadeira. O nome Matriz de Confusão vem do fato de que essa matriz torna fácil de verificar se o sistema está confundindo duas classes (ou seja, colocando elementos nas classes erradas).

³⁹ Exemplo retirado de <http://ngdweb.paginas.ufsc.br/files/2012/04/mn_DecisionTrees.pdf>. Acesso em 10 jan 2013.

Se um sistema de classificação foi treinado para diferenciar frutas, por exemplo, a Matriz de Confusão irá resumir os resultados da classificação do algoritmo a fim de realizar uma inspeção mais cuidadosa. Por exemplo, se tivermos 27 frutas – 8 morangos, 6 amoras e 13 bananas –, um possível resultado da classificação pode ser expresso pela Matriz de Confusão da Tabela 6 a seguir:

		Predição de Classe		
		Morango	Amora	Banana
Classe Verdadeira	Morango	5	3	0
	Amora	2	3	1
	Banana	0	2	11

Tabela 6 Exemplo de Matriz de Confusão.

Nessa Matriz de Confusão, vê-se que, dos 8 morangos verdadeiros, o sistema identificou e classificou corretamente 5. Das 6 amoras, o sistema identificou e classificou 3, classificando 2 como morangos e uma como banana. Pode-se ver claramente que esse sistema teve problemas para diferenciar morangos de amoras, mas executou bem a tarefa de diferenciar bananas dos outros tipos de frutas. Todas as respostas corretas estão localizadas na diagonal da tabela, de modo que fica bastante fácil encontrar os erros de classificação visualizando a tabela, já que estarão representados por qualquer valor que não seja zero e que esteja fora da diagonal.

Além da Árvore de Decisão e da Matriz de Confusão, existem três conceitos que também são importantes para compreender e avaliar o AM. São eles *recall*, *precision* e *f-measure*. Esses três conceitos dizem respeito à avaliação de desempenho do AM, ou seja, conhecendo seus valores, podemos saber o quanto o AM foi bem-sucedido.

A medida de desempenho chamada *recall*, ou cobertura, diz respeito ao número de acertos que o sistema faz computando todos os processos, inclusive os erros cometidos. A medida *precision*, ou precisão, diz respeito àquilo que o AM classificou de maneira correta, no caso deste trabalho, por exemplo, quantos textos foram avaliados como Básicos pelos corretores humanos e quantos textos o AM conseguiu classificar, também, como Básicos. E a medida *f-measure* é uma média entre essas duas outras medidas de desempenho.

6 POSICIONAMENTO DESTE TRABALHO

Retomando o que foi dito na Introdução deste trabalho, como professora de PLA e linguista de *corpus*, percebi que algo faltava de modo a aproximar as pesquisas realizadas com *corpora* e o PLA. Na tentativa de unir essas duas áreas, ao executar a pesquisa, acabei acrescentando ainda uma terceira: o PLN. Em virtude dessa multiplicidade de olhares, delimitar o escopo teórico-metodológico desta dissertação foi bastante difícil.

A delimitação ficou mais clara após decidir qual seria o *corpus* de estudo: as produções textuais do exame Celpe-Bras de 2006-1. O exame foi escolhido pelos impactos e efeitos retroativos que provoca não só na vida dos examinandos, mas em sala de aula e na formação de professores⁴⁰. Sendo um exame de larga escala, conforme apontam Diniz e Zoppi-Fontana (2006), para além de um instrumento de avaliação, o Celpe-Bras funciona também como um instrumento de política linguística e hoje tem o potencial de provocar mudanças de métodos, materiais didáticos e currículos utilizados em instituições de ensino, como bem ressaltam, também, Schlatter, Garcez e Scaramucci (2004).

Além disso, tomar conhecimento de que a parte escrita do exame é atualmente corrigida em um evento presencial, em que os avaliadores, através de um sistema de computador, procedem à leitura da prova escaneada do examinando e à marcação de sua nota. Ter um sistema que fornecesse algum apoio de análise linguística automatizada para subsidiar a correção poderia ser interessante para ajudar os corretores ao avaliarem os textos.

E, para além do escopo do exame, hoje é sabido que professores dispõem de poucos recursos que os auxiliem na tarefa de caracterizar linguisticamente os perfis das comunidades a que se dirigem, elaborar materiais didáticos e dar *feedback* aos seus estudantes. Entendo que conhecer quais formulações textuais seriam mais recorrentes nessas comunidades, quais estruturas linguísticas ofereceria mais dificuldade e como abordar essas dificuldades auxiliaria esses professores em sala de aula. Esses professores, portanto, carecem de

⁴⁰ Pude comprovar e testemunhar essa afirmação em dezembro de 2012 em Brasília, quando participei do Evento de Correção do Exame Celpe-Bras, que teve duração de sete dias, em que recebi treinamento e pude conhecer pessoalmente pesquisadores de PLA de todo o Brasil, elaboradores e coordenadores do exame. Pude, também, vivenciar as dificuldades de um corretor inexperiente que participa pela primeira vez da correção e observar ainda outras dificuldades inerentes ao processo de avaliação de desempenho.

ferramentas que os ajudem a analisar textos de estudantes. Nesse sentido, tratar de sistemas que contemplem medidas de inteligibilidade e aplicá-los à avaliação de proficiência, ainda que haja uma série de limitações a superar, poderia representar uma importante contribuição para essa ponderação.

Os estudos com *corpora* de aprendizes, feitos em Linguística de Corpus, beneficiados pelo crescente aparato tecnológico moderno do Processamento de Língua Natural, possibilitam essa abordagem. A partir de uma visão empírica e probabilística de língua, visão que é marca registrada da LC, a conjugação com o PLN permite observar o uso da linguagem e suas frequências distintas de acordo com uma multiplicidade de fatores muito grande e variada. Um ferramental gerado pela ótica do PLN (como é o caso do Coh-Matrix-Port), o qual não foi construído com a finalidade de avaliar ou classificar a proficiência de textos de estudantes de português, foi aqui usado para este fim. Foi através do contato e da troca com pesquisadores de PLN que a abordagem por técnicas de Aprendizado de Máquina foi introduzida neste trabalho, contribuição fundamental a esta dissertação, visto que nos empresta uma acurácia matemática e estatística, antes impensáveis para uma pessoa graduada em Letras, como eu. Essa acurácia, naturalmente, deve ser relativizada quando se tem em mente que o *corpus* reunido tem um tamanho bem pequeno. Entretanto, faço aqui uma aposta no bom potencial dessas técnicas.

Não pretendo, com este trabalho, estabelecer novos paradigmas para a avaliação de proficiência em língua portuguesa. Pretendo, sim, a partir de um recorte que incide apenas sobre o que está concretamente posto de modo explícito em um texto ou conjunto de textos, empreender uma comparação. Verificar, a partir dela, em que medida métricas lexicais e coesivas podem auxiliar a classificação semi-automática de níveis de proficiência do português.

Assim, em síntese, este trabalho se posiciona em meio aos trabalhos de Linguística de Corpus, acreditando na validade de seus princípios, métodos e crenças para investigar usos da língua. Entretanto, acredito também na validade de se ter em mente um objeto TEXTO, entendido como um todo de significação e de sentido, alinhando-nos com a Linguística do Texto. Esse é o objeto que faz o *corpus* e que deve não ser sublimado por ele; o texto, na sua individualidade, não deve ser subsumido em meio a um grande *corpus*. O processamento texto a texto, respeitando condições de gênero, prática hoje em voga em PLN, visto que tem máquinas e recursos capazes de cruzar um todo gigante e muitas de suas partes, nos faz relembrar isso.

PARTE III – PROCEDIMENTOS, RESULTADOS E CONCLUSÕES

Após a apresentação das teorias e metodologias que estão presentes neste trabalho, nesta parte, apresento os procedimentos realizados para a análise, os resultados obtidos e a discussão desses resultados. A seguir, as questões e hipóteses de pesquisa são retomadas e respondidas. Por fim, aponto os limites do trabalho, as perspectivas para trabalhos futuros e as considerações finais com relação à pesquisa aqui empreendida.

7 PROCEDIMENTOS

Partindo da microperspectiva estrutural do texto, isto é, considerando **apenas a tessitura coesiva e o perfil lexical dos textos** submetidos ao exame de proficiência Celpe-Bras, a pesquisa empreendida aqui é um estudo quantitativo e qualitativo sobre possíveis métricas úteis para a estimação ou pré-avaliação de proficiência escrita em português. Assim, dito de um modo muito resumido, o sistema Coh-Metrix-Port foi alimentado com 177 textos e observou-se, com apoio do AM e do sistema Weka, o quanto o seu *output* mostra correlações relevantes entre propriedades dos textos e as suas avaliações recebidas pelos corretores do Celpe-Bras.

O *corpus* possui 177 produções textuais, que foram separadas em seis classes: 4 estão na classe Iniciante, 30 estão na classe Básico, 49 estão na classe Intermediário, 43 estão na classe Intermediário Superior, 29 estão na classe Avançado e 22 estão na classe Avançado Superior.

Em trabalhos cuja metodologia segue os preceitos da LC, via de regra, os textos são todos agrupados em um único arquivo e, a partir desse *corpus*, são feitas as análises lexicais, gramaticais e semânticas. Neste trabalho, essas análises foram feitas de duas formas: a primeira, por agrupamentos, ou seja, o comportamento lexical das classes foi analisado formando um *subcorpus* de cada classe e, a partir desses arquivos, foram feitas as análises; e a segunda, tratando dos textos individualmente, agrupando os resultados que a ferramenta Coh-Metrix-Port forneceu nas classes, de modo a observar regularidades e perfis coesivos de cada uma delas.

É importante frisar, conforme mencionamos nas seções 5.2 e 5.3, que as 48 métricas adaptadas do Coh-Metrix para o Coh-Metrix-Port utilizam recursos linguísticos significativamente diferentes entre as duas línguas (português e inglês)⁴¹. Dessa forma, ao compararmos os resultados coesivos deste estudo com os do estudo realizado por Crossley e McNamara (2012), é interessante considerar que muitas das métricas utilizadas pelos autores – 600 em vez de 48 métricas –, ainda não foram adaptadas para a língua portuguesa. Isso quer

⁴¹ Outro exemplo disso seria a *Wordnet*, que é uma base de dados lexicais da língua inglesa.

dizer que boa parte das métricas disponíveis em inglês não existem em português e que, portanto, tivemos de trabalhar apenas com o que já foi implementado, faltando uma série de itens que seriam fundamentais para uma avaliação mais completa, acurada e aprofundada. Para exemplificar o que queremos dizer com “diferenças”, a métrica “incidência de palavras de conteúdo”, por exemplo, usa um banco de dados de concretude de palavras⁴² de conteúdo indisponível em português; por sua vez, as métricas de frequências fazem uso do Banco de Português⁴³, e assim por diante.

Tendo em vista o número pequeno de métricas disponíveis em comparação às utilizadas no estudo de Crossley e McNamara (2012), num primeiro momento, fiz uso de absolutamente todas as categorias de análise (lexicais, sintáticas e semânticas, tendo em vista que a categoria de medidas do tipo referencial ainda está em construção), a fim de verificar, posteriormente, o que seria mais ou menos produtivo para distinção dos níveis de proficiência.

Tendo essas categorias em mãos, a ferramenta Weka foi utilizada para apurar as relações discriminativas entre as métricas – a lista de atributos é composta pelas 48 métricas do Coh-Metrix-Port – e as classes previamente avaliadas. O algoritmo escolhido foi a implementação J48 de classificação C4.5 para construção de árvores de decisão (explicadas no Capítulo 5). Alguns exemplos dos atributos são o Índice Flesch, pronomes por sintagmas, número de palavras de conteúdo e referência anafórica adjacente.

O algoritmo escolhido construiu uma árvore de decisão, que mostrou quais relações eram discriminativas entre os atributos (no caso, as métricas do Coh-Metrix-Port) do total das instâncias (cada um dos textos analisados) em cada classe (ou seja, classe de textos que vão de Iniciante a Avançado Superior). Utilizando as seis classes, o desempenho do algoritmo foi considerado baixo, o que levou a tentativa de uma nova classificação, baseada em duas classes: textos SEM (agrupamento das classes Iniciante e Básico) e COM CERTIFICAÇÃO (agrupamento das classes Intermediário, Intermediário Superior, Avançado e Avançado Superior). Para este experimento, a árvore obteve precisão de 70% nas classificações.

As Figuras 15 e 16 a seguir representam as etapas descritas acima:

⁴² Base de dados MRC Psycholinguistic Database. Disponível em <http://www.psych.rl.ac.uk/MRC_Psych_Db.html> Acesso em fev 2013.

⁴³ Organizado por Tony Berber Sardinha. Disponível em <<http://www2.lael.pucsp.br/corpora/bp>> Acesso em abr 2012.

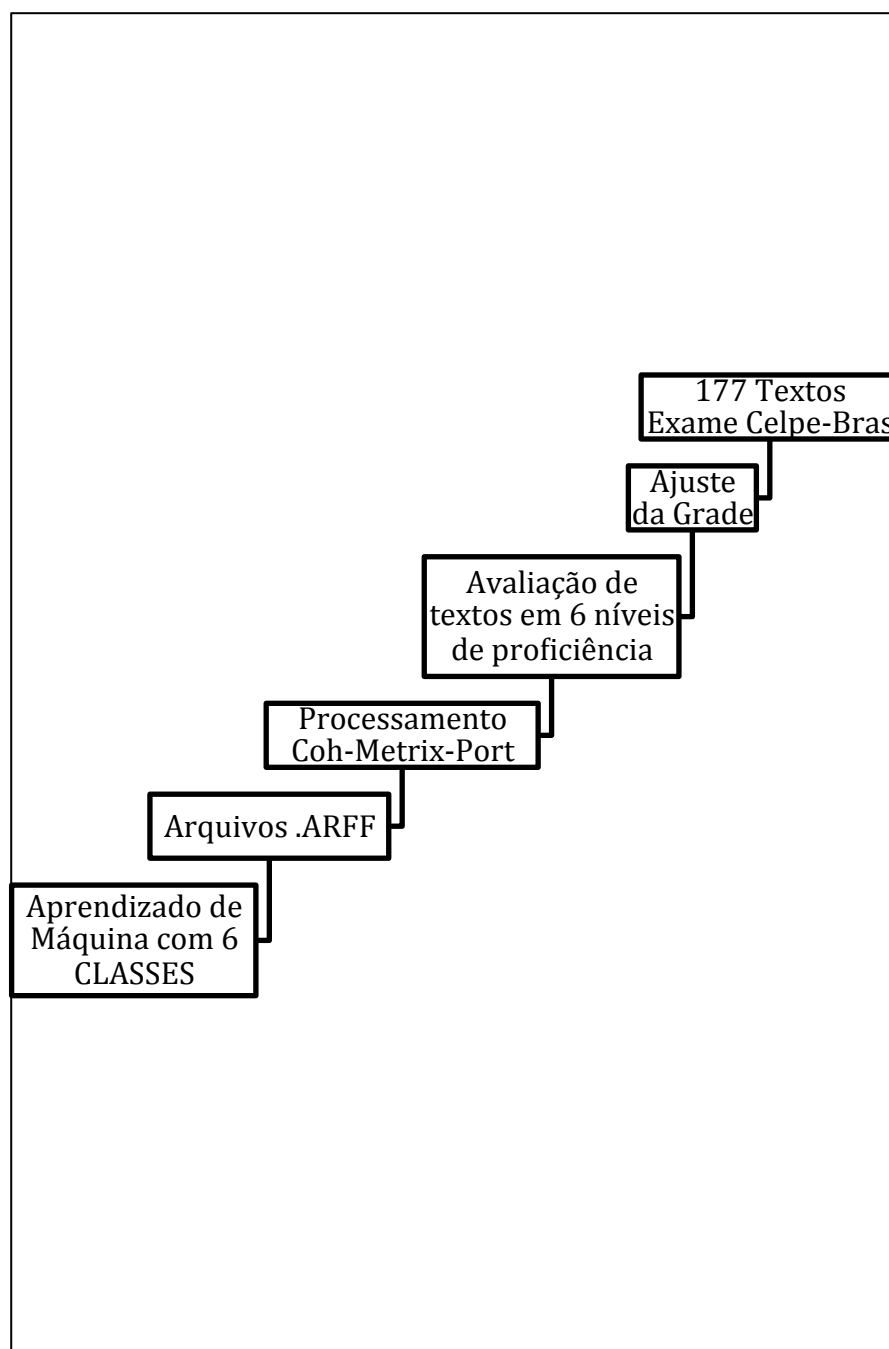


Figura 15 Etapas de processamento do *corpus* e 1ª Etapa de AM.

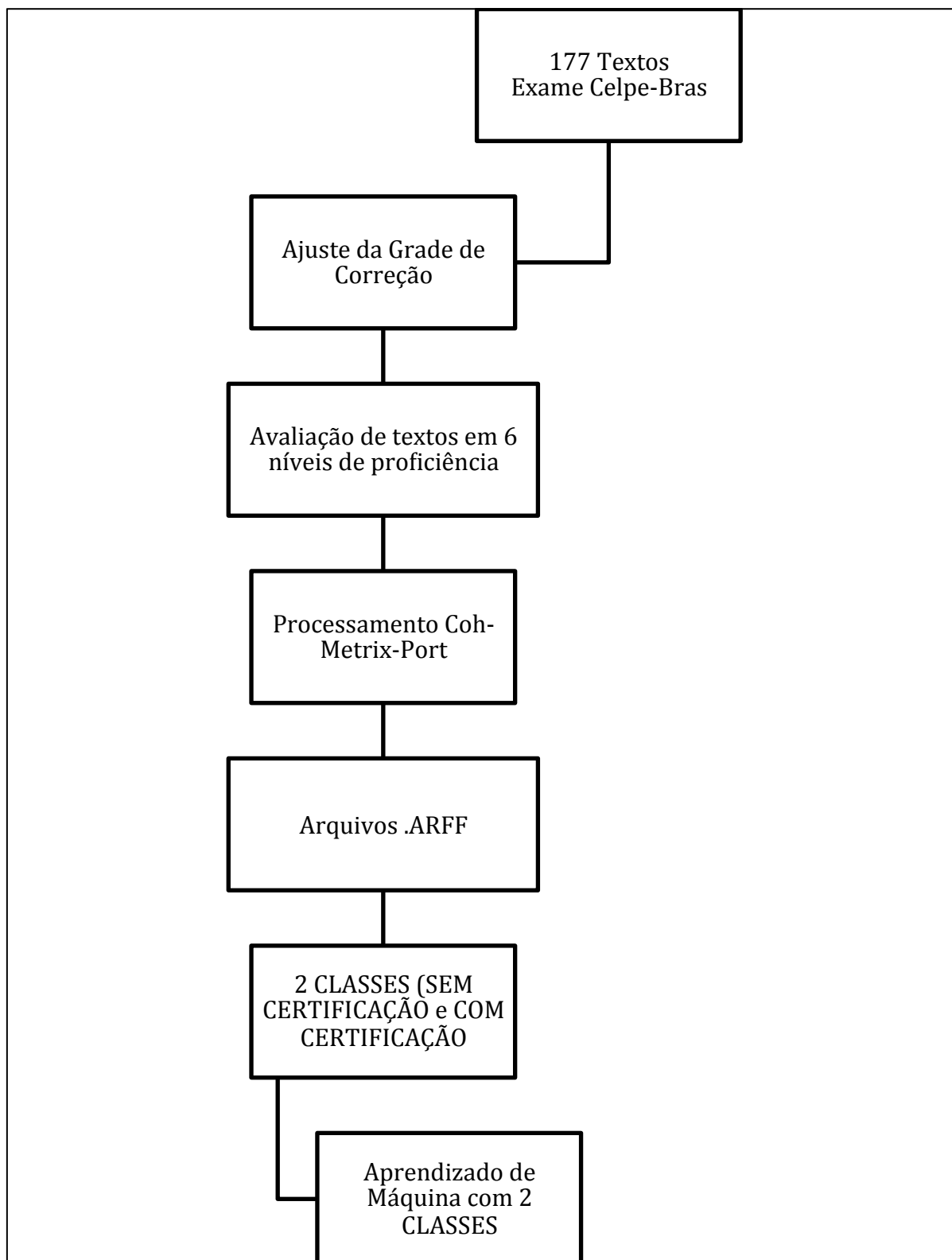


Figura 16 Etapas de processamento do *corpus* e 2ª Etapa de AM.

8 RESULTADOS E DESCRIÇÃO DOS DADOS OBTIDOS

Nesta seção, descrevo e comento os resultados obtidos após o processamento dos textos submetidos ao exame Celpe-Bras nas ferramentas Coh-Metrix-Port. As descrições foram subdivididas em duas seções: a primeira apresenta os resultados que chamamos de **resultados lexicais** – relacionados ao uso das palavras, tais como número de palavras e número de palavras distintas (TTR) e intersecção de vocabulário entre as diferentes classes –; a segunda apresenta os **resultados coesivos**: métricas número de palavras, sentenças por parágrafos, sílabas por palavras, número de palavras funcionais, hiperônimos de verbos, número de operadores lógicos, ambiguidade de substantivos, palavras antes de verbos, número de pronomes, Índice Flesch, incidência de pronomes pessoais, conectivos lógicos positivos, ambiguidade de adjetivos, tipo/tokens, pronomes por sintagmas.

Em seguida, na parte final desta seção, estão os resultados obtidos a partir da classificação das métricas por AM com a ferramenta Weka. Mostram-se os resultados obtidos com a Árvore de Decisão, a Avaliação e os problemas e soluções de classificação encontrados.

8.1 RESULTADOS LEXICAIS

Considerando a ocorrência de palavras por nível de proficiência (classes), observa-se um uso crescente do número de palavras em relação ao nível de proficiência que etiqueta os conjuntos de textos. Tal constatação pode ser observada na Tabela 7 a seguir:

Classe	Tarefa 1	Tarefa 2	Tarefa 3	Tarefa 4	Total
Iniciante (6)	295	0	50	146	491
Básico (5)	999	1023	1191	955	4168
Intermediário (4)	1065	1186	4095	675	7021
Intermediário Superior (3)	1474	1994	1105	1972	6445
Avançado (2)	685	2074	750	1546	5055
Avançado Superior (1)	1343	1209	425	1066	4043

Tabela 7 Número TOTAL de palavras por Tarefa e Nível.

Ao observarmos a Tabela 7, considerando os níveis de proficiência (Classes), vemos que há uma concentração maior de palavras nos níveis Intermediários (4 e 3), uma presença

menor de palavras no nível Iniciante (6) e uma presença de número de palavras aproximado entre os níveis Básico (5) e Avançado Superior (3). Para efeitos de visualização, o Gráfico 3 a seguir mostra os dados dessa mesma tabela por nível e por tarefa:

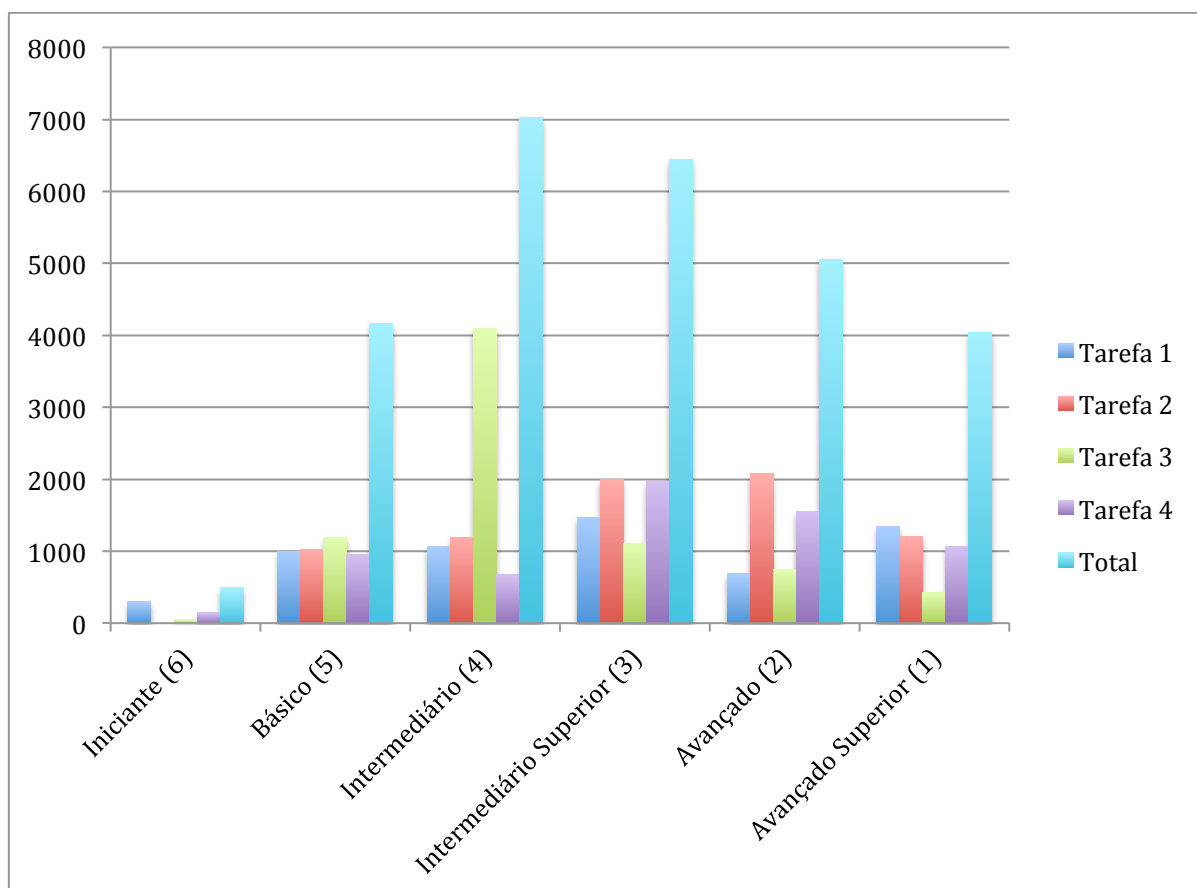


Gráfico 3 Número TOTAL de palavras por Tarefa e Nível.

Já a Tabela 8, a seguir, apresenta o número de palavras distintas dos textos por Nível e por Tarefa. Pode-se observar, novamente, que a maior concentração de uso de palavras diferentes no conjunto de textos encontra-se nos níveis Intermediários (4 e 3), apresentando uma grande queda no nível Iniciante (6) e sendo relativamente menor no nível Avançado Superior (1):

Classe	Tarefa 1	Tarefa 2	Tarefa 3	Tarefa 4	Total
Iniciante (6)	212	0	42	99	353
Básico (5)	707	710	898	661	2652
Intermediário (4)	739	834	2841	525	4547
Intermediário Superior (3)	1060	1358	778	1384	4184
Avançado (2)	481	1367	534	1084	3124
Avançado Superior (1)	894	784	288	747	2343

Tabela 8 Número de palavras DISTINTAS por Tarefa e Nível.

Para efeitos de visualização, o Gráfico 4 a seguir mostra a representação desses números em colunas:

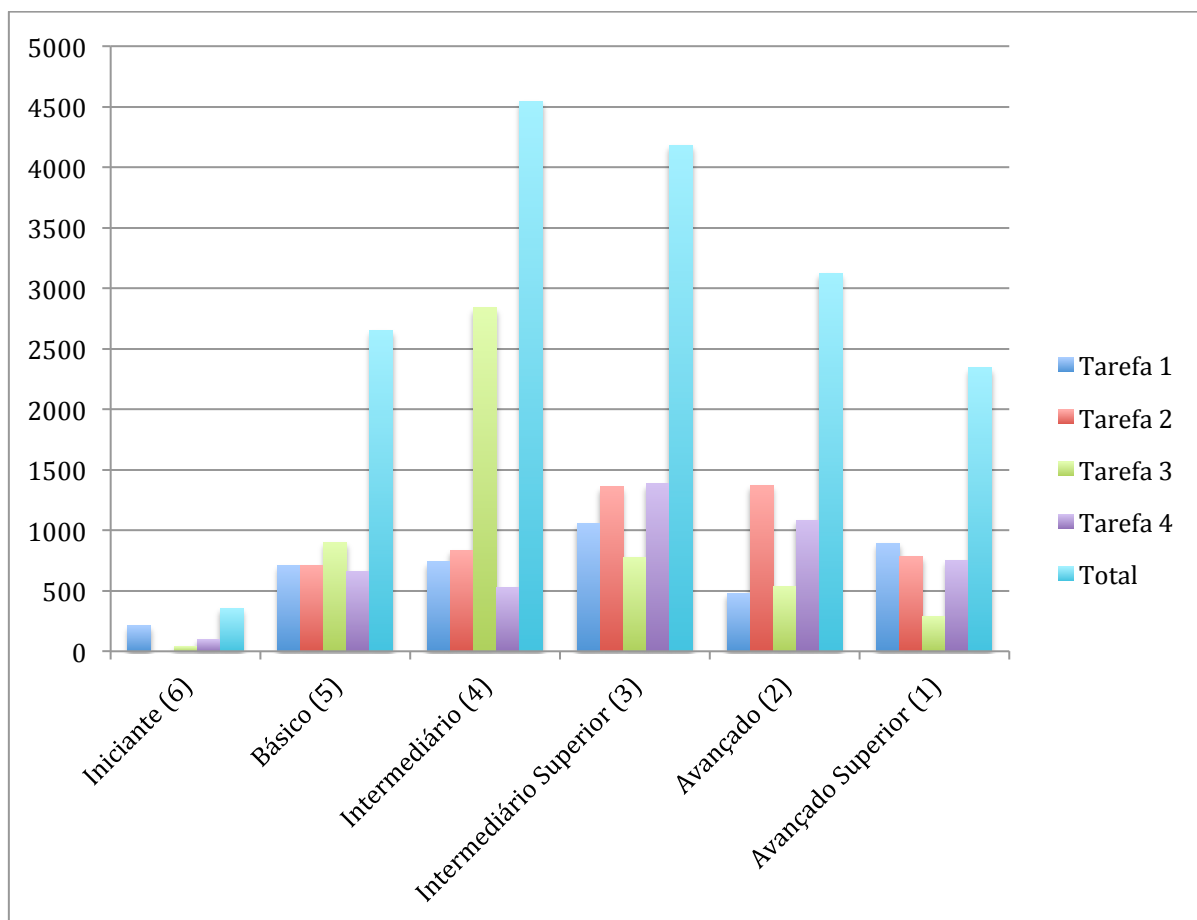


Gráfico 4 Número de palavras DISTINTAS por Tarefa e Nível.

Considerando a ocorrência de palavras por nível, observa-se um uso crescente do número de palavras em relação ao nível do conjunto de textos, e essa relação é observada tanto no número total de palavras quanto no número de palavras distintas (os Gráficos 3 e 4 são bastante parecidos quando comparados). Com relação a essa tendência do uso de palavras, verificou-se uma tendência diferente entre o uso total de palavras e o uso de palavras distintas, conforme ilustrado no Gráfico 5 a seguir:

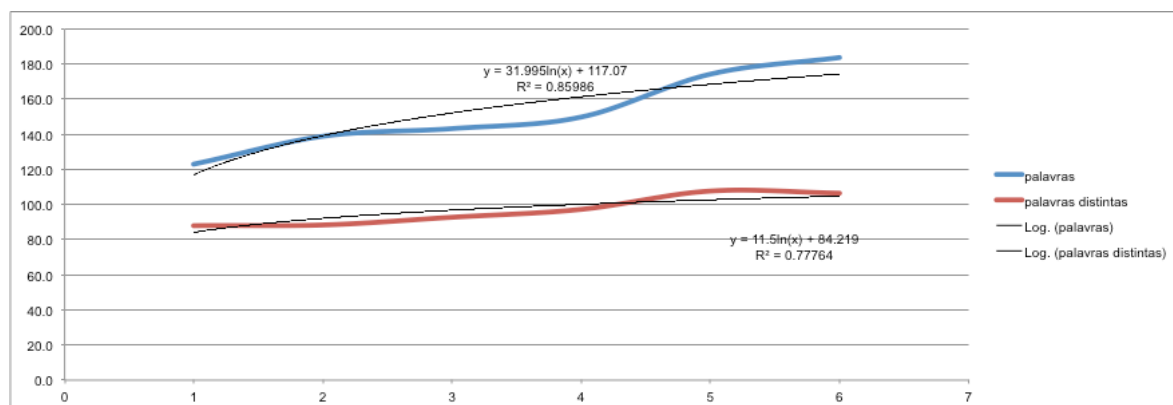


Gráfico 5 Palavras totais (linha azul) e palavras distintas (linha vermelha).

No Gráfico 5 acima, os números do eixo horizontal correspondem aos níveis de proficiência e conjuntos (1 sendo Iniciante, progredindo até o 6, Avançado Superior).

O número total de palavras, representado pela linha AZUL, pode ser separado em duas fases. A primeira, onde há um crescimento importante no número de palavras em relação ao aumento do nível de proficiência dos textos (de Iniciante a Avançado), que se estabiliza ao chegar aos níveis Intermediário e Intermediário Superior; a segunda fase, marcada do nível Intermediário Superior em diante, em que as produções textuais apresentam um número maior de palavras.

A linha VERMELHA representa o número de palavras distintas dos textos e conjuntos. Com relação ao uso de palavras distintas, verificou-se que ocorre um uso em escala logarítmica. É possível perceber que, quanto mais Básico é o nível do conjunto de textos, maior é a preocupação com o uso de palavras, ou seja, os textos de nível Básico possuem mais palavras distintas que os demais conjuntos de textos.

Ao observarmos o comportamento quanto ao uso de palavras do conjunto de textos classificado como Avançado, vemos novamente um crescimento no número de palavras, muito semelhante ao demonstrado pelo conjunto de textos Básico. No entanto, uma grande diferença marca esses dois níveis: o conjunto de textos do nível Avançado demonstra uma variação vocabular menor, ou seja, os candidatos produziram textos com maior número de palavras repetidas, não variando tanto o vocabulário.

Tendo em vista os resultados com relação ao uso de palavras, optou-se por verificar o que aconteceria com relação à intersecção de vocabulário entre níveis e tarefas. Para identificar a intersecção de vocabulário entre os conjuntos, o Índice de Jaccard⁴⁴ foi utilizado.

⁴⁴ O Índice Jaccard é usado para verificar a sobreposição de dois grupos, A e B. O objetivo de utilizar esse coeficiente é o de verificar, estatisticamente a similaridade e a diferença entre as amostras dos conjuntos. Esse Coeficiente não considera a frequência das palavras (quantas vezes a palavra ocorre em um documento).

As medidas de similaridade lexical são aplicadas entre os segmentos, estabelecendo relações mais fortes ou mais fracas. A intersecção⁴⁵ resultou em uma tabela simétrica, em que é possível observar um padrão. Quanto mais Avançado é o nível do conjunto de textos, maior é a sua similaridade, com relação às palavras usadas, com os demais grupos. A Tabela 9 a seguir mostra os resultados numéricos para essa intersecção:

	Avançado Superior	Avançado	Intermediário Superior	Intermediário	Básico	Iniciante
Avançado Superior	*	0.176	0.165	0.148	0.113	0.152
Avançado	0.176	*	0.080	0.071	0.088	0.076
Intermediário Superior	0.165	0.080	*	0.062	0.058	0.065
Intermediário	0.148	0.071	0.062	*	0.055	0.063
Básico	0.113	0.088	0.058	0.055	*	0.075
Iniciante	0.152	0.076	0.065	0.063	0.075	*

Tabela 9 Intersecção de vocabulário entre os conjuntos de textos.

Outra forma de calcular a relação entre grupos foi verificar a similaridade entre eles. Para isso, utilizou-se como métrica a similaridade pelo cosseno⁴⁶ (técnica de PLN muito comum em Recuperação de Informação para fazer comparação entre documentos). As *dimensões* são as palavras e os *valores* dos vetores são as frequências dessas palavras. Calculando a similaridade entre os grupos, foi obtida a Tabela 10 a seguir:

	Avançado Superior	Avançado	Intermediário Superior	Intermediário	Básico	Iniciante
Avançado Superior	*	0.456	0.0391	0.0346	0.0344	0.0335
Avançado	0.0456	*	0.0368	0.0332	0.0325	0.0309
Intermediário Superior	0.0391	0.0368	*	0.0295	0.0284	0.0264
Intermediário	0.0346	0.0332	0.0295	*	0.0263	0.0235
Básico	0.0344	0.0325	0.0284	0.0263	*	0.0241
Iniciante	0.0335	0.0309	0.0264	0.0235	0.0241	*

Tabela 10 Similaridade de vocabulário entre os grupos.

⁴⁵ Normalizada entre os grupos (normalização: palavras distintas em comum pelo número de palavras distintas).

⁴⁶ A Similaridade do Cosseno é uma medida de similaridade entre dois vetores que mede o cosseno do ângulo formado entre eles.

Uma forma gráfica de visualizar a Tabela 10 é dada pela Figura 17 a seguir, na qual se observa que quanto mais Avançado é o nível, maior é a similaridade com os demais níveis:

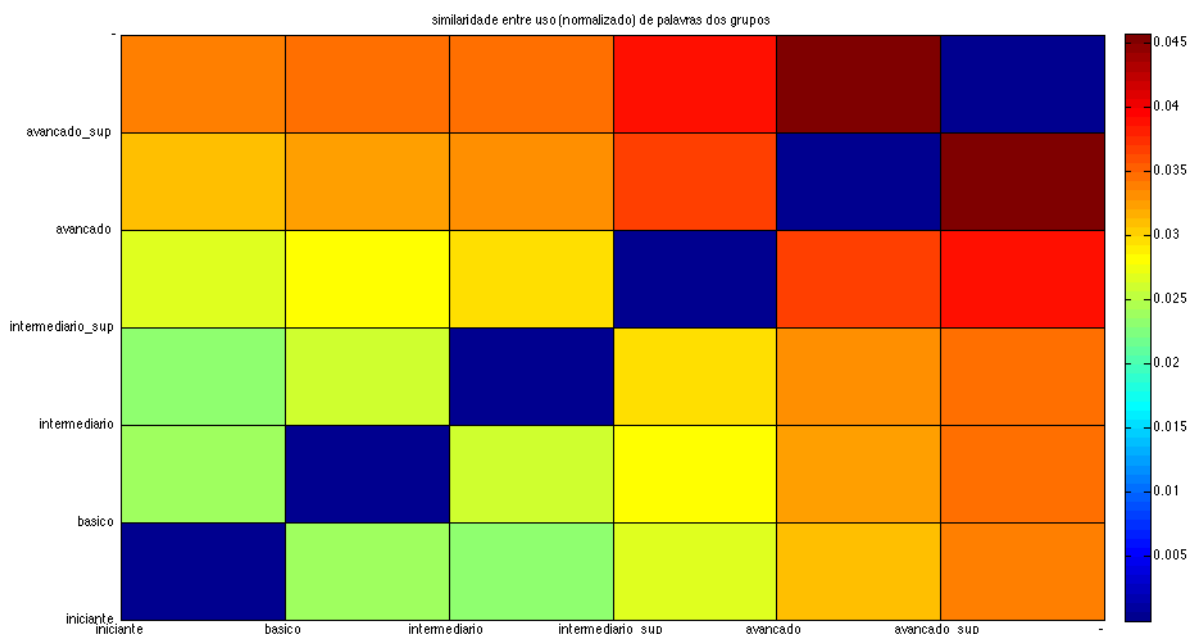


Figura 17 Similaridade de vocabulário entre os grupos.

Na imagem colorida acima existem dois eixos em que foram colocados os níveis de proficiência avaliados no exame Celpe-Bras (do Iniciante ao Avançado Superior, conforme descritos anteriormente). A cor azul escura corresponde à sobreposição do nível com ele mesmo (por exemplo, Iniciante com Iniciante, Básico com Básico, e assim por diante) em termos de léxico. A imagem mostra de forma gráfica o quão similares são os grupos quando comparados entre si, então, quanto mais próximos forem os tons de cor, mais próximo é o vocabulário que compartilham. Dessa forma, o grupo Básico do eixo vertical, quando sobreposto ao grupo Básico do eixo horizontal, tem uma relação marcada pela cor azul escura, o que significa que possuem o mesmo vocabulário. Ao observarmos a relação que se estabelece entre o grupo Avançado Superior e Avançado, veremos que os tons de vermelho prevalecem, o que quer dizer que o vocabulário utilizado por esses grupos nesta amostra são muito similares. O mesmo pode ser observado entre os demais grupos cuja proficiência é próxima (Básico e Intermediário e Intermediário Superior e Avançado, por exemplo).

A mesma comparação pode ser feita entre grupos que possuíram desempenho muito diferente. As cores frias (tons de verde que caracterizam os grupos Iniciante e Básico) se diferenciam e se distanciam das cores quentes (tons de vermelho que caracterizam os grupos Avançado e Avançado Superior) na imagem.

8.2 RESULTADOS COESIVOS

Partindo da microperspectiva estrutural do texto, reitero, considerando apenas sua tessitura coesiva, verificou-se como os conjuntos de textos se comportaram em relação às métricas do Coh-Matrix-Port. Observou-se que alguns atributos apresentam um comportamento relevante a ser verificado na classificação automática dos textos.

Analisando como as classes de textos se comportaram em relação à média, observamos que alguns atributos dados pelo Coh-Matrix-Port apresentaram um comportamento não aleatório entre os diferentes conjuntos de textos. Dos 48 atributos fornecidos pela ferramenta Coh-Matrix-Port, 15 demonstraram comportamento não aleatório, ou seja, foi possível estabelecer uma correlação entre suas presenças e os níveis de proficiência avaliados pelo exame Celpe-Bras.

Esses atributos foram descobertos através da observação das árvores resultantes na ferramenta Weka. Para estabelecer a correlação, adotamos critérios bastante flexíveis pois não há um número de significância estabelecido em linguística para apontar o tipo de correlação que queremos (na Computação, seria 1%, na Física, 0,01%). Por exemplo, geralmente, o número de palavras sozinho prevê bem classes de textos, mas não significa que isso seja verdade quando correlacionado a outros atributos, e essa situação varia de *corpus* para *corpus*. Através da observação, geramos a Tabela 11 a seguir que mostra os 15 atributos (dos 48 fornecidos pelo Coh-Matrix-Port) que seriam, em tese e a partir da nossa leitura das árvores, possíveis bons indicadores para diferenciar textos mais proficientes de textos menos proficientes, pois apresentavam uma média diferente com relação aos demais atributos.

	ATRIBUTOS	BOM INÍCIO	BOM MEIO	BOM FIM	INVERSO
1	numero_palavras	S	S	S	N
2	numero_pronomes	S	N	S	S
3	sentencas_por_paragrafos	S	S	S	N
4	silabas_por_palavras	S	S	S	N
5	Flesch	S	S	S	S
6	numero_functional_words	S	N	S	N
7	hiperonimos_verbos	S	N	S	N
8	palavras_antes_verbos	N	N	S	N
9	pronomes_pessoais	S	N	S	S
10	tipo_token	N	S	S	S
11	pronomes_por_sintagmas	N	S	S	S
12	numero_operadores_logicos	S	N	S	N
13	conectivos_logico_positivos	S	N	S	S
14	ambiguidade_substantivos	S	N	S	N
15	ambiguidade_adjetivos	S	N	S	S

Tabela 11 Atributos (medidas do Coh-Matrix-Port) que demonstraram comportamento não aleatório.

Com a observação desses 15 atributos, foi possível identificar “comportamentos” que puderam ser entendidos como relevantes para a separação dos níveis de proficiência quanto aos valores apresentados pelo sistema. De forma bastante simplificada e para fins de testagem, determinamos que:

- Quanto maior fosse o valor, maior seria a proficiência (numero_palavras, sentencas_por_paragrafos, silabas_por_palavras)
- Quanto maior fosse o valor, maior seria a proficiência (Intermediário) (numero_functional_words, hiperonimos_verbos, numero_operadores_logicos, ambiguidade_substantivos)
- Quanto maior fosse o valor, menor seria proficiência (Básico) (palavras_antes_verbos)
- Quanto maior fosse o valor, menor seria a proficiência (Intermediário) (numero_pronomes, Flesch, pronomes_pessoais, conectivos_logico_positivos, ambiguidade_adjetivos)
- Quanto maior fosse o valor, menor seria a proficiência (tipo_token, pronomes_por_sintagmas)

Dessa forma, não informamos pontualmente em qual classe um texto que apresentasse determinado padrão se encaixaria, mas apontamos tendências de comportamentos de textos mais e menos proficientes com relação à presença dos 15 atributos.

8.3 RESULTADOS DO APRENDIZADO DE MÁQUINA

A fim de identificar como esses atributos de fato seriam capazes de descrever os conjuntos de textos, ou seja, os níveis de proficiência propostos pelo exame, o algoritmo J48 (QUINLAN, 1993 – implementação do Weka) foi utilizado para a obtenção de uma Árvore de Decisão que separasse os níveis de proficiência dos textos de acordo com os 15 atributos identificados através da média e das observações descritas na seção anterior. A parametrização utilizada no algoritmo foi o padrão do Weka.

A Árvore de Decisão obtida para os seis níveis de proficiência (Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior) e os 15 atributos selecionados possui *recall* 26%, *precision* 24,3% e *f-measure* 24,3%, que apresentam um índice de acerto relativamente baixo ao classificar os textos em seis classes com os atributos apresentados. Esses resultados, porém, não são necessariamente ruins, pois em AM realizado

com seis classes é até normal que o classificador seja menos preciso. Por exemplo, se considerarmos apenas duas classes, 50% seria um valor baixo, devido às chances que o classificador teria de acertar, que são bem maiores para duas classes do que para seis. Ainda assim, a Árvore de Decisão resultante foi capaz de indicar os seguintes comportamentos dos conjuntos de textos produzidos para o exame:

- Alguns textos classificados como **Básico** pelos avaliadores humanos tendem a utilizar muitas palavras por sentenças com relação aos textos **Avançados e Intermediários**.
- Textos classificados como **Intermediário** pelos avaliadores humanos tendem a utilizar mais conectivos com relação aos outros conjuntos de textos.
- Textos classificados como **Intermediário** pelos avaliadores humanos tendem a utilizar um número menor de palavras com relação aos textos **Avançados**.
- Textos classificados como **Básico** tendem a ser menores.
- O uso de conectivos causais negativos é característico dos textos classificados como **Avançado**.
- O uso de conectivos temporais positivos é característico dos textos classificados como **Avançado**.

A Figura 18 a seguir mostra a Árvore de Decisão descrita acima.

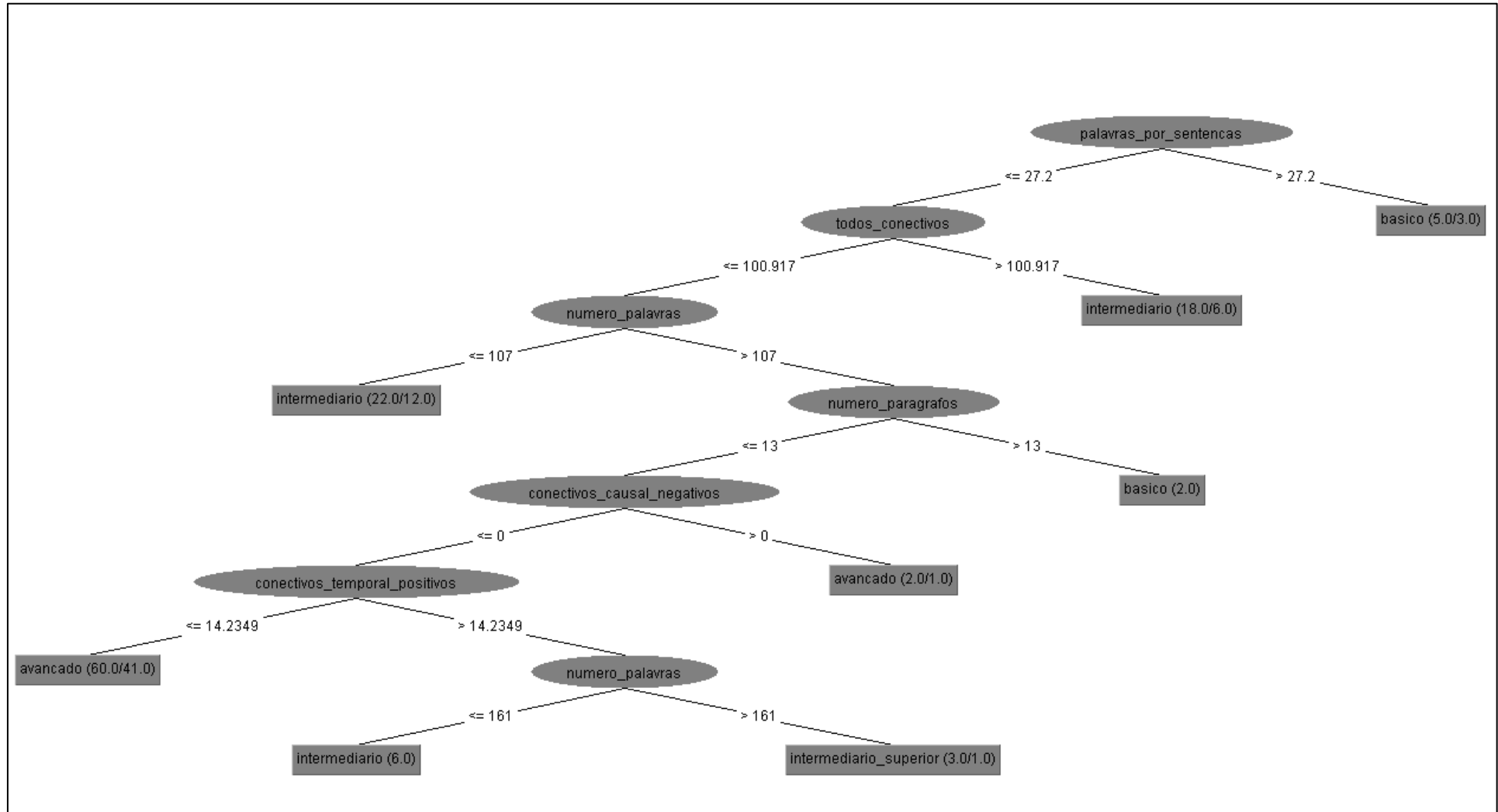


Figura 18 Árvore de Decisão para os seis níveis (Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior).

Através da análise da Matriz de Confusão⁴⁷, apresentada na Tabela 12, identificamos que o problema na classificação está no fato de as classes não serem claramente distinguíveis. Percebe-se que a confusão ocorre entre as classes mais próximas (Avançado Superior com Avançado; Avançado com Intermediário Superior e Avançado Superior; Intermediário Superior com Avançado e Intermediário; Intermediário com Intermediário Superior e Básico; e Básico com Iniciante e Intermediário):

		Predição de Classe					
		Avançado Superior	Avançado	Intermediário Superior	Intermediário	Básico	Iniciante
Classe Verdadeira	Avançado Superior	8	3	6	4	1	0
	Avançado	5	1	17	4	2	0
	Intermediário Superior	5	6	12	16	4	0
	Intermediário	4	2	13	21	9	0
	Básico	6	2	7	10	4	1
	Iniciante	0	1	1	1	1	0

Tabela 12 Matriz de Confusão do classificador.

A Matriz de Confusão está representada, também, no Gráfico 6 a seguir:

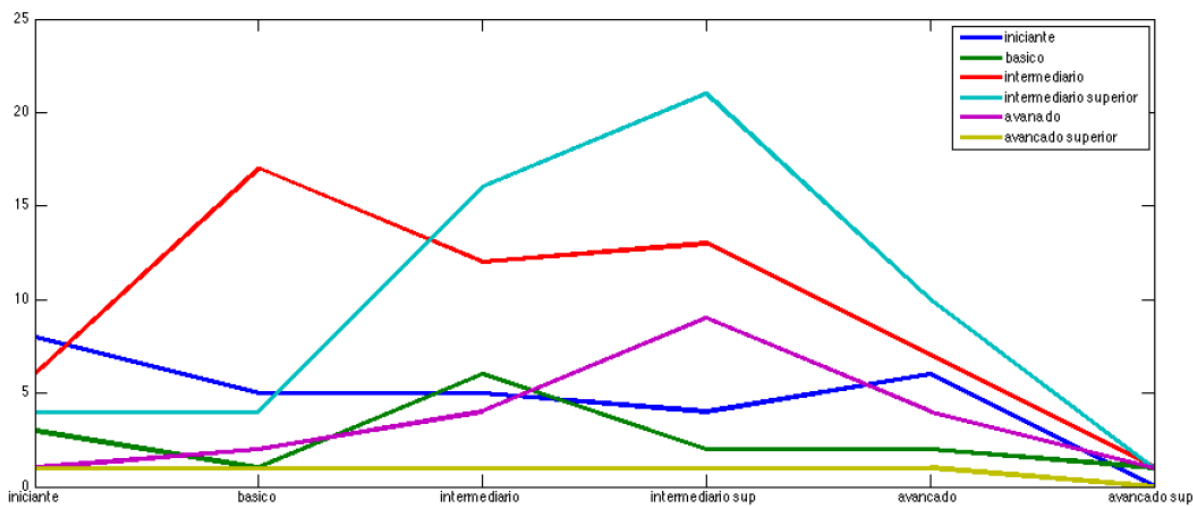


Gráfico 6 Matriz de Confusão do classificador.

Considerando a separação não clara entre as classes mais próximas, analisamos a capacidade de o classificador separar apenas o conjunto de textos que receberia certificado do

⁴⁷ Como explicado no item 5.4.2, a Matriz de Confusão é uma tabela específica cuja disposição de informações permite a visualização do desempenho de um determinado algoritmo, normalmente, um algoritmo de aprendizado supervisionado. Cada coluna dessa matriz representa uma instância de uma dada classe, enquanto cada linha dessa matriz representa as instâncias de uma classe verdadeira. O nome Matriz de Confusão vem do fato de que essa matriz torna fácil de ver se o sistema está confundindo duas classes (ou seja, colocando textos na classe errada).

exame do conjunto de textos que não receberia certificado (o exame certifica, conforme esclarecemos nos Capítulo 1, apenas examinandos que alcancem os níveis acima do Intermediário)⁴⁸. Dessa forma, utilizamos o mesmo conjunto de dados (177 textos processados pela ferramenta Coh-Metrix-Port), porém, usando apenas duas classes:

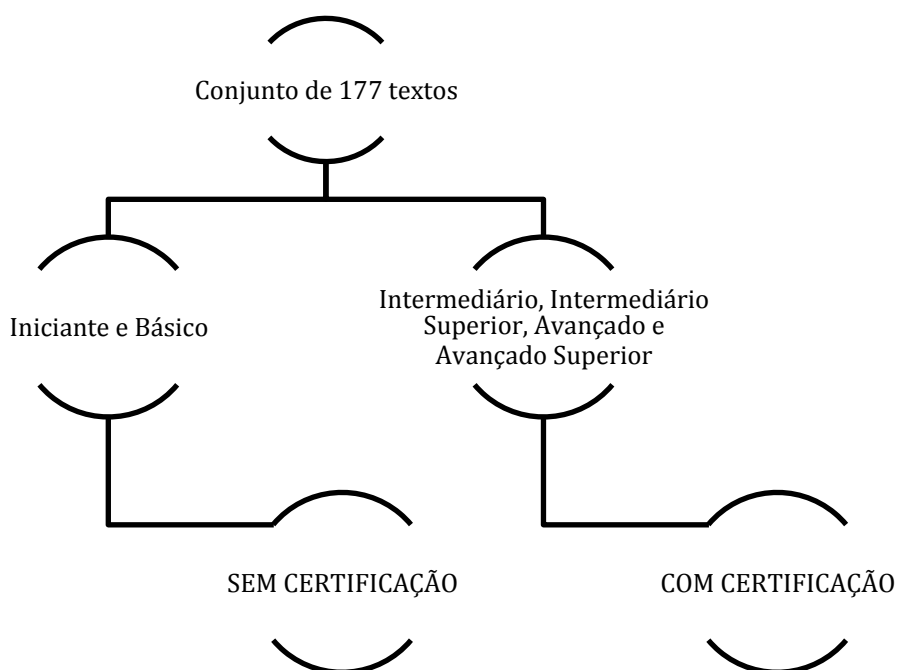


Figura 19 Novas classes para os textos: SEM CERTIFICAÇÃO e COM CERTIFICAÇÃO.

Utilizando os textos nivelados como Iniciante e Básico como SEM CERTIFICAÇÃO e os demais textos como COMO CERTIFICAÇÃO, usando a mesma configuração do algoritmo J48 do experimento anterior, obteve-se a Árvore de Decisão mostrada na Figura 20, a qual possui *recall* 73,3%, *precision* 75,4% e *f-measure* 72,1%, resultados muito melhores do que os apresentados anteriormente:

```

numero-palavras > 111
— numero-paragrafos > 3
——conectivos-adtivos-negativos <= 11.5607
— ——numero-adjetivos > 61.2245
— ——flesch > 46.5465
— —— numero-verbos > 180.851: aprovado
  
```

Figura 20 Árvore de Decisão para as duas classes (SEM CERTIFICAÇÃO e COM CERTIFICAÇÃO).

⁴⁸ Este experimento foi feito a título de curiosidade. Sabe-se que a certificação é conferida, no exame Celpe-Bras, a partir de uma média feita a partir das diferentes notas dadas para as 4 tarefas da Parte Escrita e da nota na Parte Oral do exame.

9 DISCUSSÃO DOS RESULTADOS

Após as seções anteriores, em que foram descritos os textos sob estudo, os resultados lexicais e coesivos e a análise classificatória por AM, aqui são discutidos alguns pontos de destaque ao longo desta pesquisa. No entanto, em virtude da variedade de dados levantados, será impossível ponderar detalhadamente sobre todos os tópicos. Em síntese, as discussões apresentadas neste capítulo não esgotarão todas as análises possíveis dos dados obtidos, oferecerão apenas primeiras impressões que se pode ter ao vislumbrar, pela primeira vez, o lado quantitativo dos textos do exame Celpe-Bras.

9.1 RESULTADOS LEXICAIS

No âmbito deste trabalho, não era esperado obter revelações importantes a partir da observação do comportamento lexical apresentado pelos textos das diferentes classes identificadas. No entanto, com relação aos resultados lexicais, descritos no item 8.1, observou-se que há tendências de usos de palavras (de concentração, presença e variação de palavras) que podem ser associadas aos 6 níveis de proficiência avaliados no exame Celpe-Bras da edição de 2006-1. Muito embora possam ser dados que não contribuam diretamente para uma tentativa de semi-automatização da etapa de correção do exame, esses índices são representativos dos grupos e merecem ser discutidos, especialmente ao considerar que este é o primeiro trabalho que apresenta alguma estatística com relação aos textos produzidos no contexto do exame.

A primeira tendência marcada é o maior número de palavras presente nos níveis Intermediários (Intermediário e Intermediário Superior). A segunda tendência apontada é o número menor de palavras presente no nível Iniciante. A terceira tendência é a presença de um número de palavras relativamente alto no nível Básico e similar ao número de palavras do nível Avançado Superior.

Essas tendências apontadas pelos resultados parecem estranhas a uma primeira vista – tal como a similaridade de número de palavras entre produções que estão nos dois extremos do *continuum* de proficiência, em que se esperaria um número maior de palavras em um texto

de um falante mais experiente da língua. Mas, ao pensar nos níveis de proficiência de um estudante de qualquer língua adicional, essas tendências podem ser reflexo de algo que professores relatam presenciar em sala de aula. Os grandes picos de aumento de vocabulário dos estudantes costumam ocorrer nos primeiros contatos dele com a nova língua, geralmente o que acontece nos níveis mais básicos dos cursos regulares. É evidente que, para confirmar isso, seria necessário desenvolver um novo estudo apenas com foco na aquisição de vocabulário.

O conjunto de textos dos níveis Intermediários demonstra um comportamento que pode ser entendido como mais estável ou estabilizado. Esse comportamento condiz, também, com o comportamento análogo de estudantes que, estando no nível Intermediário já há algum tempo em um curso de línguas, têm a sensação de não estarem aprendendo palavras novas. Esses estudantes, por muito tempo, passam utilizando o mesmo vocabulário, já bem estabelecido, momento em que, aparentemente, não estão aprendendo palavras novas.

Ainda seguindo a questão do uso das palavras, alguns dados mostraram-se interessantes com relação ao número de palavras distintas, que estão relacionadas ao item 5.3.2, *Type-Token Ratio* (TTR). No *corpus*, como se viu no Gráfico 5, verifica-se um uso marcado por duas fases: a primeira, do Básico até o Intermediário, em que há uma mimese do crescimento e estabilização do uso; e segundo, do Intermediário Superior em diante, em que o conjunto de textos dos examinandos apresentou um crescimento. É possível perceber que, quanto mais Básico é o nível do conjunto de textos, mais palavras distintas são usadas.

Esse é um dado interessante, visto que se poderia pensar que, quanto mais variado é o vocabulário usado pelo examinando, melhor seria a qualidade do seu texto. Na verdade, o que a amostra aponta é justamente o contrário. Uma razão para isso seria o provável desconhecimento da adequação de uso de palavras talvez apenas reconhecidas na tarefa e utilizadas para a construção do texto. Já os textos de nível Avançado repetem muitas palavras e por isso também acabam tendo um desempenho melhor. Talvez porque, justamente por se repetirem, acabam construindo um texto mais coeso e, portanto, mais claro, resultando em um desempenho geral melhor no cumprimento das tarefas propostas.

Saindo do escopo de número de palavras e de palavras repetidas, entra-se na questão da similaridade e intersecção de vocabulário. Observa-se um padrão na similaridade de uso lexical, o qual demonstra que quanto mais avançada é a proficiência do conjunto de textos, maior é a sua similaridade com os demais grupos. Constatou-se, na amostra, que a classe Avançado Superior apresenta maior similaridade com a classe Avançado, seguido por Intermediário e assim indo até o Iniciante. Esse dado aponta para um crescimento incremental

(gradual) do léxico entre as classes, o que implica em uma tendência de progressão de uso de palavras de um nível a outro de proficiência, sendo os níveis cuja proficiência foi marcada como mais avançada os que possuem palavras que existem nos outros níveis, ao menos na amostra coletada.

9.2 RESULTADOS COESIVOS

Em primeiro lugar, é preciso lembrar que a ferramenta Coh-Matrix-Port não foi planejada com a intenção de contrastar textos mais e menos proficientes ou de avaliar textos de estudantes de PLA. Isso, evidentemente, gerou uma série de dificuldades para o estabelecimento dos procedimentos e para a análise. O primeiro problema foi avaliar quais métricas, dentre as 48 do Coh-Matrix-Port, seriam aproveitáveis para a comparação dos textos previamente agrupados em níveis de proficiência. Das 48 métricas disponíveis, somente 15 mostraram alguma possibilidade de diferenciar as classes. Essas 15 métricas foram obtidas através da observação da árvore de decisão oriunda da ferramenta Weka.

Ainda que essas 15 métricas restantes sejam dignas de mérito, elas não são suficientes para uma descrição mais completa e abrangente das diferenças coesivas entre textos mais e menos proficientes, especialmente no que diz respeito a todas as 6 classes que tentamos diferenciar. Elas não se mostraram muito úteis, no entanto, quando os grupos foram separados novamente em um número menor de classes (SEM E COM CERTIFICAÇÃO), as métricas do Coh-Matrix-Port mostraram-se mais produtivas. A combinação da presença dos itens a seguir permitiu a separação de textos considerados CERTIFICADOS e NÃO CERTIFICADOS.

- mais de 100 palavras
- mais de 3 parágrafos
- valor menor ou igual a 11,56 na métrica conectivos_ativos_negativos
- valor maior que 61,22 para a métrica numero_adjetivos
- Índice Flesch maior que 46,54
- valor maior que 180,851 para a métrica numero_verbos,

Houve, portanto, a correlação entre as métricas dadas pelo sistema Coh-Matrix-Port, um processo totalmente automatizado de mensurar coesão e coerência, e as avaliações dadas no contexto de aplicação de 2006-1 do exame Celpe-Bras.

9.3 RESULTADOS DO APRENDIZADO DE MÁQUINA

Adotar a metodologia de AM nesta pesquisa foi extremamente positivo e, acredito, o ponto alto de todo o trabalho desenvolvido no mestrado. Por ser *linguista*, apesar de estar já há bastante tempo acostumada com trabalhos lexicométricos e estatísticos, nunca havia tido a oportunidade de trabalhar com estatística tão avançada e complexa. Além disso, o AM não é algo simples de ser entendido, também não é trivial delinear um estudo capaz de cumprir todas as suas etapas e ainda resultar em algo produtivo.

Apesar do número reduzido de atributos distintivos e do baixo desempenho do classificador na primeira etapa (a Árvore de Decisão obtida, com as 48 métricas, teve *recall* 26%, *precision* 24,3% e *f-measure* 24,3%, considerados muito baixos na Avaliação de AM), das 15 métricas identificadas inicialmente, 6 delas foram capazes de apontar, com bom desempenho, se os textos receberiam certificação ou não, classificação condizente com as expectativas iniciais desta dissertação.

Além disso, o Índice Flesch, apresentado na Introdução do trabalho, está, felizmente, presente entre as métricas que diferenciam os textos. Esse índice, além de ser a única métrica totalmente adaptada ao português, é a única que aponta, em uma escala hierárquica de dificuldade, níveis objetivos de complexidade textual. Assim, um exame mais detalhado dos resultados envolvendo o índice Flesch se faz necessário.

10 RETOMADA DAS QUESTÕES E HIPÓTESES DE PESQUISA

Neste capítulo, retomo e comento as hipóteses e questões de pesquisa formuladas e apresentadas no início da dissertação.

10.1 QUESTÕES DE PESQUISA

1. Como é a configuração lexical e gramatical de textos submetidos a um exame de proficiência do português brasileiro que tem como construto teórico avaliar o desempenho do examinando através da aplicação de um teste comunicativo?

A partir da abordagem computacional dos textos do exame Celpe-Bras, não foi possível verificar claramente um perfil gramatical e lexical específico de textos submetidos ao exame. No entanto, com relação ao léxico utilizado, partindo das métricas e resultados obtidos com a ferramenta Coh-Metrix-Port e Weka, obtivemos resultados que podem ser aproveitados na busca de uma resposta para a questão colocada.

Estabelecendo uma relação entre o nível avaliado pelo exame e o número de palavras presentes nos textos, é possível perceber uma tendência de que os textos com maior número de palavras localizam-se nos níveis intermediários avaliados pelo exame. Esse resultado poderia gerar uma longa discussão a respeito de, por exemplo, passar-se a impor, no contexto de aplicação do exame, uma limitação ou um número mínimo de linhas ou palavras no momento da produção do texto.

É provável que alguns dos textos mais proficientes não apresentem um número grande de palavras, como foi visto; mas os textos avaliados como menos proficientes pelos corretores humanos eram, de fato, menores. Possivelmente, muitos dos textos tenham o mesmo número de palavras, tenham eles sido considerados intermediários, básicos ou avançados. Porém, do ponto de vista do avaliador que, assim como qualquer leitor de um texto, confere uma avaliação já na primeira impressão que tem do texto que está por ler, um texto maior pode dar a impressão de apresentar mais conteúdo e de ser melhor desenvolvido. Além disso, um texto

mais longo requer do avaliador, possivelmente, um número maior de leituras, de modo a concatenar os diferentes critérios e descritores presentes nas grades de correção.

O segundo aspecto é a presença de menor repetição de palavras nos textos avaliados como menos proficientes e maior repetição nos textos avaliados como mais proficientes. Outros estudos envolvendo avaliação de língua portuguesa, especialmente no contexto de provas de redação de vestibular (FINATTO, 2008), já mostraram que as redações tidas como NOTA 10 tendem a apresentar um vocabulário mais repetitivo. Nas pesquisas relatadas, especialmente a de Crossley e McNamara (2012), essa é uma falsa verdade. Neste estudo, foi possível mostrar que a presença de um vocabulário pouco variado acabou apontando para textos mais claros e coesos (conforme infiro), marcando uma preferência por parte dos examinandos em usar adequadamente palavras que realmente conhecem em detrimento de usar muitas palavras diferentes e tentar produzir um texto com maior riqueza vocabular.

Outro aspecto, que acabou não sendo ponderado neste trabalho, e que possivelmente fosse capaz de diferenciar o texto dos examinandos frente a outros textos escritos em português, é a presença recorrente de marcas de interferência de uma segunda língua no texto em português. Conforme os estudos com *corpora* de aprendizes (SHEPHERD, 2009), esse parece ser um ponto interessante a ser verificado, talvez fora do escopo de tentar pré-classificar os textos por níveis de proficiência automaticamente.

2. Qual relação podemos estabelecer entre a configuração do léxico empregado nas produções dos examinandos e os níveis de proficiência previamente avaliados no exame?

Nesta questão, repete-se uma evidência da questão anterior: houve uma correlação forte entre o tamanho dos textos, a variedade do vocabulário e o índice Flesch e determinados níveis de proficiência atribuídos aos textos. Essa correlação coloca-se em termo estatísticos.

Por exemplo, como já explorado nas seções 9.1, 9.2 e 9.3, o número de palavras de um texto (mais de 100 palavras), atrelado ao número de parágrafos (mais de 3) foram capazes de diferenciar os níveis de proficiência (quando colocados no agrupamento SEM E COM CERTIFICAÇÃO).

3. A partir de um enfoque da LC e do PLN, quais são os melhores pontos de aproveitamento para pesquisas sobre a produção escrita em PLA?

Um dos principais pontos de aproveitamento deste trabalho para a pesquisa em PLA, tal como feita no Brasil, no meu entender, é justamente a visão de um *corpus* tal como

entendido no âmbito da LC e do PLN. Um *corpus* de estudantes, como se mostrou nos capítulos 4 e 5, tende a gerar dados extremamente úteis, tanto para a descrição dessa modalidade de uso de linguagem quanto para a elaboração de estratégias especificamente voltadas para o ensino.

A publicação *Corpora no ensino de línguas estrangeiras*, recentemente organizada e publicada por Viana e Tagnin (2011), por exemplo, traz interessantes ideias que podem ser estendidas a um *corpus* de PLA reunido e organizado em grande escala. Do mesmo modo, o acesso a ferramentas de PLN, tal como o Coh-Metrix-Port, por si só, já pode representar um recurso interessante para o professor de PLA. Penso que o uso dessa ferramenta para acompanhar o rendimento do estudante, como feito no experimento relatado na Introdução desta dissertação, pode ser um facilitador no momento em que o professor precisa dar um *feedback* ao estudante, ou quando o próprio professor não consegue enxergar os problemas que esse estudante está apresentando em suas produções textuais.

Além disso, hoje, por exemplo, já há no mercado produtos que preparam os estudantes para o exame Celpe-Bras. O recurso digital *Celpe-Bras sem segredos* (FORTES, 2012) é um conteúdo digital que traz um vasto material de preparação para o exame, com 16 exames, avaliações e ricos recursos que são possíveis graças à interatividade disponibilizada pelo material. Em uma breve avaliação do material, verifiquei que o sistema de *feedback* ao examinando não existe. O que existem são exemplos de respostas ideais às tarefas dos simulados, mas nada foi desenvolvido a fim de previamente avaliar as produções dos estudantes e mostrar ao usuário que atributos tem o texto que ele gera. Acredito que trabalhos como os desta dissertação possam subsidiar a melhoria de ferramentas didáticas como essa.

4. Partindo de textos previamente avaliados por corretores humanos, é possível apontar e formalizar elementos capazes de diferenciar os níveis automaticamente?

Em síntese, a resposta para essa pergunta é sim, mas não conforme idealizamos. Num mundo ideal, as métricas do Coh-Metrix-Port, que foram transformadas em atributos e que foram processadas por AM revelariam características particulares e específicas de cada um dos 6 níveis avaliados pelo exame, e corresponderiam às avaliações atribuídas pelos corretores humanos. O número de classes (6 níveis) foi grande demais para que a máquina fosse capaz de fazer tal distinção, mas em 2 classes, os resultados foram diferentes.

Mesmo assim, as métricas a seguir, combinadas e no valor indicado, foram capazes de separar **eficientemente** o conjunto de 177 textos, nesta nossa pequena amostra, em CERTIFICADOS e NÃO CERTIFICADOS:

1. mais de 100 palavras
2. mais de 3 parágrafos
3. valor menor ou igual a 11,56 na métrica conectivos_adtivos_negativos
4. valor maior que 61,22 para a métrica numero_adjetivos
5. Índice Flesch maior que 46,54
6. valor maior que 180,851 para a métrica numero_verbos,

Não resta dúvida de que há um longo caminho ainda a ser trilhado com a finalidade de responder completamente essa questão de pesquisa, mas tendo em mãos algumas pistas e, principalmente, a metodologia de AM, os traços escolhidos para distinguir os níveis automaticamente podem ser refinados e formalizados. Talvez mais análises lexicais possam trazer mais informações sobre o texto; talvez etiquetas relacionadas aos gêneros solicitados pelas tarefas sejam reveladoras; mas, com certeza, com um *corpus* tão pequeno e com reduzido capital humano, não será possível responder a esta pergunta tão cedo. É preciso ir além.

10.2 HIPÓTESES DE PESQUISA

I. Padrões de frequência e distribuição das palavras ao longo dos textos e do *corpus* são capazes de colaborar para distinguir níveis de proficiência em português.

Hipótese confirmada. Evidentemente, é preciso considerar as limitações desta pesquisa, sobretudo no que diz respeito ao tamanho do *corpus*. No entanto, no conjunto de textos investigados, a hipótese se confirma. Elementos, tais como **maior número de palavras** e **maior número de palavras repetidas** estão associados ao nível de proficiência **Avançado**. O **maior número de palavras distintas** está associado ao nível **Básico**. A **similaridade lexical** entre os textos mostrou que níveis **limítrofes**, como Intermediário e Intermediário Superior, estão muito próximos.

II. Textos maiores tendem a ser avaliados como mais proficientes.

Hipótese parcialmente confirmada. Nesta amostra, os textos **Intermediários** possuem o **maior número de palavras**, enquanto textos das **extremidades** do *continuum* de

proficiência, **Básicos** e **Avançados**, exibem um número **menor de palavras** em comparação aos Intermediários. O mesmo pode não se repetir em textos oriundos de outras edições do exame Celpe-Bras ou advindos de outros instrumentos avaliativos.

III. A relação entre número de palavras e palavras diferentes presentes em um dado texto (*Type Token Ratio*) é um bom indicativo para distinção de níveis de proficiência.

Hipótese confirmada. Nesta amostra, a **TTR** – medida de riqueza lexical calculada de forma bastante simples: divide-se o número de palavras diferentes de um texto (os *types* ou tipos) pelo número total de palavras desse mesmo texto (os *tokens*) – indica que textos mais proficientes – **Avançados** – possuem **menos riqueza lexical do que** os textos menos proficientes – **Básicos**.

IV. O Índice Flesch é capaz de distinguir, combinado com outros índices e métricas, níveis de proficiência de português.

Hipótese confirmada. Apesar de ser um indicativo mais “potente”, este índice deve ser considerado **SEMPRE** em combinação com outros elementos, tal como Pasqualini (2012) já havia confirmado em seu trabalho sobre traduções literárias mais ou menos complexas para diferentes tipos de leitor. O índice **Flesch**, contextualizado e **enriquecido pelo acréscimo de outros elementos textuais** e de abordagens estatísticas de análise de complexidade textual, como a técnica de AM aplicada nesta pesquisa, é **um indicador bastante confiável** para apontar se as produções textuais investigadas, quando comparadas, são mais ou menos proficientes.

11 LIMITES DO TRABALHO, PERSPECTIVAS E CONSIDERAÇÕES FINAIS

Neste trabalho, assumi como premissa que seria possível classificar textos de estudantes de PLA automaticamente, partindo de características advindas do tratamento descritivo dos textos sob consideração e de avaliações dadas por leitores humanos. O desafio seria verificar, de um modo estatisticamente válido, **quais e que tipos** de características exibidas pelos textos seriam mais profícuas para guiar essa classificação automática. Assumi, também, que a ferramenta Coh-Matrix-Port poderia apontar um bom conjunto dessas características, com a vantagem desse apontamento ocorrer de um modo automático, dado texto a texto.

Naturalmente, desde longa data, é possível classificar e identificar textos automaticamente, independente do objetivo que se tenha, e as filtragens do buscador Google nos mostram isso todos os dias. Pensando-se em uma busca mais filtrada, para além de um texto conter apenas uma palavra ou expressão, é possível separar os textos longos dos textos curtos, textos com títulos dos sem títulos, separar e-mails de contratos, entre outras tantas separações que já são possíveis automaticamente. Essas separações se dão facilmente se houver alguma etiqueta suficientemente distintiva nos textos que se acessa ou se estiver visível a, hoje singela, informação do recurso *Wordcount*, contador de palavras ao lado de cada arquivo/texto. Pela ótica da Linguística de Corpus e de suas ferramentas, usando-se um recurso como o *Wordsmith tools* ou o *AntConc*, é possível distinguir textos entre si comparando as suas listas de palavras geradas automaticamente. É possível comparar enormes *corpora* entre si apenas observando as suas estatísticas lexicais e os tipos de gêneros de textos neles contidos, desde que listados.

Sabendo disso, ao começar este trabalho minha principal pergunta foi: como se poderia fazer uma classificação automática de textos submetidos ao exame Celpe-Bras?

A junção entre os princípios da LC e as técnicas do PLN mostra que é possível selecionar determinadas características textuais e manipulá-las, de um modo eficiente, para fins dessa classificação automática. Neste trabalho, o que procurei fazer foi estabelecer uma correlação entre as características textuais medidas pela ferramenta Coh-Matrix-Port e as

características textuais usadas, mas não explicitadas sob a forma de um parecer circunstanciado, por avaliadores humanos ao atribuírem notas a produções de texto submetidas ao exame Celpe-Bras.

As características textuais medidas pelo Coh-Metrix-Port – majoritariamente coesivas – forneceram traços a serem comparados entre os textos. Esses traços e textos, por sua vez, foram processados na ferramenta Weka, com a finalidade de estabelecer uma correlação estatisticamente assistida e automática. Os resultados mostraram que algumas características automaticamente extraídas dos 177 textos desta pesquisa estão mais fortemente relacionadas a determinados níveis de proficiência atribuídos por avaliadores humanos.

Evidentemente, enfrentamos algumas limitações na proposição de uma metodologia automática, baseada em *corpus*, visando classificar automaticamente um conjunto de textos. A primeira limitação foi a falta de um *corpus* maior, o que inibe o alcance do tratamento estatístico. Foi bastante difícil encontrar o *corpus* de estudo e bastante trabalhoso prepará-lo para os experimentos. A ferramenta Coh-Metrix-Port exige um cuidadoso trabalho prévio com os textos a ela submetidos, especialmente a conferência de marcas de parágrafo e de pontuação.

A segunda limitação está na natureza da ferramenta informatizada, o Coh-Metrix-Port, que produz uma série de informações sobre os textos, tratados um a um, mas que nos fornece informações que se restringem a elementos ainda superficiais, tais como contagem de palavras, tamanho de sentença e contagens de palavras de determinados tipos. A classificação, que aproveitou apenas 15 de suas 48 métricas e índices, ficou um tanto comprometida, e não há como comparar resultados obtidos com a ferramenta Coh-Metrix para o inglês, por exemplo, muito mais robusta e que conta hoje com mais de 600 métricas de inteligibilidade textual.

A terceira limitação, e talvez a maior delas, foi justamente a tentativa de formalizar elementos de uma tarefa altamente complexa e subjetiva: a avaliação humana. Tendo em vista que a avaliação do exame Celpe-Bras é holística e que os recursos linguísticos concretamente presentes nos textos possuem um estatuto declaradamente menor em relação a outros elementos, tal com o gênero (que tem um grande peso), apontar itens específicos capazes de diferenciar os textos foi uma tarefa complexa. Entretanto, apesar dos limites, acredito que a ideia metodológica aqui posta tem grande potencial de aproveitamento, como se ressaltou na retomada das questões e hipótese de pesquisa. Afinal, foram testados índices lexicais e coesivos que se mostraram capazes de distinguir conjuntos de textos por níveis de proficiência a eles atribuídos, ao menos na mostra deste trabalho.

Diante das respostas para as questões colocadas e das evidências para as hipóteses, ficou claro que há um potencial de aproveitamento da metodologia aqui apresentada. O aproveitamento não é só para organizar e descrever *corpora* de estudantes de PLA, mas o trabalho também mostra como associar avaliações atribuídas à produção escrita e dados do conjunto de textos sob análise. No entanto, tendo em vista o tamanho reduzido do *corpus* e o caráter experimental desta dissertação, ainda é necessário verificar o alcance da metodologia para um universo maior e mais variado de textos, algo a ser feitos em larga escala e em trabalhos futuros.

Além dessa contribuição, a caracterização de elementos mais recorrentes em cada um dos níveis pode inclusive gerar bases para *feedbacks* a serem dados a estudantes que estejam se preparando para o Celpe-Bras ou mesmo para qualquer estudante de português. Assim, se poderia saber, com alguma objetividade (ainda que se pense em categorias prototípicas), o que fazer para melhorar uma produção textual que se encaixe em algum desses padrões. Evidente que a forma como esses *feedbacks* seriam dados exigiria ainda outro estudo, de modo a transformar os dados que aqui foram obtidos em informações palatáveis a professores e estudantes.

Mas ficou claro que a principal contribuição desta pesquisa refere-se à própria metodologia: como levantar informações linguísticas de um *corpus* de estudantes de português, com todas as suas peculiaridades, e como utilizá-las para implementação de sistemas computacionais. Como trabalhos futuros, sugerem-se:

- a) Realizar pesquisas sobre expressões classificadoras de um domínio/gênero;
- b) Comparar os traços de um *corpus* de falantes nativos de português com um *corpus* de estudantes ou contrastar um *corpus* de estudantes de português brasileiro com um *corpus* de estudantes de português europeu;
- c) Ampliar o *corpus* e testar o papel de novos atributos não tratados no sistema Coh-Matrix-Port – como as elipses na correlação com níveis de proficiência pré-estabelecidos. Também aqui caberia, talvez, estabelecer-se uma taxonomia de traços para avaliação de proficiência escrita do português;
- d) Classificar o *corpus* em função de gêneros textuais/discursivos e tipos de tarefas, correlacionando-as aos níveis de proficiência maiores ou menores. A utilidade de um estudo que viabilizasse isso ficou bastante clara ao estudar o construto teórico do exame e as noções de gêneros do discurso;
- e) Detectar em um *corpus* de examinandos quais características linguísticas mais “finas” estariam associadas àquilo que em geral tende a ser

subjetivamente reconhecido pelos avaliadores como “fluência” maior ou menor. Comentários como “o texto avançado é mais fluente” são um tanto vagos para descrever um nível de desempenho não se entendendo que caracterize essa “fluência”;

- f) Correlacionar características linguísticas do texto da tarefa (e do enunciado da tarefa) e a produção textual dos examinandos, também com vistas a detectar cópia de grandes segmentos. Com a sistemática apresentada na seção 8.1, sobre resultados lexicais, poder-se-ia implementar outro estudo buscando a similaridade de um conjunto de textos de uma mesma tarefa e a própria tarefa (enunciado, textos-base). Essa similaridade encontrada seria fundamental para a descrição do item Propósito e Informações da grade de correção do exame.

Enfim, há muito o que fazer. Viu-se que, para o português do Brasil, o PLN ainda engatinha na produção de recursos linguísticos e ferramentas de amplo uso, e para impulsionar a área, nós, linguistas, somos fundamentais. Este trabalho produziu muito recurso linguístico, que era preocupação constante durante sua realização, no entanto, não chegou-se ao objetivo de classificar automaticamente textos nos seis níveis avaliados no exame Celpe-Bras por falta de *corpora*. O ensino de PLA, apesar de já ter trilhado um longo caminho e administrar um exame que, a cada ano, fica maior em termos de aplicação e examinandos, parece ainda não ter dialogado com os estudos de *corpora* empreendidos, de longa data, em línguas estrangeiras como o inglês, especialmente no âmbito da Linguística de Corpus, geralmente associada pelos pesquisadores de Ensino apenas a “pesquisas estatísticas” menores ou de pouca valia frente às grandes questões da área. Mostrou-se, aqui, que esse investimento vale a pena e que pode trazer resultados produtivos.

Cabe, ainda, ressaltar que toda a análise aqui empreendida foi baseada em textos já corrigidos por humanos e já classificados por eles nos níveis pré-determinados. Em momento algum essa avaliação foi questionada. Foi a partir dela que todos os testes, refinamentos, medidas e escores estatísticos foram realizados, tendo como índice máximo de confiabilidade o quanto os dados da pesquisa se aproximavam ou se encaixavam nas medidas humanas dadas pela avaliação do exame Celpe-Bras. Ocorre que a correção humana, como enfatizado em vários momentos deste texto, é absolutamente subjetiva, mas em momento algum pensou-se em não tomar como ponto de partida e chegada a correção e classificação em níveis estabelecidos por corretores humanos. Repletos de subjetividade e idiossincrasias, talvez as avaliações humanas não pudessem ser comparáveis ao método de análise aqui desenvolvido,

totalmente empírico, baseado em dados concretos e observáveis. Talvez não se devesse esperar que ele se aproximasse de resultados provenientes de uma análise subjetiva e holística de corretores humanos. Como mencionado, o “objetivo era o de identificar elementos que fossem capazes de contribuir para diferenciar esses níveis automaticamente, de modo a poder propor uma pré-classificação automática de textos para posterior avaliação humana”, então, a classificação aqui proposta deveria vir antes da correção humana.

Talvez se a classificação prévia dos avaliadores humanos fosse desprezada, se fossem consideradas todas as produções textuais, observados todos os aspectos que foram observados, e só depois os resultados fosse cotejados com os dos humanos, outros fenômenos poderiam ter sido observados e que podem ter sido apagados por conta da necessidade de trabalhar com os níveis pré-determinados. Ficará para um próximo trabalho responder o que teria acontecido se a avaliação dada pela equipe de correção do Celpe-Bras tivesse sido deixada de lado.

Encerro esta dissertação retomando, como na Introdução, a citação de Perini:

Mas dificuldades há, embora nem sempre sejam mencionadas nos livros de introdução à biologia. Uma delas é o ornitorrinco. Esse estranho bicho australiano bota ovos, mas amamenta os filhotes; tem a temperatura do corpo parcialmente dependente da temperatura ambiente e tem pêlos. Será um mamífero, um réptil ou outra categoria qualquer? Isso depende de darmos mais importância a um ou outro dos critérios que definem as classes. De qualquer forma, é necessário admitir que as categorias "mamífero" e "réptil", embora convenientes e úteis, não são perfeitas. A maioria dos animais se coloca claramente em uma ou outra das diversas classes reconhecidas pelos zoólogos; mas há alguns, como o ornitorrinco, que ficam mais ou menos no meio. (PERINI, 1997, p. 39-40)

Este trabalho é, visivelmente, um “ser híbrido”, fruto da junção e da parceria produtiva da Linguística de Corpus, da Avaliação em PLA e das técnicas de PLN. Cada uma das áreas participantes pode sair beneficiada desse longo diálogo. O pesquisador do PLN, dada a emergência de se tratar o português do Brasil, pode encontrar usos concretos para seus trabalhos; o linguista de *corpus* pode perceber um tratamento estatístico da linguagem oriundo do PLN, de condução mais facilitada e confiável do que as tradicionais metodologias da LC; por fim, o pesquisador e professor de PLA, especialmente o professor preocupado com metodologias de avaliação, pode descobrir na exploração em larga escala respostas para perguntas que não encontramos ainda na gramática e nos livros didáticos. A partir das informações de grandes universos textuais e também com a informação sobre o que seja específico, pontual e de menor distribuição, esse pesquisador será capaz de propor estratégias de ensino que reconhecem elementos mais e menos presentes na produção escrita em PLA.

Como dito no início desta dissertação, esse estudo *léxico-estatístico textual*, que abordou de forma inédita as produções textuais elaboradas por estudantes de português que

participaram do exame Celpe-Bras em 2006-1, gerou inúmeros dados. Eles devem ser entendidos aqui como **referência**, e as considerações aqui apresentadas não são conclusivas, apenas reflexivas. Este trabalho, portanto, é apenas um **auxílio** e um primeiro passo. Para que possa ser chamado de **caminho** ou de **referência**, ainda outros tantos passos precisam ser dados, que não dependem mais da autora deste texto, mas daqueles que o lerem.

REFERÊNCIAS

AIJMER, K. **Discourse Particles**: evidence from a *corpus*. Amsterdã: John Benjamins, 2002.

AMARAL, D. **A perspectiva dos examinadores sobre o uso da grade de avaliação oral do IELTS**. 2011. 172 f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre 2011.

AZEREDO, S. de. **Expressões anunciadoras de paráfrases em manuais acadêmicos de Química**: um estudo baseado em *corpus*. 2007. 222 f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2007.

BAKER, M. Corpus Linguistics and Translation Studies – Implications and Applications. In: BAKER, M.; FRANCIS, M. G.; TOGNINI-BONELLI, E. **Text and Technology**: In Honour of John Sinclair. Amsterdã e Filadélfia: John Benjamins, 1993.

BEAUGRANDE, R.; DRESSLER, W. **Introduction to Text Linguistics**. 1981. Disponível em: <http://beaugrande.com/introduction_to_text_linguistics.htm>. Acesso em: 13 mai 2012.

BERBER SARDINHA, T. **Linguística de corpus**. São Paulo: Manole, 2004.

BIBER, D.; CONRAD, S. **Register, genre and style**. Cambridge: Cambridge, 2009.

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus linguistics**: investigating language structure and use. Nova York: Cambridge University Press, 1998.

BIDERMAN, M. T. C. Léxico e Vocabulário Fundamental. **Alfa**, São Paulo, v. 40, p. 27-46, 1996.

BIDERMAN, M. T. C. **Teoria Linguística**: Linguística Quantitativa e Computacional. Rio de Janeiro: Livros Técnicos e Científicos, 1978.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Manual do examinando ao exame Celpe-Bras 2013/1**. Brasília, 2013. Disponível em: <<http://www.inep.gov.br/celpebras/default.asp>>. Acesso em: 12 mar. 2013.

CANDIDO JR., A. **Criação de um ambiente para o processamento de córpus de português histórico**. 2007. Projeto de Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2007.

COURA-SOBRINHO, J.; DELL'ISOLA, R. L. P. O contrato de comunicação na avaliação de proficiência em língua estrangeira. In: JÚDICE, N.; DELL'ISOLA, R. L. P. **Português – língua estrangeira**: novos diálogos. Niterói: Intertexto, p. 89-102, 2009.

CROSSLEY, S.; MCCARTHY, M.; MCNAMARA, D. A linguistic analysis of simplified and authentic texts. **The Modern Language Journal**, v. 91, n. 1, p. 15-30, 2007.

CROSSLEY, S.; MCNAMARA, D. Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. **Journal of Research in Reading**, v. 35, n. 2, p. 115-135, 2012.

DAMAZO, L. O. **A modalização na produção de textos em português como língua estrangeira**. 2012. 220 f. Dissertação (Mestrado em Letras) – Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2012.

DIAS DA SILVA, B. **A face tecnológica dos estudos da linguagem**: o processamento automático de línguas naturais. 1996. 274 f. Tese (Doutorado em Linguística) – Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 1996.

DIAS DA SILVA, B. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 103-138, junho de 2006.

DINIZ, L. A. R.; ZOPPI-FONTANA, M. G. Política linguística no MERCOSUL: o caso do certificado de proficiência em língua portuguesa para estrangeiros (Celpe-Bras). In: HORA, D. D. **Língua(s) e Povos**: Unidade e Diversidade. João Pessoa: Ideia, p. 150-156, 2006.

EVERS, A.; ALLE, C. M. O.; MARCOLIN, P. Causalidade expressa via conectores em Química, Física e Pediatria: um estudo exploratório. In: XX Salão de Iniciação Científica da UFRGS, 2008, Porto Alegre. **Caderno de Resumos do XX Salão de Iniciação Científica da UFRGS, XVII Feira de Iniciação Científica e III Salão UFRGS Jovem**. Porto Alegre: UFRGS, 2008.

EVERS, A.; FINATTO, M. J. B.; PASQUALINI, B. F. Corpus de Aprendizes de Português como Língua Estrangeira (PLE): Compilação Inicial e Primeiros Resultados. In: 18a INPLA – Intercâmbio de Pesquisa em Linguística Aplicada, 2011, São Paulo, **Caderno de Resumos do 18a INPLA**, p. 148-149, 2011.

FERRIS, D. **Treatment of error in second language student writing**. Ann Arbor: University of Michigan Press, 2002.

FILLMORE, C. J. "Corpus linguistics" or "Computer-aided armchair linguistics". In: **Proceedings of Nobel Symposium**: Directions in corpus linguistics. Estocolmo: Jan Svartvik, p. 35-60, 1991.

FINATTO, M. J. B.; AZEREDO, S.; CREMONESE, L. O vocabulário na redação de vestibular: do enfoque estatístico às especificidades da enunciação. In: UFRGS/COPERSE. (Orgs.). **A Redação no Vestibular**: do leitor ao produtor do Texto. Porto Alegre: Editora da UFRGS, p. 95-108, 2008.

FINATTO, M. J. B.; EVERS, A.; ALLE, C. M. O. Do uso de expressões de causalidade como um elemento caracterizador do gênero textual artigo científico. In: V SIGET - Simpósio Internacional de Estudos de Gêneros Textuais, 2009, Caxias do Sul. **Anais... SIGET**. Caxias do Sul: Editora da UCS, 2009.

FINATTO, M. J. B.; EVERS, A.; ALLE, C. M.; ALENCAR, M. C. Das terminologias às construções recorrentes: um percurso de estudos sobre linguagens especializadas. **Ikala Revista de Lenguaje y Cultura**, Antioquia, v. 15, p. 223-258, 2010.

FLESCHE, R. A new readability yardstick. **Journal of Applied Psychology**, 32, p. 221-233, 1948.

FORTES, G. **Celpe-Bras sem segredos**. São Paulo: HUB Editorial, 2012.

GOMES, M. S. **A complexidade de tarefas de leitura e produção escrita no exame Celpe-Bras**. 2009. 109 f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

GRAESSER, A. C.; GERNSBACHER, M. A.; GOLDMAN, S. (Eds.). **Handbook of discourse processes**. Nova Jersey: Taylor & Francis e-Library, 2008.

GRAESSER, A. C.; McNAMARA, D.; LOUWERSE, M.; CAI, Z. Coh-Metrix: Analysis of text on cohesion and language. **Behavior Research Methods, Instruments, & Computers**, v. 36, n. 2, p. 193-202, 2004.

GRANGER, S. A bird's-eye view of learner *corpus* research. In: GRANGER, S.; HUNG, J.; PETCH-TYSON, S. **Computer learner corpora, Second language acquisition and Foreign language teaching**. Amsterdã: John Benjamins, p. 3-33, 2002.

GRANGER, S. The Learner Corpus: A Revolution in Applied Linguistics. **English Today**, v. 10, n. 3, p. 25-33, julho de 1994.

GRIES, S. Th. **Corpora in cognitive linguistics**: Corpus-based approaches to syntax and lexis, 1–18. Berlim, Heidelberg, Nova York: Mouton de Gruyter, 2006.

HALLIDAY, M. A. K.; HASAN, R. **Language, Context, and Text**: Aspects of language in a social-semiotic perspective. Londres: Oxford University Press, 1989.

HAWKINS, J. A.; BUTTERY, P. Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In: TAYLOR, L.; WEIR, C. J. (Orgs.). **Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment**, p. 158-175. Cambridge: Cambridge University Press, 2009.

HOEY, M. **Patterns of lexis in text**. Londres: Oxford University Press, 1991.

HOFFMANN, L. Possibilidades de aplicação e aplicação atual de métodos estatísticos na pesquisa de linguagens especializadas. Tradução: Leonardo Zilio. **Cadernos de Tradução**, Porto Alegre, v. 20, p. 61-76, junho de 2007.

HULSTIJN, J. **Linking L2 proficiency to L2 acquisition: opportunities and challenges of profiling research**. 2010. Disponível em: <<http://eurosla.org/monographs/EM01/233-238Hulstijn.pdf>>. Acesso em: 12 out 2012.

HYMES, D. H. On Communicative Competence. In: PRIDE, J. B.; HOLMES, J. (Org.). **Sociolinguistics**. Harmondsworth: Penguin, 1972.

JARVIS, S.; GRANT, L.; FERRIS, D. Exploring multiple profiles of highly rated learner compositions. **Journal of Second Language Writing**, v. 12, n. 4, p. 377-403, 2003.

KOCH, I. G. V. Linguística textual: quo vadis? **D.E.L.T.A**, São Paulo, v. 17, 2001.
Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502001000300002&lng=en&nrm=iso>. Acesso em: 09 mar. 2013.

KOCH, I. G. V.; TRAVAGLIA, L. C. **A coerência textual**. 17 ed. São Paulo: Contexto, 2009.

MARTINS, R. O pecado original da linguística computacional. **Alfa**, São Paulo, v. 55, n. 1, p. 287-307, 2011.

MARTINS, T. B. F.; GHIRALDELO, C. M.; NUNES, M. G.; OLIVERIA JÚNIOR, O. N. Readability formulas applied to textbooks in Brazilian-Portuguese. **Relatório Técnico**, São Carlos: ICMC/USP, 1996.

MCNAMARA, D.; LOUWERSE, M.; GRAESSER, A. Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. **Grant Proposal**. 2002. Disponível em: <<ftp://129.219.222.66/pdf/IESproposal.pdf>>. Acesso em: 30 fev 2012.

MCNAMARA, D.S., KINTSCH, E., BUTLER-SONGER, N., & KINTSCH, W. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. **Cognition and Instruction**, 14, p. 1-43, 1996.

PASQUALINI, B. F. **Leitura, tradução e medidas de complexidade textual em contos da literatura para leitores com nível de letramento básico**. 2012. 159 f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

PERINI, M. A. Sobre língua, linguagem e Linguística: uma entrevista com Mário A. Perini. **ReVEL**, v. 8, n. 14, p. 1-12, 2010.

PERINI, M. A. **Sofrendo a gramática**. São Paulo: Ática, 1997.

REHM, G.; USZOREIT, H. **A língua portuguesa na era digital**. Berlin: META-NET, 2012.

ROSA, J. L. G. **Fundamentos da inteligência artificial**. Rio de Janeiro: LTC, 2011.

SÁNCHEZ, A.; CANTOS, P. Predictability of word forms (types) and lemmas in linguistic corpora: a case study based on the analysis of the CUMBRE corpus – an 8 million word corpus of contemporary Spanish. **International Journal of Corpus Linguistics**, v. 2, n. 2, p. 259-280, 1996.

SAUSSURE, F. **Curso de linguística geral**. 27 ed. Tradução: Antônio Chelini, José Paulo Paes e Izidoro Blikstein. São Paulo: Cultrix, 2006.

SCARTON, C.; ALMEIDA D. M.; ALUISIO, S. **Coh-Metrix-Port**. Projeto de Pesquisa. 2009. Disponível em: <<http://caravelas.icmc.usp.br:3000/>>. Acesso em: 13 ago. 2010.

SCARTON, C.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o português. **LinguaMática**, v. 1, n. 2, p. 45-62, 2010.

SCHLATTER, M.; GARCEZ, P. D. M.; SCARAMUCCI, M. O papel da interação na pesquisa sobre aquisição e uso de língua estrangeira: implicações para o ensino e para a avaliação. **Letras de Hoje**, v. 39, n. 3, p. 345-378, 2004.

SCHLATTER, M.; GARCEZ, P. M. Línguas adicionais (espanhol e inglês). In: Rio Grande do Sul, Secretaria de Estado da Educação, Departamento Pedagógico. (Orgs.). **Referenciais curriculares do Estado do Rio Grande do Sul: linguagens, códigos e suas tecnologias**. Porto Alegre: Secretaria de Estado da Educação, Departamento Pedagógico, v. 1, p. 127-172, 2009.

SCHOFFEN, J. R. **Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no exame Celpe-Bras**. 2009. 192 f. Tese (Doutorado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

SEPÚLVEDA-TORRES, L. **Escrita científica em português por hispanofalantes: recursos linguístico-computacionais baseados em métodos de alinhamento de textos paralelos**. 2010. Projeto de Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2010.

SHEPHERD, T. M. G. Corpora de aprendiz de língua estrangeira: um estudo de n-gramas. **Veredas** (UFJF. On-line), v. 2, p. 100-116, 2009.

SIDI, W. **Níveis de proficiência em leitura e escrita de falantes de espanhol no exame Celpe-Bras**. 2002. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

SINCLAIR, J. M. **Corpus, concordance, collocation**. Londres: Oxford University, 1991.

SOUZA, J. A. **Tipologia de traços linguísticos de textos do português do Brasil dos séculos XVI, XVII, XVIII e XIX: uma proposta para a classificação automática de gêneros textuais**. 2010. Dissertação (Mestrado em Letras) – Centro de Educação e Ciências Humanas, Universidade Federal de São Carlos, São Carlos, 2010.

STUBBS, M. **Text and corpus analysis: computer assisted studies of language and culture**. Oxford: Blackwell Publishers, 1996.

SWALES, J. M. **Genre analysis: English in academic and research settings**. Cambridge: Cambridge University Press, 1990.

TAGNIN, S. E. O. **O jeito que a gente diz: expressões convencionais e idiomáticas**. São Paulo: Disal, 2005.

TAGNIN, S. E. O.; FROMM, G. CoMAprend: a experiência da construção de um corpus de aprendizes para estudos. **Domínios de Lingu@gem**, Uberlândia, v. 2, n. 2, 2008.

- TURNER, C. E.; UPSHUR, J. A. Rating scales derived from student samples: effects of the scale marker and the student sample on scale content and student scores. **TESOL Quarterly**, v. 36, n. 1, 2002.
- VAN DIJK, T. **Texto y contexto: semantica y pragmatica del discurso**. Madri: Catedra, 1984.
- VIANA, V.; TAGNIN, S. E. O. **Corpora no ensino de línguas estrangeiras**. São Paulo: HUB Editorial, 2011.
- VIEIRA, R.; STRUBE DE LIMA, V. Linguística Computacional: princípios e aplicações. In: IX Escola de Informática da SBC-Sul, 2001, Porto Alegre, **Anais da IX Escola de Informática da SBC-Sul**, p. 27-61, 2001.
- VOLPE NUNES, M. G. **O processamento de línguas naturais: para quê e para quem?** 2008. Notas Didáticas – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2008.
- WEIGLE, S. C. **Assessing Writing**. Cambridge: Cambridge University Press, 2002.
- WIDDOWSON, H. G. **O ensino de línguas para a comunicação**. Campinas: Pontes, 1991.
- WITTEN, I.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. São Francisco: Elsevier, 2005.
- YUQI, S. **A produção de hedges por falantes brasileiros de português e aprendizes chineses de LA**. 2011. Dissertação (Mestrado em Letras) – Faculdade de Letras, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2011.

ANEXO I – TAREFAS E TEXTOS-BASE

Tarefa I – FUNDAÇÃO DARCY RIBEIRO

ENUNCIADO DA TAREFA

Você vai assistir duas vezes a uma entrevista com Tatiana Memória, Presidente da Fundação Darcy Ribeiro (Documentário *O Povo Brasileiro*, Superfilmes, 2000), podendo fazer anotações enquanto assiste. Imagine que você tenha sido convidado/a para fazer um **texto de apresentação** da Fundação Darcy Ribeiro, para ser publicado em um guia sobre centros culturais do Rio de Janeiro. Seu texto deverá conter informações sobre **Darcy Ribeiro** e sobre a **criação e objetivos da Fundação**.

TEXTO-BASE DA TAREFA: VÍDEO

Informações escritas no vídeo:

Darcy Ribeiro

Nascido em MG em 1922

Antropólogo, educador e escritor

Capas de livros mostradas:

Diários índios

Maíra

UNB: invenção e descaminhos

Os índios e a civilização

A invenção do Brasil

O povo Brasileiro

Transcrição da fala de Tatiana Memória:

O grande desejo da vida do Darcy era ser imortal. Ele desejava demais a imortalidade. Na impossibilidade de conseguir isso, ele criou em 1996 a Fundação Darcy Ribeiro com o objetivo principal de continuar trabalhando as ideias dele, as obras dele. Ele se considerava um homem de fazimentos e era. Eu acho que poucos intelectuais brasileiros tiveram uma

atuação tão diversificada em tantas áreas e foram capazes de executar tanta coisa, tantas obras quanto ele foi. Ele teve um apoio indispensável pra que isso tudo acontecesse que foi o apoio dos dois governos Brizola. A Fundação foi criada com a intenção de trabalhar principalmente educação, antropologia, as obras literárias dele, manter os livros dele, principalmente, à disposição. Ele não teve nenhuma intenção filantrópica. Ele não pretendeu com a fundação exercer nenhuma atividade filantrópica. O Darcy acreditava principalmente no trabalho, e muito pouco na caridade. Ele achava que o indispensável era dar ao povo trabalho, porque com trabalho o povo teria dignidade e autossuficiência. A Fundação é uma instituição de direito privado, sem fins lucrativos. Ela nunca recebeu valor nenhum em termos de doação. Tudo o que a fundação é hoje, esse prédio em que ela tá instalada, com dois andares, com restaurante, com salão multimídia, tudo isso foi conquistado com o trabalho de uma equipe que acredita realmente que é preciso realizar alguma coisa, que trabalha com uma enorme dedicação. São quase todas elas aqui dentro pessoas que trabalharam com o Darcy depois que ele voltou do exílio.

Tarefa II – FAMÍLIAS ACOLHEDORAS

ENUNCIADO DA TAREFA

Você vai ouvir duas vezes uma entrevista da rádio *Revista CBN* (fevereiro de 2005) com Cláudia Cabral, diretora executiva de uma organização não-governamental (ONG), podendo fazer anotações enquanto ouve. Convencido/a da importância do projeto, escreva uma **carta aberta** aos moradores de seu bairro, **estimulando** o cadastramento de famílias acolhedoras. Sua carta deverá explicar **o que é** e **como funciona** o projeto e **quais são** as motivações das famílias que dele participam.

TEXTO-BASE DA TAREFA: ÁUDIO

Rodrigo: Você ouvinte do Revista CBN sabe do número de crianças e adolescentes com o que é chamada situação de risco social ou pessoal. São jovens e crianças que por algum motivo não podem viver com a família. Bom, numa situação como essa, num primeiro momento nós pensaríamos que a solução seria um orfanato, até que uma família se interessasse pela adoção. Mas há uma alternativa muito interessante pra isso: são as famílias acolhedoras. Se você quer saber exatamente como elas funcionam, eu convido você a ouvir conosco aqui o nosso papo com a Cláudia Cabral, ela que é diretora executiva da ONG Terra dos Homens. Cláudia, muito boa tarde, como vai?

Cláudia: Boa tarde, Rodrigo, tudo bem.

Rodrigo: O Cláudia, eu queria começar te pedindo pra explicar exatamente o que é isso, como ela funciona?

Cláudia: Rodrigo, é assim, o acolhimento familiar é um tipo de alternativa pra criança assim que precisa estar afastada da família já difundida no mundo todo. É quando uma criança está numa situação de risco o conselho tutelar detecta a situação, precisa tirar a criança da família, e em vez de colocar ela numa instituição, já teria uma lista de famílias cadastradas, que são famílias acolhedoras, ou famílias guardiãs, e essas famílias ficam digamos que em contato com uma equipe, que é uma equipe do poder público, que acompanha a colocação dessas crianças nessas famílias por um tempo provisório, sempre em contato com a família de origem até que ela volte a ter o contato com a sua família quando reverter a situação inicial.

Rodrigo: Então quer dizer que não é adoção, é por um período determinado, enfim, até que a família original possa receber de volta essa criança ou esse adolescente, né?

Cláudia: É, ela fica provisoriamente, ela tem consciência disso, ela faz isso por opção, ela abre um espaço na sua casa, normalmente, tem muitas famílias acolhedoras que já receberam várias crianças, e essas famílias têm consciência de que o papel delas é apoiar a criança por aquele período.

Rodrigo: O Cláudia, o que a família acolhedora ganha acolhendo essa criança, esse adolescente, em casa?

Cláudia: O projeto prevê sempre um subsídio financeiro de apoio pra educação da criança, e essas famílias elas têm diversas motivações, né cada caso é um caso, mas na maioria das vezes são pessoas que gostam, né, de uma casa cheia, gostam de afiliação, de crianças, de educação, e têm sempre esse espaço aberto. E é claro que tem sempre um lado de muita solidariedade.

Tarefa III – VOCÊ SABE O QUE ESTÁ COMENDO?

ENUNCIADO DA TAREFA

Com base na reportagem da *Revista ÉPOCA*, de 25 de julho de 2005, escreva um *e-mail* a uma amiga que consome apenas alimentos naturais, **alertando-a** sobre os mitos a respeito desse tipo de alimentação.

TEXTO-BASE DA TAREFA: REPORTAGEM

Tarefa III

Você sabe o que está comendo?

Carne branca ou vermelha? Pão ou bolacha?

Na busca obsessiva por alimentos saudáveis, o consumidor é vítima de vários mitos

LIA BOCK

Bolacha cream cracker com queijo branco é uma comidinha leve, de dieta, certo? Talvez não. Cinco bolachas dessas têm três vezes mais gordura saturada que um pão francês. E uma fatia de queijo branco, um dos ícones da alimentação light, tem três vezes mais colesterol que a mesma porção de requeijão integral. Pouca gente imaginaria. Mas esses exemplos são, na verdade, apenas uma amostra da enorme vala que separa o que os nutricionistas sabem sobre alimentação do senso comum. A dica da voz, o anúncio de televisão e até mesmo interesses políticos se misturam com pesquisas sérias e confundem o consumidor, ávido por receitas para emagrecer e viver mais. Essa mistura de interesses se transformou em um grande refogado de mitos.

DIGA-ME O QUE COMES E EU TE DIREI...

No mundo das neuroses alimentares, não é difícil identificar os tipos abateo, sempre prontos a defender suas convicções. Mas nem sempre eles têm razão.



Amigo dos animais

O que defende

Grande parte dos vegetarianos diz que não come carnes porque, consumindo apenas vegetais, tem uma dieta mais saudável.

O que ele não diz

Carnes são as principais fontes de ferro e de vitamina B12, nutrientes fundamentais para a saúde. Além disso, as proteínas encontradas nas carnes têm melhor valor biológico, em taxas que dificilmente são obtidas consumindo apenas vegetais. Verduras, legumes e frutas também possuem substâncias prejudiciais – de gorduras a compostos cancerígenos.



Viva a natureza

O que defende

À procura pelos chamados alimentos orgânicos cresceu assustadoramente nos últimos anos, assim como a oferta desses produtos nos supermercados. Seus consumidores defendem um mundo sem agrotóxicos.

O que ele não diz

Até não há consenso científico sobre as vantagens nutricionais dos alimentos orgânicos. Muitos dos benefícios desse tipo de dieta podem ser obtidos com uma higienização adequada dos produtos não-orgânicos – o que permite eliminar a maior parte dos resíduos de defensivos agrícolas.



Não à indústria

O que defende

Muita gente acredita que a indústria alimentícia inclui substâncias impúblicas no preparo de seus produtos para torná-los mais duráveis, saborosos e atraentes. Evitar todo e qualquer alimento industrializado seria a melhor maneira de evitar malefícios à saúde.

O que ele não diz

A população de países industrializados, que consome os conservantes dos alimentos processados, tem uma longevidade invejável. O maior perigo são os produtos ricos em gorduras e açúcares, mas isso vale para qualquer comida, não apenas as industrializadas.



Dieta já

O que defende

Lavados pela preocupação com a obesidade, os integrantes desse grupo passaram a consumir apenas os produtos classificados como diet ou light.

O que ele não diz

Produtos são diet quando sua formulação exclui um determinado ingrediente, como açúcar ou sódio. Ele pode não conter açúcar, mas ser rico em gorduras, portanto, ruim para quem quer perder peso. O ideal é ficar atento ao que ele faz o rótulo com relação ao valor energético ou calórico e à quantidade de gorduras totais.

Época 25 de julho, 2005

Fonte: RGR Nutri Consultoria Nutricional

Tarefa IV – BIG BROTHER CORPORATIVO

ENUNCIADO DA TAREFA

Com o intuito de estimular a discussão sobre a decisão do Tribunal Superior do Trabalho (TST) apresentada na reportagem *da Revista Você S.A.* (junho de 2005), escreva um texto para ser afixado no quadro de avisos de sua empresa, argumentando contra a invasão de privacidade.

TEXTO-BASE DA TAREFA: REPORTAGEM

Tarefa IV

VOCÊ BEM INFORMADO | privacidade

Big Brother CORPORATIVO

O Tribunal Superior do Trabalho deu o aval para as empresas vigiarem e-mails dos empregados. Mas isso ainda vai render muita discussão

POR JOSÉ EDUARDO COSTA



Cameras de segurança nos corredores, crachá eletrônico, controle de ligações. Se você olhar à sua volta, perceberá que algumas organizações viraram uma espécie de Big Brother. Elas ficam de olho em cada movimento dos funcionários. No mês passado, o Tribunal Superior do Trabalho (TST), em Brasília, abriu um precedente ainda maior em favor das corporações. Elas foram autorizadas a vasculhar o correio eletrônico de todo mundo do escritório. A decisão partiu do julgamento de um processo envolvendo o HSBC Seguros e um de seus empregados, que foi demitido por justa causa, em maio de 2000, depois de enviar, usando o e-mail da empresa, uma mensagem aos colegas com fotos de mulheres nuas.

"Embora poucas companhias admitam, hoje é comum vasculhar e-mails", diz Paulo Perez, gerente de engenharia de segurança da Open Communication Security, especializada em softwares de segurança corporativa. Os motivos apontados pelas organizações para tanto controle são os seguintes:

- a) averiguar se o empregado está sendo improdutivo;
- b) examinar se há mensagem com anexos do tipo ".exe", que podem conter vírus ou programas sem licença;
- c) constatar se o funcionário não está visitando sites inseguros, que não tenham relação com sua atividade profissional;
- d) checar e-mails contendo informação sigilosa da empresa (...)

Há quem veja nisso um atentado à privacidade. "Esse é um direito fundamental garantido pela Constituição a todo ser humano. A pessoa não deixa de ser cidadã por que está em seu ambiente de trabalho", rebate o advogado paraense Caio Túlio Vianna, autor do livro *Fundamentos de Direito Penal Informático* (Editora Forense).

Algumas organizações têm procurado uma solução menos hostil. Uma prática comum nos Estados Unidos, que vem ganhando cada vez mais adeptos por aqui, é a assinatura de um termo de compromisso no ato da contratação. Nele, o funcionário se compromete a usar seu e-mail profissional para fins somente de trabalho. E nada mais. □



ANEXO II – GRADES DE CORREÇÃO

GRADE DE CORREÇÃO DA TAREFA I

ENUNCIADO

Você vai assistir duas vezes a uma entrevista com Tatiana Memória, Presidente da Fundação Darcy Ribeiro (Documentário *O Povo Brasileiro*, Superfilmes, 2000), podendo fazer anotações enquanto assiste. Imagine que você tenha sido convidado/a para fazer um **texto de apresentação** da Fundação Darcy Ribeiro, para ser publicado em um guia sobre centros culturais do Rio de Janeiro. Seu texto deverá conter informações sobre **Darcy Ribeiro** e sobre a **criação e objetivos da Fundação**.

RESPOSTA ESPERADA

Gênero discursivo: formato do texto, propósito e interlocutor: o candidato deve escrever um texto de apresentação para um guia. O texto deverá conter informações sobre:

- Darcy Ribeiro;
- A criação e objetivos da Fundação.

Conteúdo informativo:

1. Darcy Ribeiro: antropólogo, educador e escritor, nasceu em MG no ano de 1922;
2. Darcy Ribeiro desejava a imortalidade e por isso criou a Fundação Darcy Ribeiro em 1996.
3. Objetivos da Fundação: Continuidade ao trabalho de Darcy e suas ideias de educação e antropologia, além de manter seus livros à disposição.
4. Darcy Ribeiro acreditava no trabalho e muito pouco na caridade, pois com o trabalho o povo teria dignidade e autossuficiência. Não acreditava em filantropia.

Informações extras:

- Livros publicados por Darcy Ribeiro: Maíra, Diários índios, Os índios e a civilização, A fundação do Brasil, O Povo Brasileiro e UNB- Invenção e descaminhos;

- Darcy teve apoio dos dois governos de Brizola (para a criação de sua Fundação);
- É uma Fundação de direitos privados (não recebe doações);
- Tudo o que existe na Fundação foi conquistado com o trabalho de uma equipe, sendo que a maioria deles trabalhou com Darcy depois que ele voltou do exílio.

TAREFA I - FUNDAÇÃO DARCY RIBEIRO						
NÍVEL	5 - Avan Super	4 - Avançado	3 - Inter Super	2 - Intermediário	1 - Básico	0 - Iniciante
GÊNERO: Formato: Texto de apresentação para guia (com título ou não) Interlocutor: Público em geral.	Adequado	Adequado	Adequado (Pode apresentar, também, interferência do gênero notícia, desde que o formato geral do texto seja de apresentação).	Adequado/Parcialmente adequado (formato híbrido: pode começar com data e nome de outra cidade, mas resgata o texto de apresentação).	Parcialmente Adequado (Pode apresentar formato híbrido como no nível 2)	Inadequado (se apresentar o formato de outro gênero. EX: carta)
GÊNERO: Propósito: Texto de apresentação sobre a Fundação Darcy Ribeiro. INFORMAÇÕES: 1, 2, 3, 4 (informações extras/opcionais)	Adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2, 3 e 4 (Mesmo que falte algum detalhe de algum tópico)	Adequado	Adequado OU Parcialmente adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2 e 3 ou 1, 3 e 4 (Pode apresentar alguns equívocos quanto às informações OU apresentar informações incompletas).	Adequado OU Parcialmente adequado: Apresenta Fundação e Darcy Ribeiro com informações 1, 2 e 3 ou 1, 3 e 4 (com frequentes problemas nessas informações OU 2 e 3 - fala apenas sobre a Fundação).	Adequado OU Parcialmente adequado OU Inadequado: Apresenta Fundação e Darcy Ribeiro com muitas (ou algumas) informações, porém equivocadas.	Inadequado Apresenta Fundação e Darcy Ribeiro com muitas (ou algumas) informações, porém equivocadas. OU Não apresenta informações contidas no vídeo.
CLAREZA E COESÃO	Texto muito bem desenvolvido com clareza e coesão	Texto bem desenvolvido com clareza e coesão, com deslizes na construção da coesão	1. Texto bem desenvolvido com poucos problemas de clareza e coesão OU 2. Texto pouco desenvolvido com clareza e coesão	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão (É necessário um mínimo de articulação do texto)	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão	Texto pouco desenvolvido com muitos problemas graves de clareza e coesão
ADEQUAÇÃO LEXICAL E GRAMATICAL	Raras inadequações e/ou interferências da língua materna	Algumas inadequações e/ou interferências da língua materna	1. Inadequações e/ou interferências da língua materna frequentes no uso de estruturas mais complexas e algumas nas elementares OU 2. Poucas inadequações e/ou interferências da língua materna	Inadequações e/ou interferências da língua materna frequentes no uso de estruturas complexas e elementares	Inadequações e/ou interferências da língua materna mais frequentes	Muitas inadequações e/ou interferências da língua materna

AJUSTES (observações)

Propósito: Embora o propósito do texto seja apresentar a Fundação (e para isso explicar quem foi Darcy Ribeiro), também será considerado adequado o texto que iniciar focalizando Darcy Ribeiro (biografia) para, depois, apresentara Fundação.

Informação 2: o candidato poderá deixar implícita a questão da imortalidade de Darcy, utilizando outra palavra/ expressão: “Darcy criou a Fundação para que seguissem sua obra; com o objetivo principal de trabalhar as obras dele”.

GRADE DE CORREÇÃO DA TAREFA II

ENUNCIADO

Você vai ouvir duas vezes uma entrevista da rádio *Revista CBN* (fevereiro de 2005) com Cláudia Cabral, diretora executiva de uma organização não-governamental (ONG), podendo fazer anotações enquanto ouve. Convencido/a da importância do projeto, escreva uma **carta aberta** aos moradores de seu bairro, **estimulando** o cadastramento de famílias acolhedoras. Sua carta deverá explicar **o que é** e **como funciona** o projeto e **quais são** as motivações das famílias que dele participam.

RESPOSTA ESPERADA

Gênero discursivo: formato do texto, propósito e interlocutor: o candidato deve escrever uma carta aos moradores do bairro. A carta deverá estimular o cadastramento de famílias acolhedoras,

- explicando o que é o projeto;
- explicando como funciona o projeto;
- explicitando as motivações das famílias que dele participam.

Conteúdo informativo:

1. O que é?

Projeto (alternativo) de acolhimento familiar: famílias acolhedoras/guardiãs (para crianças em risco social e pessoal)

2. Como funciona?

2a. conselho tutelar ou ONG ou órgão público: detecta/encaminha crianças - cadastra famílias acolhedoras

2b. situação provisória (informação facultativa: acompanhamento do processo)

3. Motivações:

3a. subsídio financeiro (não importa de onde vem)

3b. solidariedade

3c: gosta de criança, educação, filiação, casa cheia

TAREFA II – FAMÍLIAS ACOLHEDORAS						
NÍVEL	5 - Avan Super	4 - Avançado	3 - Inter Super	2 - Intermediário	1 - Básico	0 - Iniciante
GÊNERO: Formato: carta aberta Interlocutor: moradores do bairro Autor: ONG, órgão, pessoa que ouviu entrevista, família acolhedora	Adequado Abertura: vocativo e fechamento OU Abertura: título Apresenta marcas de interlocução no texto (autor e destinatário)	Adequado OU Parcialmente adequado Apresenta abertura OU fechamento (idem)	Adequado OU Parcialmente adequado (idem)	Adequado OU Parcialmente adequado (sem abertura) Com poucas marcas de interlocução no texto.	Adequado OU Parcialmente Adequado Não atende a proposta Sem interlocutor ou marcas de interlocução Destinatário inadequado	Parcialmente adequado OU Inadequado
GÊNERO: Propósito: Estimular famílias acolhedoras	Adequado	Adequado	Adequado	Adequado OU Parcialmente adequado	Adequado/ parcialmente adequado/inadequado	Parcialmente adequado OU inadequado
INFORMAÇÕES 1, 2, 3 (informações facultativas)	1, 2, 3 Adequado	Adequado 1, 2, 3 (em relação à informação 3, apresentar pelo menos o item 3b + qualquer outro)	1. famílias acolhedoras 2. período provisório 3. item 3b + qualquer outro Algumas informações equivocadas, desde que não as acima.	Adequado ou 1 e 2: idem ao nível 3 3. mencionar uma motivação, mesmo que implícita (idem)	Adequado/ Parcialmente adequado/Inadequado Várias informações incompletas e/ou inadequadas	Inadequado Muitas informações incompletas e/ou inadequadas
CLAREZA E COESÃO	Texto muito bem desenvolvido com clareza e coesão	Texto bem desenvolvido com clareza e coesão deslizes na construção da coesão	1. Texto bem desenvolvido com poucos problemas de clareza e coesão OU 2. Texto pouco desenvolvido com clareza e coesão	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão (É necessário um mínimo de articulação do	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão	Texto pouco desenvolvido com muitos problemas graves de clareza e coesão
ADEQUAÇÃO LEXICAL E GRAMATICAL	Raras inadequações e/ou interferências da língua materna	Algumas inadequações e/ou interferências da língua materna	1. Inadequações e/ou interferências da língua materna frequentes no uso de estruturas mais complexas e algumas nas elementares OU 2. Poucas inadequações e/ou interferências da língua materna	Inadequações e/ou interferências da língua materna frequentes no uso de estruturas complexas e elementares	Inadequações e/ou interferências da língua materna mais frequentes	Muitas inadequações e/ou interferências da língua materna

AJUSTES (observações)

Propósito:

Adequado: estimular explicitamente, com a manutenção do propósito e, conseqüentemente, com marcas de interlocução ao longo do texto

Parcialmente adequado: não há manutenção do propósito ao longo do texto

Motivação:

3b (solidariedade): para avançado superior mencionar explicitamente; para os demais níveis, também considerar este item de forma implícita, ou seja, por meio da utilização de expressões como "ajudar...", "contribuir..." e outras

3c: pode ser qualquer "gostar de"

GRADE DE CORREÇÃO DA TAREFA III**ENUNCIADO**

Com base na reportagem da *Revista ÉPOCA*, de 25 de julho de 2005, escreva um *e-mail* a uma amiga que consome apenas alimentos naturais, **alertando-a** sobre os mitos a respeito desse tipo de alimentação.

RESPOSTA ESPERADA

Gênero discursivo: formato do texto, propósito e interlocutor: o candidato deve escrever um e-mail para uma amiga, alertando-a sobre os mitos a respeito de alimentos naturais.

Conteúdo informativo:**A) Informações imprescindíveis:**

- mito referente aos alimentos orgânicos e seus contra-argumentos: 1) não há consenso científico das vantagens de seu consumo E 2) boa higienização dos produtos não-orgânicos garante eliminação de resíduos tóxicos.
- mito referente ao consumo de produtos industrializados e seus contra-argumentos: 3) população de países em que há o consumo de produtos industrializados tem uma longevidade invejável E 4) gorduras e açúcares, os grandes perigos para a saúde, são encontrados não apenas em produtos industrializados, mas em qualquer tipo de alimento.
mito sobre a não ingestão de carnes e seus contra-argumentos: 5) Verduras, legumes e frutas também possuem substâncias prejudiciais - de gordura a compostos cancerígenos.

B) Informações periféricas (podem aparecer ou não):

- anúncios de TV, dicas da vovó, pesquisas científicas confundem o consumidor.
mito sobre a não ingestão de carnes e seus contra-argumentos: carnes possuem ferro e vitamina B12, além de proteínas com melhor valor biológico.
mito sobre o consumo de produtos light ou diet e seus contra-argumentos: mesmo sem açúcar, podem ser ricos em gordura (o ideal é ficar atento ao que diz nos rótulos dos produtos).

TAREFA III – VOCÊ SABE O QUE ESTÁ COMENDO?						
NÍVEL	5 - Avançado Superior	4 - Avançado	3 - Intermediário Superior	2 - Intermediário	1 - Básico	0 - Iniciante
GÊNERO: Formato e interlocutor e-mail para amiga que somente come alimentos naturais	Adequado	Adequado	Adequado Ou Parcialmente adequado(o interlocutor é amiga cuja única preocupação é a dieta ou alimentação vegetariana OU não deixa explícito que amiga é essa.)	Adequado Parcialmente adequado(idem)	Parcialmente Adequado ou Inadequado(outro interlocutor)	Inadequado
GÊNERO: propósito Alertar apenas alimentos naturais sobre o consumo de explicitamente	Adequado	Adequado	Adequado	Adequado OU Parcialmente adequado (alerta implicitamente)	Adequado/Parcialmente adequado OU Inadequado (texto informativo)	Inadequado
INFORMAÇÕES Apresenta no mínimo 3 contra-argumentos dentre os cinco principais OU Apresenta somente os dois contra-argumentos específicos a respeito dos alimentos orgânicos	Adequado	Adequado	Adequado OU Parcialmente adequado(apresenta apenas 2 contra-argumentos dentre os 5)	Adequado Ou Parcialmente adequado(apresenta somente 1 ou 2 contra-argumentos quaisquer OU apresenta informações equivocadas)	Parcialmente adequado OU Inadequado(apenas tangencia as informações do texto- base OU não faz referência ao texto- base)	Parcialmente adequado OU Inadequado(idem)
CLAREZA E COESÃO	Texto muito bem desenvolvido com clareza e coesão	Texto bem desenvolvido com clareza e coesão, com deslizes na construção da coesão	1. Texto bem desenvolvido com poucos problemas de clareza e coesão OU 2. Texto pouco desenvolvido com clareza e coesão	Texto pouco/mal desenvolvido com problemas de clareza e coesão (É necessário um mínimo de articulação do texto)	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão	Texto pouco desenvolvido com muitos problemas graves de clareza e coesão
ADEQUAÇÃO LEXICAL E GRAMATICAL	Raras inadequações e/ou interferências da língua materna	Algumas inadequações e/ou interferências da língua materna	1. Inadequações e/ou interferências da língua materna frequentes no uso de estruturas mais complexas e algumas nas elementares OU. Poucas materna interferências inadequações e/ou 2 da língua	Inadequações e/ou interferências da língua materna frequentes no uso de estruturas complexas e elementares	Inadequações e/ou interferências da língua materna mais frequentes	Muitas inadequações e/ou interferências da língua materna

AJUSTES (observações)

Formato: Parcialmente adequado: e-mail sem abertura. Inadequado: e-mail sem abertura e sem fechamento.

Propósito: Parcialmente adequado: normalmente os textos que alertam implicitamente têm problemas na interlocução, que acontece apenas nos primeiros e últimos parágrafos. Inadequado: não há marca de interlocução.

GRADE DE CORREÇÃO DA TAREFA IV**ENUNCIADO**

Com o intuito de estimular a discussão sobre a decisão do Tribunal Superior do Trabalho (TST) apresentada na reportagem *da Revista Você S.A.* (junho de 2005), escreva um texto para ser afixado no quadro de avisos de sua empresa, argumentando contra a invasão de privacidade.

RESPOSTA ESPERADA

Gênero discursivo: formato do texto, propósito e interlocutor: o candidato deve escrever um texto para um quadro de avisos de uma empresa, argumentando contra a invasão de privacidade.

Conteúdo informativo: Contextualizar a questão (falar da invasão de privacidade e/ou das empresas estarem controlando os funcionários, vasculhando e-mails, etc.) e mencionar a decisão do TST.

TAREFA IV – BIG BROTHER CORPORATIVO						
NÍVEL	5 - Avan Super	4 - Avançado	3 - Inter Super	2 - Intermediário	1 - Básico	0 - Iniciante
GÊNERO: formato e interlocutor Texto argumentativo para quadro de avisos	Adequado	Adequado	Adequado ou Parcialmente Adequado (Carta para alguém do trabalho)	Adequado ou Parcialmente Adequado (Carta para alguém do trabalho)	Adequado ou Parcialmente Adequado ou Inadequado	Inadequado
GÊNERO: propósito Posiciona-se e argumenta contra a invasão de privacidade	Adequado	Adequado	Adequado ou Parcialmente adequado (menciona ser contra, mas não argumenta)	Adequado ou Parcialmente adequado (menciona ser contra, mas não argumenta)	Adequado ou Parcialmente adequado ou Inadequado	Inadequado
INFORMAÇÕES Dizer que as empresas estão invadindo a privacidade dos funcionários e mencionar a decisão do TST	Adequado (Menciona a decisão do TST explicitamente)	Adequado (Faz referência implícita à decisão do TST)	Adequado ou Parcialmente Adequado (Contextualiza a questão, mas não faz referência à decisão do TST)	Adequado ou Parcialmente Adequado (Não contextualiza a questão e/ou não faz referência à decisão do TST - fala somente de privacidade)	Adequado ou Parcialmente Adequado ou Inadequado ou copia muito do texto original	Inadequado
CLAREZA E COESÃO	Texto muito bem desenvolvido com clareza e coesão	Texto bem desenvolvido com clareza e coesão, com deslizes na construção da coesão.	1. Texto bem desenvolvido com problemas de clareza e coesão OU 2. Texto pouco desenvolvido com clareza e coesão	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão (E necessário um mínimo de articulação do texto)	Texto pouco/mal desenvolvido com problemas graves de clareza e coesão	Texto pouco desenvolvido com muitos problemas graves de clareza e coesão
ADEQUAÇÃO LEXICAL E GRAMATICAL	Raras inadequações e/ou interferências da língua materna	Algumas inadequações e/ou interferências da língua materna	1. Inadequações e/ou interferências da língua materna frequentes no uso de estruturas mais complexas e algumas nas elementares OU 2. Poucas inadequações e/ou interferências da língua materna	Inadequações e/ou interferências da língua materna frequentes no uso de estruturas complexas e elementares	Inadequações e/ou interferências da língua materna mais frequentes	Muitas inadequações e/ou interferências da língua materna

ANEXO III – ARQUIVOS ARFF

ARFF – 15 atributos e 2 classes

```
@relation "
1,3,1,3,1,3,3,1,3,1,3,1,1,1,3,reprovado
3,3,3,1,3,1,1,1,3,3,3,3,3,3,3,aprovado

@attribute numero_palavras numeric
3,1,1,3,1,3,3,3,1,1,1,3,1,1,3,aprovado
@attribute numero_pronomes numeric
1,3,1,1,3,1,1,3,3,1,3,1,3,3,3,aprovado
@attribute sentencas_por_paragrafos
numeric
3,3,1,1,3,3,1,1,1,1,1,3,3,3,3,aprovado
1,1,1,1,1,1,3,1,1,1,1,1,3,3,3,aprovado
@attribute silabas_por_palavras numeric
1,3,3,3,1,1,1,1,3,1,1,1,1,1,3,aprovado
@attribute flesch numeric
1,1,1,1,1,1,1,1,3,1,1,1,3,1,3,reprovado
@attribute numero_functional_words
numeric
3,1,1,3,1,1,3,1,1,3,1,3,1,1,3,aprovado
3,1,1,3,1,3,3,3,1,3,1,3,3,3,3,aprovado
@attribute hiperonimos_verbos numeric
1,1,1,1,3,1,1,1,1,3,1,3,1,3,3,aprovado
@attribute palavras_antes_verbos numeric
1,3,1,1,3,1,1,1,3,3,3,1,3,1,3,reprovado
@attribute pronomes_pessoais numeric
1,1,3,1,3,1,1,1,1,3,1,3,3,1,3,aprovado
@attribute tipo_token numeric
1,1,3,1,3,1,1,1,3,1,1,3,3,3,3,reprovado
@attribute pronomes_por_sintagmas
numeric
1,3,1,1,3,3,1,1,3,3,3,1,3,1,3,aprovado
1,1,3,3,1,1,1,1,1,3,1,3,1,3,3,aprovado
@attribute numero_operadores_logicos
numeric
3,1,3,3,1,3,1,3,1,3,1,3,1,3,3,aprovado
1,1,3,1,1,1,1,3,1,1,1,3,1,1,3,aprovado
@attribute conectivos_logico_positivos
numeric
1,1,1,1,3,1,3,3,1,3,3,3,1,3,3,aprovado
3,3,3,3,3,3,3,1,3,3,3,3,1,3,3,aprovado
@attribute ambiguidade_substantivos
numeric
1,1,1,1,3,3,1,3,3,3,3,3,3,1,3,aprovado
3,3,1,3,1,3,1,3,1,1,1,1,1,1,3,aprovado
@attribute ambiguidade_adverbios
numeric
3,1,1,3,1,3,3,1,1,1,1,1,1,1,3,aprovado
1,3,1,3,1,3,1,3,3,3,3,1,3,3,3,aprovado
@attribute class {aprovado,reprovado}
1,1,1,1,3,3,3,3,3,1,3,3,1,3,3,aprovado
1,3,1,3,1,3,1,1,3,3,3,1,1,1,3,aprovado
@data
3,1,1,1,3,3,1,3,1,3,1,3,3,1,3,aprovado
1,1,1,3,1,1,3,1,1,3,1,1,3,3,3,aprovado
1,1,1,3,1,3,1,1,1,3,1,3,1,3,3,aprovado
1,3,1,3,1,3,3,3,1,3,3,1,1,3,aprovado
3,3,3,3,1,1,3,1,3,3,3,1,1,1,3,aprovado
3,1,1,1,3,3,1,3,3,1,1,1,1,3,3,aprovado
1,3,1,3,1,3,3,1,1,3,3,3,3,1,3,aprovado
1,3,1,3,1,3,3,1,1,3,3,3,3,1,3,aprovado
3,3,3,1,3,1,1,3,1,1,1,1,1,1,3,aprovado
3,3,3,1,3,1,1,3,1,1,1,1,1,1,3,aprovado
1,3,1,3,3,3,1,1,3,1,3,1,3,1,3,aprovado
3,3,3,3,3,3,3,1,1,3,3,1,1,3,3,aprovado
1,3,1,1,3,3,1,1,3,1,3,1,3,1,3,aprovado
3,3,1,1,3,3,3,1,3,3,1,3,1,1,3,aprovado
3,1,3,3,1,3,3,1,1,3,1,1,3,1,3,aprovado
3,3,3,1,3,3,3,1,3,3,1,1,3,1,3,aprovado
1,1,3,1,3,3,1,1,1,3,1,3,3,3,3,aprovado
1,3,1,3,1,3,1,1,3,1,3,1,3,3,3,aprovado
```

1,3,1,3,1,1,3,1,1,1,3,1,3,1,3,aprovado
 1,1,1,1,3,3,3,1,1,1,3,1,1,3,3,aprovado
 1,1,1,1,3,1,1,1,3,1,3,1,1,1,3,aprovado
 3,3,1,3,1,3,1,1,1,1,3,1,1,3,3,aprovado
 1,1,1,3,1,3,1,1,3,3,3,3,1,1,3,reprovado
 1,3,1,1,3,1,3,1,3,3,3,1,3,1,3,aprovado
 1,1,1,1,3,3,3,1,1,1,1,1,3,1,3,aprovado
 3,1,1,3,1,3,3,3,1,1,1,3,1,1,3,aprovado
 3,3,1,3,1,3,3,3,3,3,1,1,1,3,aprovado
 3,1,1,3,3,3,3,1,1,1,3,1,3,3,aprovado
 3,1,3,3,1,1,3,1,3,3,1,1,3,3,3,reprovado
 1,1,1,3,3,1,1,3,1,1,1,1,1,3,aprovado
 1,1,1,3,1,1,3,1,3,1,1,1,3,1,3,reprovado
 3,1,3,3,3,1,1,1,1,1,1,1,1,3,3,aprovado
 3,1,1,3,1,3,3,3,1,1,1,1,3,3,3,aprovado
 3,1,3,3,1,3,1,3,3,1,1,3,1,3,3,aprovado
 1,1,1,1,3,1,1,1,1,3,3,1,1,1,3,aprovado
 3,3,1,3,1,3,3,3,3,3,3,1,1,3,aprovado
 3,3,3,1,3,3,1,1,3,3,3,3,1,3,3,aprovado
 3,1,3,3,1,3,1,1,3,1,1,3,3,1,1,3,aprovado
 3,3,3,1,3,1,1,3,1,1,3,1,1,3,3,aprovado
 3,3,3,1,3,1,1,3,1,1,3,3,3,3,aprovado
 3,3,3,1,3,1,1,3,1,1,3,3,3,3,aprovado
 3,1,3,3,1,3,1,1,3,1,1,3,1,1,3,reprovado
 3,3,3,1,3,1,1,3,1,1,3,3,3,3,aprovado
 3,3,3,1,3,1,1,3,1,1,3,3,3,3,aprovado
 3,1,3,3,1,3,1,1,3,1,1,3,3,3,3,aprovado
 1,3,1,3,3,3,1,1,3,3,3,1,3,1,3,reprovado
 1,3,1,1,3,1,1,3,1,1,3,1,3,3,3,aprovado
 1,1,1,1,1,1,3,1,1,1,3,1,3,3,3,reprovado
 1,1,1,3,1,3,1,1,1,3,1,3,1,1,3,aprovado
 3,3,3,1,3,3,3,1,3,1,1,1,3,3,3,aprovado
 3,3,3,1,3,3,3,1,3,1,1,1,3,3,3,aprovado
 1,1,1,1,3,1,1,1,1,3,1,3,1,1,3,aprovado
 1,1,1,3,1,1,1,1,1,1,1,3,3,3,3,reprovado
 1,1,3,1,1,1,1,3,1,1,1,3,1,1,3,aprovado
 3,1,1,1,1,1,3,3,1,3,1,3,1,1,3,aprovado
 1,3,1,1,3,1,3,1,1,1,3,3,3,1,3,aprovado
 3,1,1,1,3,1,1,1,1,1,1,1,3,3,aprovado
 1,1,3,3,1,1,1,1,1,3,1,3,1,1,3,aprovado
 1,3,1,3,3,3,3,1,3,3,3,3,3,3,reprovado
 1,3,1,1,3,3,3,1,3,3,3,3,1,3,reprovado
 1,1,1,3,1,3,1,1,1,3,3,3,1,3,aprovado

3,1,3,3,1,1,1,3,1,1,3,1,1,3,aprovado
3,3,3,1,3,3,3,1,3,3,3,1,3,1,3,aprovado
3,1,1,3,1,3,1,3,1,3,1,3,3,3,3,aprovado
3,1,3,3,1,1,3,3,1,1,1,1,1,1,3,aprovado
3,3,3,1,3,1,1,1,3,3,1,3,1,1,3,reprovado
1,1,1,3,1,1,3,3,1,1,1,1,1,3,3,aprovado
3,1,3,3,1,1,1,1,1,1,1,1,3,1,3,aprovado
1,3,3,1,3,3,1,1,3,3,3,3,1,3,aprovado
3,1,3,1,1,1,1,3,1,1,1,3,1,3,3,aprovado
3,1,3,1,1,1,1,1,1,1,1,1,3,3,3,aprovado
1,3,1,1,3,3,1,3,3,1,3,3,1,3,3,aprovado
3,3,1,1,3,3,3,1,3,3,3,3,1,3,aprovado
1,3,3,1,3,3,3,1,3,3,3,1,1,3,3,reprovado
1,1,1,3,1,1,1,1,3,3,1,3,3,1,3,aprovado
1,3,1,3,3,3,3,1,1,1,3,1,3,1,3,reprovado
1,1,1,3,3,3,1,3,1,1,1,3,3,3,3,aprovado
3,3,1,3,1,3,1,3,1,3,1,1,1,1,3,reprovado
1,1,1,1,3,3,3,3,1,3,1,1,3,3,3,aprovado
1,1,1,3,1,3,3,3,3,3,1,3,1,1,3,aprovado

ARFF – 48 atributos e 6 classes (amostra)

```
@relation "  
@attribute texto STRING  
@attribute numero_palavras numeric  
@attribute numero_sentencas numeric  
@attribute numero_paragrafos numeric  
@attribute numero_verbos numeric  
@attribute numero_substantivos numeric  
@attribute numero_adjetivos numeric  
@attribute numero_adverbios numeric  
@attribute numero_pronomes numeric  
@attribute palavras_por_sentencas numeric  
@attribute sentencas_por_paragrafos numeric  
@attribute silabas_por_palavras numeric  
@attribute flesch numeric  
@attribute numero_content_words numeric  
@attribute numero_functional_words numeric  
@attribute frequencia_content_words numeric  
@attribute minimo_frequencia_content_words numeric  
@attribute hiperonimos_verbos numeric  
@attribute incidencia_sintagmas_nominais numeric  
@attribute modificadores_sintagmas numeric  
@attribute palavras_antes_verbos numeric  
@attribute pronomes_pessoais numeric  
@attribute tipo_token numeric  
@attribute pronomes_por_sintagmas numeric  
@attribute numero_e numeric  
@attribute numero_ou numeric  
@attribute numero_se numeric  
@attribute numero_negacoes numeric  
@attribute numero_operadores_logicos numeric  
@attribute todos_conectivos numeric  
@attribute conectivos_adtivos_positivos numeric  
@attribute conectivos_adtivos_negativos numeric  
@attribute conectivos_temporal_positivos numeric  
@attribute conectivos_temporal_negativos numeric  
@attribute conectivos_causal_positivos numeric  
@attribute conectivos_causal_negativos numeric  
@attribute conectivos_logico_positivos numeric  
@attribute conectivos_logico_negativos numeric  
@attribute ambiguidade_verbos numeric  
@attribute ambiguidade_substantivos numeric  
@attribute ambiguidade_adverbios numeric  
@attribute ambiguidade_adjetivos numeric  
@attribute correferencia_argadjovl numeric  
@attribute correferencia_argovl numeric  
@attribute correferencia_stmadjovl numeric  
@attribute correferencia_stmovl numeric  
@attribute correferencia_cwovl numeric
```

@attribute anafora_refanaadj numeric
 @attribute anafora_refana numeric
 @attribute class
 {avancado,intermediario,basico,avancado_superior,intermediario_superior,iniciante}

@data

T01E09.txt,136.0,9.0,3.0,161.765,330.882,58.8235,66.1765,66.1765,15.1111,3.0,2.71429,48.7458,617.647,352.941,212231.0,5798.0,0.4,257.353,0.628571,4.33333,0.0,0.666667,1.89076,14.7059,14.7059,7.35294,7.35294,44.1176,66.1765,29.4118,14.7059,0.0,0.0,29.4118,0.0,29.4118,0.0,4.18182,1.2,0.0,0.625,0.75,0.916667,0.875,0.861111,0.875,0.222222,0.222222,avancado

T01E10.txt,123.0,7.0,4.0,170.732,325.203,81.3008,16.2602,81.3008,17.5714,1.75,2.9863,40.4781,593.496,373.984,140768.0,5300.57,0.111111,308.943,0.473684,4.85714,16.2602,0.890411,2.1395,24.3902,0.0,0.0,0.0,24.3902,65.0406,40.6504,0.0,8.13008,0.0,40.6504,0.0,24.3902,0.0,7.2381,2.16129,0.0,1.1,0.333333,0.380952,0.5,0.47619,0.5,0.571429,0.571429,avancado

T01E11.txt,122.0,12.0,8.0,131.148,385.246,65.5738,32.7869,73.7705,10.1667,1.5,2.84,54.7535,614.754,368.852,217913.0,27248.4,0.153846,311.475,0.394737,4.41667,0.0,0.746667,1.94133,24.5902,8.19672,0.0,0.0,32.7869,65.5738,32.7869,0.0,0.0,0.0,40.9836,0.0,16.3934,0.0,5.0,1.40625,0.0,2.75,0.636364,0.378788,0.636364,0.424242,0.636364,0.0,0.0,avancado
 T01E12.txt,179.0,7.0,3.0,150.838,312.849,67.0391,61.4525,78.2123,25.5714,2.33333,2.84906,36.6655,592.179,346.369,237190.0,3088.43,0.173913,268.156,0.395833,4.14286,5.58659,0.811321,1.62942,22.3464,0.0,5.58659,5.58659,33.5196,72.6257,44.6927,0.0,0.0,0.0,27.933,0.0,27.933,5.58659,4.88889,2.46512,0.0,0.75,0.833333,0.857143,1.0,0.952381,0.833333,0.285714,0.285714,avancado

T01E13.txt,115.0,8.0,2.0,173.913,347.826,78.2609,52.1739,0.0,14.375,4.0,2.8,45.9174,652.174,339.13,285758.0,2346.38,0.2,286.957,0.515152,4.375,0.0,0.72,0.0,69.5652,0.0,0.0,0.0,69.5652,78.2609,52.1739,0.0,8.69565,0.0,26.087,0.0,17.3913,0.0,7.1,1.66667,0.0,1.22222,0.857143,0.75,0.857143,0.785714,0.857143,0.0,0.0,avancado

T02E07.txt,248.0,12.0,3.0,209.677,245.968,64.5161,52.4194,100.806,20.6667,4.0,2.71831,50.1301,572.581,411.29,255082.0,2711.08,0.574468,237.903,0.661017,2.25,16.129,0.767606,1.70858,24.1935,8.06452,4.03226,0.0,36.2903,72.5806,28.2258,4.03226,8.06452,0.0,36.2903,0.0,40.3226,0.0,8.88461,2.30909,0.0,2.625,0.727273,0.621212,1.0,0.878788,0.636364,0.0,0.0833333,avancado

T02E08.txt,206.0,9.0,6.0,165.049,223.301,72.8155,33.9806,155.34,22.8889,1.5,2.94118,46.5465,495.146,480.583,152022.0,4224.22,0.424242,203.883,0.595238,4.22222,38.835,0.862745,3.69857,43.6893,0.0,4.85437,4.85437,53.3981,72.8155,38.835,4.85437,4.85437,0.0,29.1262,0.0,24.2718,4.85437,10.2353,2.04348,0.0,2.46667,0.625,0.555556,0.5,0.444444,0.25,0.22222,1.22222,avancado

T02E09.txt,240.0,15.0,8.0,166.667,270.833,58.3333,54.1667,116.667,16.0,1.875,2.67424,61.985,550.0,437.5,298327.0,4090.27,0.589744,262.5,0.396825,2.4,29.1667,0.825758,1.85185,37.5,8.33333,4.16667,0.0,50.0,95.8333,41.6667,0.0,12.5,4.16667,45.8333,0.0,25.0,0.0,6.7,1.63793,0.0,1.0,0.357143,0.247619,0.642857,0.361905,0.5,0.0666667,0.533333,avancado

T02E10.txt,186.0,13.0,8.0,134.409,284.946,102.151,48.3871,86.0215,14.3077,1.625,3.06604,42.8256,569.892,430.108,186445.0,6204.85,0.73913,252.688,0.617021,4.53846,0.0,0.745283,1.83024,37.6344,5.37634,0.0,0.0,43.0108,80.6452,32.2581,0.0,5.37634,0.0,43.0108,0.0,16.129,0.0,6.28,1.04,0.0,2.15789,0.5,0.294872,0.583333,0.371795,0.583333,0.0,0.0,avancado

T02E11.txt,158.0,14.0,7.0,196.203,278.481,94.9367,18.9873,56.962,11.2857,2.0,2.97849,46.2268,588.608,392.405,154457.0,149751.0,0.607143,272.152,0.627907,1.71429,6.32911,0.795699,1.3247,25.3165,0.0,0.0,0.0,25.3165,63.2911,25.3165,0.0,12.6582,6.32911,25.3165,0.0,25.3165,0.0,7.51613,0.97619,0.0,1.06667,0.153846,0.197802,0.307692,0.263736,0.153846,0.0,0.0,avancado

T02E12.txt,196.0,11.0,6.0,168.367,250.0,107.143,40.8163,61.2245,17.8182,1.83333,2.99099,39.1046,566.327,387.755,197206.0,103540.0,0.4,265.306,0.769231,3.54545,0.0,0.837838,1.17739,35.7143,10.2041,5.10204,0.0,51.0204,86.7347,40.8163,0.0,5.10204,5.10204,56.1224,0.0,45.9184,0.0,6.66667,2.17778,0.0,1.52381,0.2,0.2,0.2,0.290909,0.2,0.181818,0.181818,avancado

T02E13.txt,142.0,9.0,7.0,133.803,253.521,91.5493,63.3803,77.4648,15.7778,1.28571,3.31169,33.8318,542.254,401.408,314340.0,9133.89,0.388889,253.521,0.555556,2.77778,0.0,0.831169,2.1518,42.2535,14.0845,0.0,0.0,56.338,77.4648,42.2535,0.0,0.0,0.0,42.2535,0.0,28.169,0.0,6.42105,1.82857,0.0,1.30769,0.25,0.305556,0.625,0.555556,0.25,0.0,0.0,avancado

T02E14.txt,214.0,14.0,8.0,154.206,285.047,112.15,23.3645,93.4579,15.2857,1.75,2.88618,45.1443,574.766,383.178,171862.0,74510.4,0.612903,266.355,0.701754,4.14286,9.34579,0.788618,1.63961,37.3832,4.6729,4.6729,0.0,46.729,98.1308,37.3832,0.0,4.6729,0.0,65.4206,0.0,46.729,0.0,6.87879,2.32692,0.0,3.0,0.461538,0.318681,0.538462,0.373626,0.384615,0.0,0.0,0.0714286,avancado

T02E15.txt,145.0,6.0,3.0,213.793,282.759,75.8621,34.4828,75.8621,24.1667,2.0,2.80682,38.7693,606.897,344.828,105841.0,527.833,0.366667,268.966,0.410256,2.83333,0.0,0.943182,1.94518,34.4828,6.89655,6.89655,0.0,55.1724,68.9655,20.6897,0.0,6.89655,0.0,34.4828,0.0,27.5862,0.0,6.16129,2.78378,0.0,0.454545,0.2,0.533333,0.6,0.733333,0.0,0.0,0.0,avancado

T02E16.txt,165.0,9.0,9.0,175.758,266.667,84.8485,18.1818,109.091,18.3333,1.0,3.02222,43.5939,545.455,442.424,177219.0,115539.0,0.392857,242.424,0.65,1.66667,12.1212,0.811111,2.72727,30.303,6.06061,0.0,0.0,36.3636,66.6667,24.2424,0.0,0.0,0.0,42.4242,0.0,30.303,0.0,7.37931,2.25581,0.0,3.5,0.375,0.166667,0.375,0.166667,0.375,0.111111,0.111111,avancado

T02E17.txt,170.0,12.0,6.0,182.353,276.471,111.765,41.1765,105.882,14.1667,2.0,2.94231,38.3829,611.765,388.235,181880.0,21161.5,0.777778,229.412,0.923077,1.58333,5.88235,0.759615,2.71493,29.4118,0.0,5.88235,0.0,35.2941,82.3529,29.4118,0.0,5.88235,0.0,47.0588,0.0,41.1765,0.0,4.77419,1.34146,0.0,4.26316,0.636364,0.393939,0.727273,0.439394,0.545455,0.0833333,0.0833333,avancado

T03E03.txt,154.0,18.0,8.0,188.312,253.247,116.883,84.4156,110.39,8.55556,2.25,2.9798,41.8355,642.857,324.675,436436.0,121565.0,0.689655,240.26,0.594595,1.94444,25.974,0.838384,2.9835,0.0,0.0,0.0,6.49351,51.9481,12.987,0.0,6.49351,0.0,25.974,0.0,19.4805,0.0,7.0,2.02857,0.0,2.22222,0.411765,0.189542,0.470588,0.24183,0.294118,0.0,0.0,avancado

T03E04.txt,193.0,12.0,10.0,155.44,227.979,72.5389,82.9016,165.803,16.0833,1.2,2.82692,5
4.1052,538.86,430.052,336417.0,103027.0,0.482759,207.254,0.4,2.83333,15.544,0.855769,4.
14508,36.2694,5.18135,5.18135,5.18135,51.8135,108.808,56.9948,0.0,25.9067,0.0,36.2694,0
.0,36.2694,5.18135,7.03333,1.55814,0.0,1.14286,0.636364,0.454545,0.727273,0.454545,0.18
1818,0.333333,0.333333,avancado

T03E05.txt,246.0,18.0,7.0,227.642,252.033,97.561,44.7154,146.341,13.6667,2.57143,2.6535
9,51.3194,621.951,361.789,245878.0,59390.2,0.418182,199.187,0.489796,3.0,20.3252,0.856
209,2.98656,12.1951,16.2602,4.06504,8.13008,40.6504,73.1707,36.5854,4.06504,0.0,0.0,40.
6504,0.0,40.6504,0.0,8.51786,2.79167,0.0,3.45833,0.411765,0.392157,0.352941,0.418301,0.
176471,0.0,0.222222,avancado

T03E06.txt,153.0,15.0,11.0,183.007,222.222,111.111,65.3595,117.647,10.2,1.36364,2.91011
,48.2702,581.699,392.157,294703.0,334924.0,0.296296,222.222,0.735294,2.53333,13.0719,0
.853933,3.46021,6.53595,0.0,6.53595,0.0,13.0719,58.8235,19.6078,6.53595,6.53595,0.0,45.7
516,0.0,32.6797,6.53595,4.21429,2.28125,0.0,2.64706,0.357143,0.2,0.357143,0.2,0.142857,0
.2,0.266667,avancado

T04E07.txt,220.0,12.0,8.0,159.091,272.727,77.2727,81.8182,63.6364,18.3333,1.5,2.74615,4
9.4903,590.909,386.364,254710.0,4036.08,0.290323,245.455,0.481481,5.66667,13.6364,0.81
5385,1.17845,22.7273,4.54545,0.0,0.0,27.2727,68.1818,40.9091,4.54545,9.09091,0.0,27.272
7,4.54545,22.7273,4.54545,5.51429,2.7037,0.0,1.41176,0.454545,0.19697,0.545455,0.45454
5,0.454545,0.0,0.0,avancado

T04E08.txt,173.0,10.0,5.0,167.63,289.017,69.3642,28.9017,86.7052,17.3,2.0,2.83333,55.229
3,554.913,416.185,273902.0,1610.7,0.615385,248.555,0.604651,5.1,0.0,0.822917,2.0164,52.
0231,0.0,5.78035,5.78035,63.5838,80.9249,52.0231,11.5607,0.0,0.0,17.341,0.0,17.341,0.0,4.
72414,2.88095,0.0,1.0,0.111111,0.177778,0.333333,0.444444,0.333333,0.0,0.3,avancado

T04E09.txt,117.0,6.0,5.0,153.846,316.239,102.564,59.8291,76.9231,19.5,1.2,2.97297,35.981
,632.479,367.521,354897.0,5322.33,0.8,307.692,0.361111,7.16667,0.0,0.810811,2.13675,8.5
4701,0.0,0.0,0.0,8.54701,8.54701,8.54701,0.0,0.0,0.0,0.0,0.0,0.0,5.88889,2.53125,0.0,0.9
16667,0.6,0.4,0.8,0.733333,0.8,0.0,0.0,avancado

T04E10.txt,154.0,11.0,4.0,246.753,227.273,45.4545,64.9351,129.87,14.0,2.75,2.7,59.9315,5
84.416,402.597,285921.0,3444.0,0.0810811,220.779,0.5,2.27273,25.974,0.855556,3.81971,3
2.4675,12.987,0.0,0.0,45.4545,90.9091,19.4805,12.987,6.49351,0.0,51.9481,0.0,45.4545,0.0,
9.44737,2.40625,0.0,1.71429,0.1,0.290909,0.5,0.454545,0.3,0.0,0.0,avancado

T04E11.txt,135.0,10.0,7.0,214.815,251.852,51.8519,0.0,125.926,13.5,1.42857,3.12857,42.11
92,518.518,437.037,167380.0,16677.8,0.111111,281.481,0.315789,2.3,14.8148,0.8,3.31384,2
9.6296,0.0,0.0,0.0,29.6296,88.8889,51.8519,0.0,0.0,0.0,59.2593,0.0,51.8519,0.0,11.3103,2.25
806,0.0,1.14286,0.333333,0.444444,0.777778,0.688889,0.333333,0.0,0.1,avancado

T04E12.txt,159.0,11.0,3.0,169.811,295.597,81.761,25.1572,100.629,14.4545,3.66667,2.9120
9,49.0014,572.327,427.673,189861.0,6534.55,1.08333,264.151,0.619048,3.36364,12.5786,0.
835165,2.39593,25.1572,0.0,0.0,0.0,25.1572,31.4465,18.8679,0.0,0.0,0.0,12.5786,0.0,0.0,0.0,
5.62963,1.7561,0.0,2.23077,0.3,0.290909,0.5,0.581818,0.4,0.0,0.0,avancado

T04E13.txt,219.0,11.0,6.0,178.082,283.105,100.457,31.9635,73.0594,19.9091,1.83333,3.16923,30.0684,593.607,401.826,237166.0,3036.36,0.486486,246.575,0.537037,5.63636,0.0,0.815385,1.35295,4.56621,9.13242,22.8311,0.0,36.5297,59.3607,13.6986,0.0,4.56621,0.0,50.2283,0.0,41.0959,0.0,5.74359,3.33333,0.0,1.36364,0.3,0.272727,0.5,0.472727,0.2,0.0,0.0,avancado

T04E14.txt,139.0,10.0,7.0,143.885,287.77,71.9424,71.9424,71.9424,13.9,1.42857,2.9875,43.0071,575.54,395.683,329473.0,12062.8,1.23529,244.604,0.558824,5.0,0.0,0.85,2.11595,7.19424,0.0,0.0,0.0,7.19424,79.1367,14.3885,7.19424,7.19424,0.0,50.3597,0.0,43.1655,7.19424,5.1,2.45714,0.0,1.4,0.111111,0.155556,0.444444,0.355556,0.222222,0.0,0.0,avancado

T04E15.txt,224.0,16.0,7.0,174.107,254.464,58.0357,58.0357,178.571,14.0,2.28571,2.86066,52.9616,544.643,437.5,287927.0,128318.0,0.526316,218.75,0.387755,2.8125,22.3214,0.827869,3.64431,44.6429,4.46429,0.0,13.3929,62.5,84.8214,40.1786,8.92857,8.92857,0.0,22.3214,4.46429,17.8571,4.46429,6.94872,2.33929,0.0,2.07692,0.4,0.308333,0.533333,0.408333,0.4,0.0625,0.125,avancado

T01E01.txt,115.0,8.0,2.0,182.609,339.13,78.2609,52.1739,0.0,14.375,4.0,2.8,45.9174,652.174,339.13,317982.0,3148.38,0.1875,286.957,0.515152,4.375,0.0,0.72,0.0,69.5652,0.0,0.0,0.0,69.5652,78.2609,52.1739,0.0,8.69565,0.0,26.087,0.0,17.3913,0.0,7.57143,1.60714,0.0,1.22222,0.714286,0.714286,0.857143,0.821429,0.714286,0.0,0.0,avancado_superior