DOI: 10.14393/Lex2-v1n2a2016-3

Linguística de *Corpus*, Léxico-Estatística Textual e Processamento de Linguagem Natural: perspectiva para estudos de vocabulário em produções textuais[1]

Corpus Linguistics, Lexicostatistics and Natural Language Processing: perspective for vocabulary studies about essays

Aline Evers [*]
Maria José Bocorny Finatto [**]

RESUMO: Partindo da visão teórica e metodológica da Linguística de *Corpus* (LC), conjugada com metodologias do Processamento de Linguagem Natural (PLN), apresenta-se aqui um trabalho de léxico-estatística textual com produções textuais escritas por estudantes de português como língua adicional (PLA). Inicialmente, discute-se a relevância de aspectos quantitativos da linguagem, especialmente a característica de frequência de palavras, conforme propostos por Biderman (1978, 1996) e Hoffmann (2007). Em seguida, situa-se a LC e o PLN e relata-se uma pesquisa (EVERS, 2013) que propôs uma metodologia de avaliação automática aplicada a textos produzidos no contexto do exame Celpe-Bras – um exame de proficiência do português brasileiro. Fazendo uso do Aprendizado de Máquina (AM) supervisionado, uma técnica de PLN, cotejaram-se padrões lexicais e coesivos para distinguir níveis de proficiência e calcularam-se parâmetros de coesão, de coerência e de inteligibilidade textual de uma amostra de textos. Por fim, a proposta de metodologia que associa LC e PLN é problematizada e são apontados seus limites, vantagens e futuras aplicações.

PALAVRAS-CHAVE: Linguística de *Corpus*. Léxico-estatística textual. Português como língua adicional.

ABSTRACT: Based on the theoretical and methodological framework of *Corpus* Linguistics (CL), and allied to Natural Language Processing (NLP) techniques, we present a lexicostatistical study about textual productions written by students of Portuguese as an additional language. We begin by discussing the relevance of quantitative language studies, specially regarding word frequencies, as proposed by Biderman (1978, 1996) and Hoffmann (2007). Then, we situate CL and NLP and their role in the proposition of a methodology (EVERS, 2013) for automatic essay score applied to texts produced in the context of Celpe-Bras – a Brazilian Portuguese as an additional language proficiency exam. By using supervised Machine Learning (ML), a NLP technique, it was possible to identify lexical cohesive patterns and distinguish levels of proficiency using such patterns. Cohesion, coherence and intelligibility parameters were used and the text sample was submitted for examination. At the end, the proposed methodology combines CL and NLP and it is problematized: we point out limits, advantages and future applications for the results found with this research.

KEYWORDS: *Corpus* Linguistics. Lexicostatistic. Portuguese as an additional language.

1 Este trabalho foi realizado durante o projeto RITA (Rich Text Analysis through Enhanced Tools Based on Lexical Resources), financiado pela Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) no âmbito do programa Stic-AmSud (Ciências e Tecnologias da Informação e da Comunicação), processo 047/2014.
* Doutoranda do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.
** Docente do PPG-Letras-UFRGS e do Departamento de Linguística, Filologia e Teoria Literária do Instituto de Letras da Universidade Federal do Rio Grande do Sul.

1. Introdução

Para Halliday, Sinclair e Gries (1989, 1991 e 2006), grandes nomes da Linguística de *Corpus* (LC), a única forma segura para se descrever uma língua é através da observação dessa língua em uso, analisando-se registros autênticos em larga escala. Complementando essa percepção, Hoey (1991) já salientava que, adotando essa postura descritiva, a LC não seria apenas mais uma nova ramificação da Linguística, mas sim um novo caminho para os Estudos da Linguagem.

Considerando a ideia de um caminho e das direções que se buscam em meio aos Estudos da Linguagem, cabe lembrar também aqui as ideias do linguista brasileiro Mário Perini (2010). Para Perini, a Linguística atual ainda estaria em um estágio de desenvolvimento incompleto, comparável ao da História Natural frente à Biologia, dado que ainda se buscam generalizações e se testam hipóteses para o estabelecimento de paradigmas. Assim, tal como entende o autor, um linguista ainda **precisa** (*grifo nosso*) recolher muitos dados, buscar padrões e especificidades, testar hipóteses e agir guiado pelo método da tentativa e do erro. Para Perini, esse estágio de observação direta é ainda muito fundamental.

Tal como vemos, os direcionamentos da LC harmonizam-se muito com esse tipo de percepção de Perini sobre o que é relevante em uma pesquisa linguística. Afinal, na LC, vamos dos fatos às teorizações, ainda que partamos de algumas crenças sobre o que seja a língua e seu funcionamento. Como o que colocamos neste artigo tem caráter experimental e exploratório, o que relatamos bem poderia ser situado nesse cenário. Ao lidar com um pequeno *corpus*, trazemos um exercício de observação que associa à LC o Processamento de Linguagem Natural (PLN) e que visa demonstrar o potencial de uso do Aprendizado de Máquina (AM) na tarefa de gerar sistemas capazes de traduzir elementos de coesão textual em números e, com isso, realizar uma tarefa de classificação automática de textos. Embora tratar desses temas possa parecer desnecessário para algum crítico mais rigoroso, entendemos que, a despeito de a LC estar "em ação" no Brasil pelo menos desde 2000 (conforme marca o artigo já clássico de BERBER SARDINHA, 2000), muitos estudantes e pesquisadores da área de Letras ainda entendem que PLN e LC seriam uma mesma coisa. Por vários motivos, poucos linguistas, nos dias de hoje, mostram alguma familiaridade com o PLN ou mesmo com procedimentos básicos de estatística linguística.

Resgatamos aqui também, ainda que de modo breve, as importantes contribuições da Análise Multidimensional (AMD) proposta por Biber (1988) no cenário da LC dos anos 90. A

AMD foi planejada, justamente, para dar conta do tratamento cruzado de múltiplas variáveis e propriedades da língua e do discurso-texto em um estudo com *corpus*. Assim, cremos que ainda vale um contraponto entre a AMD e os tratamentos de correlações com muitas variáveis do PLN. Assim, pedimos que se considere como moldura para este texto o seguinte pensamento de Perini:

> 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. (PERINI, 2010, p. 11-12)

Portanto, alertamos o leitor que o que segue neste texto deve ser visto como uma sugestão de práticas metodológicas, partindo-se da LC, e não como um trabalho encerrado com certezas postas. Não trazemos neste texto, portanto, um ensaio científico *stricto sensu*, com hipóteses e análises de dados que as confirmam ou não. O foco é mais a reflexão sobre caminhos teóricos e práticas metodológicas partindo-se de uma experiência pontual de estudo com um dado *corpus* (EVERS, 2013) de textos produzidos no contexto do exame Celpe-Bras (Certificação de Proficiência em Língua Portuguesa para Estrangeiros).

2. Encaminhamentos teóricos

2.1 A Linguística de *Corpus* e a Visão Probabilística da Linguagem

Considerando como ponto de partida levantamentos **probabilísticos**, **estatísticos** e **quantificáveis** de dados linguísticos, trazemos neste artigo parte dos resultados de um estudo léxico-estatístico realizado a partir do levantamento de recursos linguísticos presentes em textos submetidos ao exame Celpe-Bras (edição 2006-1). Esses textos foram produzidos por estudantes de Português como Língua Adicional (PLA) que precisam da certificação para variados fins – trabalhar no Brasil, ingressar em uma universidade brasileira, conseguir aumento de salário em suas empresas, entre outros.

Por se tratar de uma análise que lida com a face quantitativa dos estudos da linguagem, é comum que alguns pesquisadores insistam em associar esse tipo de estudo quantitativo e léxico-estatístico a "meros" estudos de vocabulário ou a "simples trabalhos quantitativos com *corpus*". Por essa razão, queremos deixar sublinhado que não é de hoje, nem apenas no âmbito da LC, que se afirma que o aspecto quantitativo é uma das propriedades do léxico – e da língua – e que a frequência é uma característica típica das palavras. Afirmações como essas estão registradas, por exemplo, nos pioneiros trabalhos realizados em língua portuguesa por Maria Tereza Camargo Biderman (CAMARGO, 1967; BIDERMAN, 1978 e 1996).

É na esteira do legado de Biderman, marcado especialmente no seu trabalho fundador de 1967, que exemplificamos aqui 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 nossos*). Nesse sentido, importa alertar que o enfoque estatístico deve ser entendido como uma **referência**, e não como um fim em si mesmo. O resultado da análise estatística é 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. Assim, estamos descrevendo e não prescrevendo.

A LC foi e ainda é capaz de fornecer boas pistas para o esclarecimento do que pode ser considerado "uso da língua" através da análise de *padrões* detectados em grandes conjuntos de textos autênticos. A análise proposta pela LC mostra que há uma visão de língua como *probabilidade* por trás dos procedimentos metodológicos adotados, contrapostos pela visão de língua como *possibilidade*, fortemente presente nos estudos gerativistas. Dessa forma, os estudos de LC não têm como objetivo mostrar um exemplo de uso perfeito/ideal ou um modelo de língua: eles têm como objetivo, em sua maioria, descrever o que ocorre num dado uso, nos mais diferentes recortes de registros, mostrando diferentes nuances e possibilidades, sejam esses usos considerados mais ou menos adequados, formais ou informais.

De acordo com Berber Sardinha, "embora muitos traços linguísticos sejam possíveis teoricamente, eles não ocorrem com a mesma frequência" (2004, p. 350). Isso significa que a frequência de ocorrência de determinados elementos linguísticos é maior do que a de outros, e que essas frequências estão atreladas a contextos. Vale aqui, a propósito, um exemplo ilustrativo. Ao considerarmos o uso de um grupo de conectivos causais a partir dessa

pressuposição, foi possível mostrar que existe uma probabilidade maior de uso desses conectivos causais em textos didáticos universitários brasileiros de Química do que em artigos científicos de Química escritos em português (EVERS; ALLE; MARCOLIN, 2008; FINATTO, EVERS; ALLE, 2009). Para fazer esse tipo de constatação, algo difícil de se perceber diretamente pela leitura de alguns poucos textos, foram necessárias grandes quantidades de dados e ferramentas computacionais de apoio. Vale salientar que foi possível confirmar, estatisticamente, uma *tendência* para uso de um elemento linguístico-textual, associando-se diferentes gêneros, no caso, discursos produzidos que eram mais ou menos didáticos.

Com esse método de observar textos autênticos em busca de padrões, a LC vai de encontro, por exemplo, àquelas típicas observações que tentam mapear discrepâncias em relação ao que as gramáticas tradicionais postulam como "boa" norma ou que buscam confirmações para algumas certezas. Isso significa que um *corpus* de estudo, no âmbito da LC, não é apenas um lugar para confirmar ideias colocadas *a priori*, tampouco deve servir de fonte de exemplos para o que já se sabia. Um *corpus* é, antes de tudo, uma fonte de novas perguntas e de respostas muitas vezes inesperadas.

Com isso, podemos afirmar, com certa tranquilidade, que os trabalhos de LC fazem uso de ferramentas de análise lexical que evidentemente oferecem dados interessantes para a observação em larga escala, que mostram padrões de uso e que permitem que o linguista faça descrições interessantes sobre o uso da linguagem em um enfoque extensivo. No entanto, sua metodologia e algumas ferramentas, conforme vemos, parecem um tanto complexas para um uso direto quando comparadas a algumas ferramentas computacionais estatísticas disponíveis no âmbito do PLN. É sobre a natureza e as técnicas de PLN que tratamos a seguir.

2.2 Processamento de Linguagem Natural

O PLN é uma subárea da Inteligência Artificial, ramo da Ciência da Computação, que agrupa métodos formais para analisar textos e gerar frases escritas em língua natural. O objetivo final do PLN é capacitar computadores a "entender" e a "compor" textos em língua natural. Por "entender", queremos dizer serem capazes de reconhecer o contexto, fazerem a análise sintática, semântica, léxica e morfológica, criarem resumos, extraírem informação, gerarem traduções automáticas, interpretarem os sentidos e até "aprenderem" conceitos fazendo uso de textos processados. Não se sabe se um dia os computadores poderão igualar a capacidade humana de entender e compor textos. Afinal, atualmente, essas capacidades computacionais ainda são

bastante limitadas, mas muitos resultados práticos já existem e são utilizados por diversos tipos de programas e para muitas finalidades, com ótimos desempenhos para os fins a que se propõem.

Para demonstrar algo que o PLN possibilitou e que usamos cotidianamente ao redigir *e-mails* e outros tipos de texto, podemos citar os corretores ortográficos. Qualquer pessoa que tenha utilizado uma ferramenta de processamento de texto, como o MS Word, Open Office ou Pages, sabe que ela contém um corretor ortográfico, que destaca desvios ortográficos e gramaticais e propõe correções. Longe de perfeitos, são reconhecidamente úteis.

Os primeiros corretores ortográficos, desenvolvidos no âmbito do PLN, funcionavam através da comparação simples de uma lista de palavras extraídas do texto em escrita com uma lista de palavras (dicionário de palavras) corretamente grafadas, previamente armazenada no computador. Era uma tarefa extremamente simples e que não exige processamento complexo. Essas ferramentas, hoje, tornaram-se bem mais sofisticadas e são capazes de detectar desvios 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). Esses são os léxicos computacionais, conhecidos como *bases de conhecimento lexical*. Esses léxicos foram já tratados por Zavaglia (2006), autora que nos apresentou, pioneiramente, um tal ponto de encontro entre linguistas e cientistas da Computação em sua tese de doutorado (ZAVAGLIA, 2002), orientada por M. T. C. Biderman.

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

2.3 Da Linguística de *Corpus* ao Processamento de Linguagem Natural

Os grandes problemas encontrados atualmente por linguistas que fazem uso de *corpora* e da LC em suas pesquisas são: a) lidar com estatísticas complexas e mais acuradas[2]; e b) gerar estudos que analisem texto a texto, e não somente grandes "pacotes de palavras" ou "pacotes de textos". Esses problemas tendem a ser enfrentados também por pesquisas de PLN, mas com alguns refinamentos metodológicos e ferramentas estatísticas diferenciadas, as quais acreditamos que vale a pena conhecer.

No cenário brasileiro, alguns exemplos recentes, entre vários, de trabalhos que deram um passo no sentido de um encontro da LC com o PLN 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 como o Philologic[3] e o Unitex[4]. No entanto, para realizar a classificação automática, fez uso de algoritmos comuns ao PLN e ao Aprendizado de Máquina, que trataremos mais a seguir.

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 que tratam de medidas de coesão e de coerência do texto em português e em inglês, tais como os sistemas Coh-Metrix e o Coh-Metrix-Port. Fruto da parceria entre informatas e linguistas, esse trabalho foi capaz de mostrar que as traduções para o português resultaram em textos mais complexos do que seus textos-fonte, considerando a comparação de algumas das métricas aferidas pela ferramenta.

No âmbito de pesquisas internacionais sobre a produção textual de estudantes de língua inglesa como língua adicional, há trabalhos recentes que demonstram a viabilidade de se mensurar proficiência escrita de forma automática. Essas pesquisas fazem uso de técnicas que não pertencem à LC e utilizam o ferramental do PLN para obter soluções apuradas e confiáveis. Ferris (2002), Jarvis *et al*. (2003) e Hulstijn (2012), por exemplo, utilizaram medidas como

[2] Aqui cabe reler o trabalho de Biderman (CAMARGO, 1967) e notar que pouco mudamos nesse quesito. Texto disponível em : http://seer.fclar.unesp.br/alfa/article/view/3300/3027
[3] Disponível em https://sites.google.com/site/philologic3/. Acesso em 10 set 2013.
[4] Disponível em http://www-igm.univ-mlv.fr/~unitex/. Acesso em 10 set 2013.

variedade lexical, repetição de palavras e tamanho de texto para aferir proficiência escrita em língua inglesa. Outros autores, como Crossley e McNamara (2012), utilizam métricas coesivas para avaliar automaticamente níveis de proficiência escrita em inglês considerando-se o encaixe ou não de um texto X em um dado padrão coesivo previamente formalizado. O "encaixe" é feito automaticamente, mediante o uso de uma ferramenta de análise de textos abastecida com padrões recorrentes de frases e de textos retirados de *corpora* de produções textuais de estudantes.

Esses trabalhos já apontam que, por exemplo, quanto maior o nível de proficiência em língua adicional, maior é a variedade lexical empregada pelo autor em seus textos. É evidente que muitos desses apontamentos podem ser discutidos e relativizados enquanto ***tendências*** que são. Essa afirmação, por exemplo, pode ser contestada em outros contextos, como já foi verificado em redações de vestibulandos brasileiros (FINATTO *et al.*, 2008) em contexto de português língua materna. Nesse caso, conforme a amostra construída, verificou-se que as redações que receberam nota maior no concurso vestibular da Universidade Federal do Rio Grande do Sul (UFRGS) tenderam a exibir menor variedade lexical. A tendência dos vestibulandos que receberam escores mais altos em suas redações foi a de maior coesão, estabelecida através da repetição de palavras.

Essas questões advindas dos estudos em *corpus*, portanto, não estão encerradas e sempre instigam nossa curiosidade e nos fazem recolocar novas e antigas perguntas. Até que ponto confirma-se uma tendência X em uma dada amostra Y e Z? Afinal, a variedade lexical é ou não um critério capaz de separar textos mais bem-sucedidos em contextos avaliativos? Até que ponto o léxico empregado nas produções textuais é capaz de nos dizer onde devem ser encaixados os textos em um *continuum*? Assim inspiradas, a seguir, relatamos os passos de uma pesquisa que buscou classificar automaticamente textos submetidos a um exame de proficiência de PLA. Trata-se de uma pesquisa bastante marcada pelo estilo de trabalho em PLN.

3. Metodologias de LC combinadas com as de PLN

3.1 Um *Corpus* de Produções Textuais de Falantes de Português como Língua Adicional (PLA)

Ainda são poucas as pesquisas que exploram textos produzidos por estudantes de PLA (SIDI, 2002; SCHOFFEN, 2009; GOMES, 2009; YUQI, 2011; DAMAZO, 2012), especialmente que utilizem pressupostos da LC e de técnicas de PLN. A ideia do trabalho aqui

relatado foi a de buscar no PLN metodologias e técnicas de pesquisa que viabilizassem uma observação descritiva, quantitativa, automatizada e acurada desse tipo de texto, de modo a construir uma base para futuras pesquisas qualitativas e quantitativas em larga escala sobre avaliação.

Uma das razões para recorrer ao PLN foi justamente a pequena dimensão do *corpus* reunido naquele estudo. A amostra continha apenas 177 textos, ou seja, um conjunto com pouco mais de 25 mil palavras. Outro ponto que favoreceu o diálogo com o PLN foi o propósito de avaliar texto a texto, em detalhe, frente ao conjunto reunido. No PLN, há recursos capazes de produzir uma análise multifatorial com muitos elementos em correlação, mesmo que se parta de um número relativamente pequeno de textos. Assim, é possível testar, matematicamente, e verificar se há correlação entre diferentes características linguísticas e entre diferentes textos. Dessa forma, por exemplo, torna-se possível confirmar se há alguma correlação estatística entre a avaliação que um texto recebeu em um dado exame – de proficiência, vestibular ou concurso público –, seus números de parágrafos, tamanho do texto e variedade lexical, além de outras características formalizáveis. Em LC, em geral, lidam-se com amostras de textos bem maiores para compor um *corpus* de estudo. Esse *corpus* será colocado frente a um *corpus* de referência pelo menos cinco vezes maior.

3.2 Uma ideia de estudo: Avaliação de Proficiência usando Estatística Lexical

Partindo da microperspectiva estrutural do texto, isto é, considerando **apenas a tessitura coesiva e o perfil lexical dos textos**, a pesquisa que relatamos aqui foi um estudo quantitativo e qualitativo sobre possíveis métricas úteis para a predição de graus de proficiência escrita em português. O *corpus* utilizado teve 177 produções textuais, que já tinham sido previamente avaliadas por professores-avaliadores e separadas em seis níveis, que chamaremos aqui de classes: Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior.

Em trabalhos cuja metodologia segue os preceitos da LC, em geral, os textos são agrupados em um único arquivo e, a partir desse grande *corpus*, são feitas as análises lexicais, gramaticais e semânticas de um todo amostral. Nossas análises foram feitas de duas formas: a primeira, por pequenos agrupamentos, e a segunda, texto a texto. Para averiguar o comportamento estatístico-lexical de cada uma das seis classes (níveis de proficiência), agrupamos os textos por níveis. Depois, na segunda maneira de análise, os textos são tratados

individualmente, agrupando **os resultados** que a ferramenta Coh-Metrix-Port – que indica medidas de coerência e de coesão dos textos – forneceu para cada texto. Por fim, essas medidas são relacionadas às classes/aos níveis de proficiência, de modo a observarem-se regularidades e perfis coesivos presentes.

A ferramenta Coh-Metrix-Port, utilizada para gerar os resultados estatísticos sobre recursos coesivos empregados nos textos, foi desenvolvida, por um grupo brasileiro de pesquisa de PLN, para auxiliar a tarefa de simplificação de textos e facilitação do acesso à informação para analfabetos funcionais e para pessoas com deficiências cognitivas. Assim, essa ferramenta foi originalmente imaginada para "medir" automaticamente a complexidade coesiva de um dado texto. Sua mais recente versão opera com 48 métricas disponíveis gratuitamente. Ela foi adaptada para o português do Brasil por Scarton e Aluísio (2010) partindo-se de um sistema feito para o inglês. Suas medidas principais são:

- Contagens básicas: número de palavras, número de sentenças, número de parágrafos, sentenças por parágrafos, palavras por sentenças e sílabas por palavras.
- Índice Flesch, que é um índice de inteligibilidade bastante corrente em PLN, pouco conhecido por linguistas no Brasil, ainda que estudos sobre leitura já o tenham apresentado (como em LEFFA, 1996).
- 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, *types* e *tokens*: incidência de pronomes pessoais, pronomes por sintagmas nominais e relação *type-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.

O sistema Coh-Metrix-Port gerou, portanto, as medidas acima mencionadas para cada um dos 177 textos da amostra de produções textuais submetidas ao exame Celpe-Bras. Essas medidas são devolvidas em formato de arquivo textual, e uma parte do arquivo gerado pode ser visto na Figura 1 a seguir:

- **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 *e* em um texto.
Incidência de OU	12.3457	Incidência do operador lógico *ou* em um texto.
Incidência de SE	0	Incidência do operador lógico *se* em um texto.
Incidência de Negações	0	Incidência de Negações. Consideramos como negações: *não, nem, nenhum, nenhuma, nada, nunca e jamais*.

- **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 córpus 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 1. Parte do arquivo textual gerado pelo sistema Coh-Metrix-Port.

Como é possível visualizar na Figura 1, cada medida é traduzida por um número. Por exemplo, no arquivo recortado, o item "Incidência de Operadores Lógicos" teve como contagem 55.55. Essa numeração é comparada entre os arquivos e correlacionada às seis classes/aos níveis de proficiência definidos neste estudo.

Para realizar essa comparação multifatorial, recorremos a uma área do PLN chamada de Aprendizado de Máquina (AM). O AM nada mais é do que o desenvolvimento de algoritmos que permitem que o computador "aprenda" padrões. Algumas partes do AM estão intimamente ligadas à mineração de dados e à estatística, e são essas as partes que interessam ao linguista. Uma tarefa de mineração de dados recorrente em AM, por exemplo, dado que envolve

identificação de padrões de atributos de dados associados entre si, é a Classificação. Neste estudo, a Classificação foi utilizada para "ensinar" o computador a diferenciar os níveis de proficiência a partir dos resultados que a ferramenta Coh-Metrix-Port forneceu. Para realizar essa Classificação automática, um sistema chamado Weka foi utilizado: ele possibilita a observação estatística de dados e permite observar correlações relevantes entre atributos e classes, por exemplo. Esse sistema é utilizado por diferentes áreas de pesquisa, e no caso da pesquisa linguística aqui descrita, procuramos observar as correlações entre as propriedades coesivas dos textos (traduzida em números) e os níveis de proficiência do português avaliados.

O Weka foi utilizado para apurar as relações discriminativas entre as métricas; esse sistema trabalha somente com um tipo de arquivo, que precisa ser preparado pelo pesquisador. O nome deste arquivo é .ARFF, e possui 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. Os **atributos**, no caso da pesquisa relatada, são as métricas do Coh-Metrix-Port, descritas anteriormente; as **instâncias** são cada um dos textos analisados; e as **classes** são os seis níveis de proficiência com os quais está-se trabalhando: Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior. Um trecho do arquivo .ARFF utilizado neste trabalho é apresentado na Figura 2 a seguir. Nele é possível ver os atributos descritos em **@attribute** (numero_palavras, flesch_numeric, etc), as instâncias em **@data** (T01E09 significa "Tarefa 01, texto número 09"), em que são vistas as "numerações" para cada um dos atributos listados e, por fim, a classe, que aparece ao final da mesma linha que começa em @data:

```
@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_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.
0,5798.0,0.4,257.353,0.628571,4.33333,0.0,0.666667,1.89076,14.7059,14.7059,7.35294,7.35294,44.1176,66.1765,29.4118,1
4.7059,0.0,0.0,29.4118,0.0,29.4118,0.0,4.18182,1.2,0.0,0.625,0.75,0.916667,0.875,0.861111,0.875,0.222222,0.222222,avanc
ado
```

Figura 2. Parte do arquivo .ARFF utilizado na pesquisa, delineado para que funcionasse no sistema Weka.

4. Resultados

4.1 Resultados da análise: agrupamento por nível

Alguns dos resultados mais interessantes da análise recaem nos dados lexicais dos textos. Esses dados foram obtidos através dos agrupamentos por níveis de proficiência, informação dada pelos professores-avaliadores. Dessa forma, os textos foram analisados quantitativamente e dados como os presentes no Gráfico 1 a seguir foram gerados. No Gráfico 1, os números do eixo horizontal correspondem aos níveis de proficiência avaliados no exame (1- Iniciante, 2- Básico, 3- Intermediário, 4- Intermediário Superior, 5- Avançado e 6- Avançado Superior). O número total de palavras é representado pela linha AZUL. A linha AZUL mostra que existem duas fases de crescimento com relação ao número de palavras utilizadas nos textos nos diferentes níveis de proficiência. Na primeira observação, vemos que há um crescimento importante no número de palavras nos níveis Iniciante e Avançado (há picos de crescimento visíveis).

Esses picos no número de palavras utilizadas já não existem nos níveis de proficiência Intermediários (números 3 e 4 no eixo horizontal), e vemos que o crescimento do número de palavras se estabiliza. O número volta a crescer nos níveis avançados, números 5 e 6 do eixo, em que as produções textuais são, portanto, mais extensas.

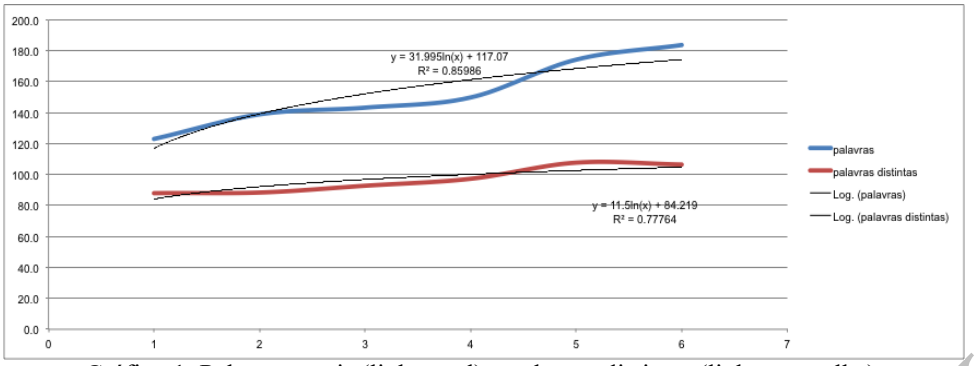

Gráfico 1. Palavras totais (linha azul) e palavras distintas (linha vermelha).

A linha VERMELHA, por sua vez, representa o número de palavras diferentes dos textos nos conjuntos. Com relação ao uso de palavras diferentes, é possível perceber que, quanto mais Básico é o nível do conjunto de textos, maior é o número de palavras diferentes nele contido, comparado aos 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 apresentando grande variação de vocabulário.

Tendo em vista os resultados interessantes com relação ao uso de palavras, optamos por verificar o que aconteceria com relação à intersecção de vocabulário entre os níveis. O que chamamos aqui de intersecção é o uso das mesmas palavras nos diferentes níveis da classificação. Para identificar a intersecção de vocabulário entre esses conjuntos, foi utilizado o Índice de Jaccard, que possibilita a verificação da sobreposição de dois grupos, A e B. O objetivo de utilizar esse coeficiente é o de verificar, estatisticamente, a similaridade e a diferença entre os textos presentes nos conjuntos. As medidas de similaridade lexical são aplicadas entre os segmentos, estabelecendo relações mais fortes ou mais fracas. A intersecção 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 a similaridade, com relação às palavras usadas, com os demais grupos. A Tabela 1 a seguir mostra os resultados numéricos dessa intersecção. O símbolo (*) significa que a célula representa o encontro do grupo com ele mesmo (por exemplo, Avançado com Avançado ou Iniciante com Iniciante). Desse encontro, o valor

só pode ser (*) porque é evidente que as palavras utilizadas pelo grupo, quando comparadas a ele mesmo, resultam em um espelhamento do uso de palavras:

Tabela 1. Intersecção de vocabulário entre os conjuntos de textos.

	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	*

Uma forma gráfica de visualizar a Tabela 1 é dada pela Figura 3 a seguir, na qual observa-se que quanto mais Avançado é o nível de proficiência do grupo de textos, maior é a sua similaridade com os demais níveis:

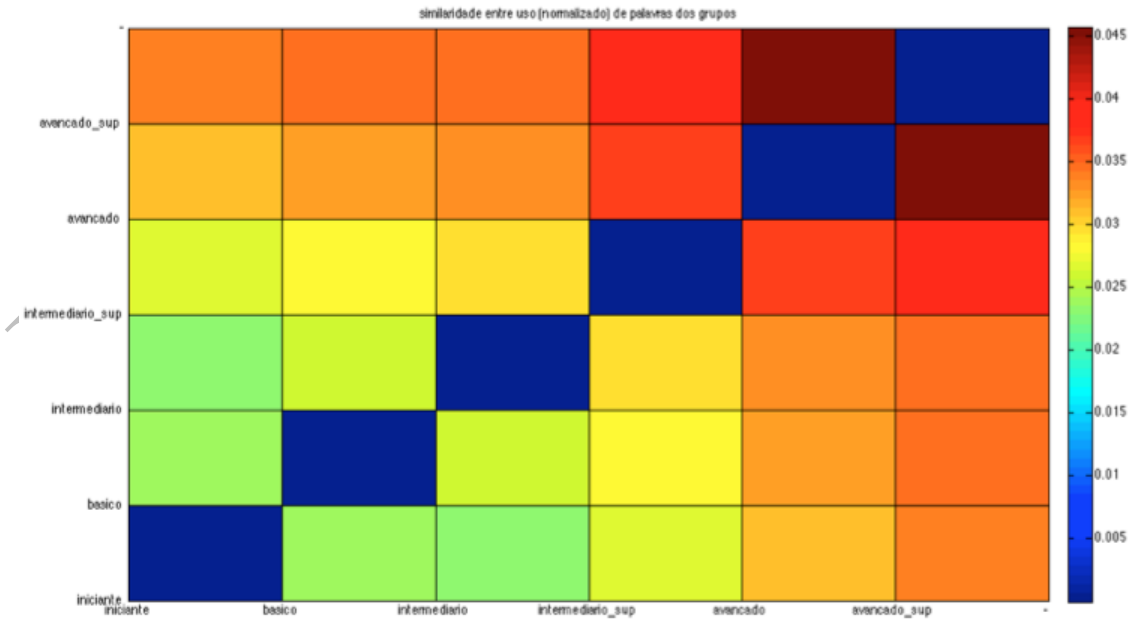

Figura 3 Similaridade de vocabulário entre os grupos.

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 semelhante 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 os vocabulários utilizados 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.

4.2 Resultados da análise: atributos Coh-Metrix-Port e correlações entre textos e níveis

Através da observação da Árvore de Decisão[5], mostrada na Figura 4, foi gerada a Tabela 2 a seguir, que mostra os 15 atributos (dos 48 fornecidos pelo Coh-Metrix-Port) que seriam, em tese e a partir da leitura das Árvores de Decisão, possíveis bons indicadores para diferenciar textos mais proficientes de textos menos proficientes, pois apresentavam uma média diferente com relação aos demais atributos.

[5] Uma árvore de decisão é formada por um conjunto de nós de decisão, ou seja, espécieis de perguntas, que permitem a classificação de cada texto em um nível de proficiência.

Tabela 2. Atributos (medidas do Coh-Metrix-Port) que demonstraram comportamento não aleatório.

	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**

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 significam 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**.

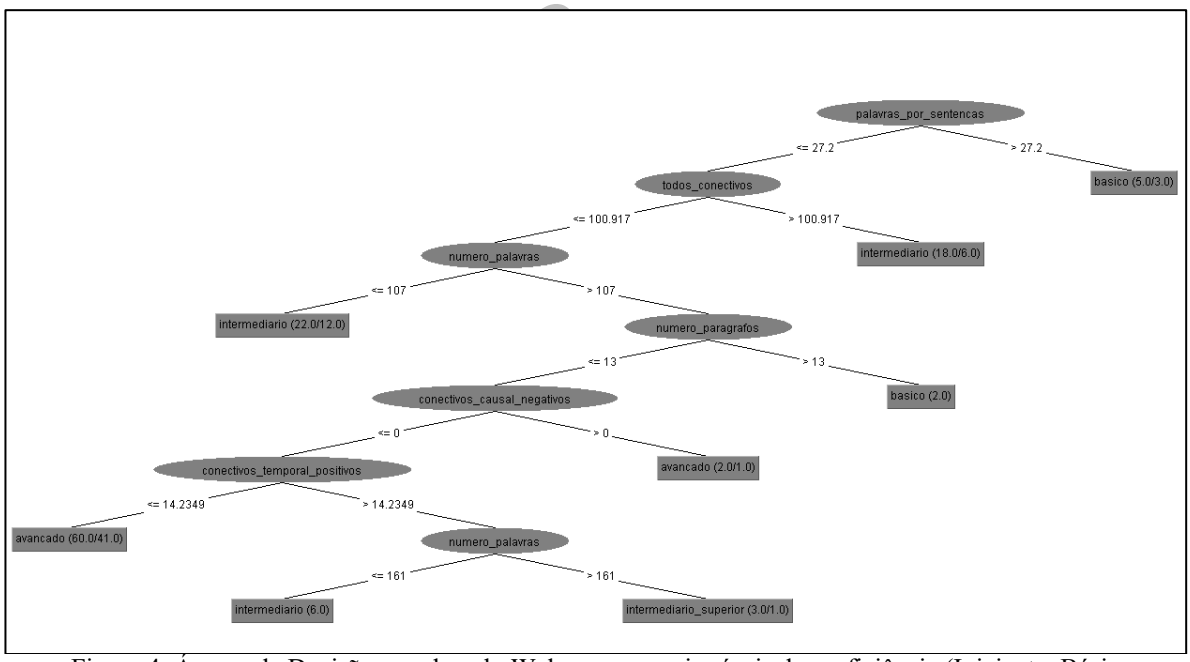

Figura 4. Árvore de Decisão gerada pelo Weka para os seis níveis de proficiência (Iniciante, Básico, Intermediário, Intermediário Superior, Avançado e Avançado Superior).

4.3 Principais Resultados da Pesquisa

Nosso desafio foi verificar, de um modo estatisticamente válido, **quais** e **que tipos** de características exibidas pelos textos seriam mais profícuas para guiar uma classificação automática da amostra de textos. Partimos do pressuposto de que a ferramenta Coh-Metrix-Port poderia apontar um bom conjunto dessas características, com a vantagem desse apontamento ocorrer de um modo automático, dado texto a texto.

A junção entre os princípios da LC e as técnicas do PLN mostrou que é possível selecionar determinadas características textuais e manipulá-las, de um modo eficiente, para fins de uma classificação automática. Estabelecemos uma correlação entre as características coesivas medidas pela ferramenta Coh-Metrix-Port e as características dos textos desse *corpus*. Evidentemente, é possível encontrar algumas limitações na proposição de uma metodologia automática, baseada em *corpus*, visando classificar automaticamente esse tipo de produção textual. A primeira limitação foi a falta de um *corpus* maior, o que inibe, de certa forma, mesmo em PLN, o alcance do tratamento estatístico.

Mas, apesar dos limites, ficou claro que a principal contribuição desta pesquisa refere-se à própria metodologia experimentada: como levantar informações linguísticas de um *corpus* de textos produzidos por estudantes de PLA, com todas as suas peculiaridades, e como utilizá-las para implementação de sistemas computacionais. A partir dessa experiência, como trabalhos futuros, sugerem-se:

a) realizar pesquisas sobre expressões linguísticas – especialmente construções retóricas ou argumentativas – que fossem capaz de funcionar como potenciais classificadoras de um texto de um dado domínio do conhecimento ou de um gênero textual/discursivo;
b) comparar os traços de um *corpus* de falantes nativos de português com um *corpus* de estudantes estrangeiros ou contrastar um *corpus* de estudantes de português brasileiro como LA com um *corpus* de estudantes de português europeu como LA;
c) ampliar o *corpus* e testar o papel de novos atributos – características coesivas ou outras - não tratados no sistema Coh-Metrix-Port – tais como, por exemplo, as elipses na correlação com níveis de proficiência pré-estabelecidos. Também aqui caberia, talvez, estabelecer-se uma taxonomia de traços mais produtivos - em termos estatísticos - para avaliação de características de proficiência escrita do português;

d) classificar o *corpus* em função de gêneros textuais/discursivos por tipos de tarefas ou por propostas de redação em diferentes tipos, 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 e explicitamente apontado pelos avaliadores como "fluência" maior ou menor. Comentários de um avaliador, tal como, como "o texto avançado é mais fluente" servem como um estímulo para que tentemos perseguir, num *corpus*, o que realizaria essa "fluência";
f) correlacionar características linguísticas do texto da tarefa (que é o comando que gera a redação nesse tipo de prova - o enunciado da tarefa) e a produção textual dos examinandos, também com vistas a detectar eventuais cópias de grandes segmentos.

Cabe, ainda, ressaltar que toda a análise empreendida foi baseada em textos já corrigidos por avaliadores humanos e já classificados por eles nos níveis pré-determinados desse exame. Foi a partir dessa classificação humana que todos os testes, refinamentos, medidas e escores estatísticos foram cruzados, tendo como índice máximo de confiabilidade o quanto os dados da pesquisa se aproximavam ou se encaixavam nas medidas dadas pela avaliação humana dos textos. As avaliações humanas, perpassadas por diferentes condições, não se comparam a um método computacional de análise, totalmente empírico, baseado em dados concretos e diretamente observáveis.

5. Considerações finais: a análise multidimensional frente a técnicas de PLN

Como dito no início deste texto, um estudo *léxico-estatístico textual* com *corpus* pode gerar inúmeros dados. Esses dados devem ser entendidos como uma **referência**, apontando tendências no que se refere a características da língua em uso. Estamos aqui aventando possibilidades para quem se interessa pela LC e por diferentes metodologias de estudo com *corpora* e para quem esteja interessado em não perder de vista a especificidade dos textos sob exame em meio ao todo de um *corpus*. Entretanto, ao finalizar este texto, cabe dizer que esse tipo de preocupação não é uma novidade para o linguista de *corpus*.

Metodologias de estudo que usam técnicas e princípios de PLN conjugadas com LC muito se aproximam do que já vem sendo feito, há vários anos, em LC através da AMD. Tal como proposta por Douglas Biber em 1988 (BIBER, 1988), a AMD permite, também, a investigação dos padrões principais da variação por características linguísticas dos registros falado e escrito.

Esse tipo de análise, operacionalizada pela primeira vez por Biber, já visava permitir uma descrição rica e complexa de *corpora* permitindo uma *clusterização* ou agrupamento de textos, derivando-se essa classificação do conceito de dimensão de variação. Dimensão constitui um conjunto de traços que subjazem a um *corpus*. O método de análise da AMD possibilita utilizar concomitantemente uma variedade de traços linguísticos empregados na análise textual e aplicar a codificação desses traços a um número de textos maior do que se poderia fazer manualmente, utilizando ferramentas computacionais e estatísticas. Por meio da AMD, a variação entre textos e registros pode ser descrita por meio de múltiplos parâmetros, possibilitando a utilização de um aparato quantitativo de descrição, o qual permite a especificação da coocorrência dos traços linguísticos de modo preciso (BERBER SARDINHA, 2004).

O enfoque da AMD foi inovador no cenário da LC por combinar análises de nível macroestrutural com análises de nível microestrutural, ou seja, da macrodimensão do *corpus* chega-se à microdimensão do texto, e a microdescrição dos traços de cada texto revela macroagrupamentos textuais, que caracterizam os gêneros (FINATTO, 2011). Para quem se interesse em conhecer mais sobre a AMD, podendo inclusive comparar a metodologia com a de Evers (2013), o trabalho de Zuppardo (2013) pode ser um bom ponto de partida. Esse é um trabalho bastante atual que exemplifica a técnica e os passos de pesquisa da AMD, tendo lidado com manuais de aeronáutica em inglês. Outro trabalho, também exemplar, embora mais antigo, é o de Shergue (2003).

Assim, além do que colocamos aqui sobre PLN, a AMD também é uma opção metodológica interessante para levantar traços linguísticos coocorrentes em textos em determinados agrupamentos. Afinal, a possibilidade de identificar padrões particulares a um e outro tipo de texto seria uma forma de apontar eventuais diferenças, por exemplo, de níveis de proficiência em um conjunto de textos. Para além desse avanço dos estudos com *corpora,* conforme vemos, o PLN e o AM poderiam também ser muito úteis para quem lida com LC. Do

mesmo modo, acreditamos que a AMD pode oferecer alguns *insights i*nteressantes para o pesquisador de PLN que lida com discursos/textos para a geração de diferentes recursos.

Finalizando este artigo, cabe retomar a ideia de que esse tipo de estudo exploratório, feito apenas com 177 redações em português produzidas por estrangeiros, faz pensar sobre diferentes rumos da pesquisa linguística. Longe de darmos conta do todo de significação que é o texto, é inegável a contribuição dos pequenos passos e dos modos de fazer da Estatística Linguística que enfatiza o léxico.

Referências Bibliográficas

BERBER SARDINHA, T. Linguística de *Corpus*: histórico e problemática. **DELTA**, São Paulo, v. 16, n. 2, p. 323-367, 2000. crossref http://dx.doi.org/10.1590/s0102-44502000000200005

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

BIBER, D. **Variation Across Speech and Writing**. Cambridge: Cambridge University Press, 1988. crossref http://dx.doi.org/10.1017/CBO9780511621024

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.

BIDERMAN, M. T. C. Estatística linguística. **Alfa**, São Paulo, v. 11, p. 117-128, 1967.

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. crossref http://dx.doi.org/10.1111/j.1467-9817.2010.01449.x

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.

EVERS, A. **Processamento de língua natural e níveis de proficiência do português:** um estudo de produções textuais do exame Celpe-Bras. 2013. 174 f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

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.

FERRIS, D. **Treatment of error in second language student writing**. Ann Arbour: 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. Complexidade textual em artigos científicos: contribuições para o estudo do texto científico em português. **Organon** (UFRGS), 2011, p. 30-45.

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.

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.

GRIES, S. Th. **Corpora in cognitive linguistics**: Corpus-based approaches to syntax and lexis, 1–18. Berlim, Heidelberg, Nova York: Mouton de Gruyter, 2006. crossref http://dx.doi.org/10.1515/9783110197709

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

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.

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. crossref http://dx.doi.org/10.1016/j.jslw.2003.09.001

LEFFA, V. J. **Fatores da compreensão na leitura**. Projeto ELO, Ensino de línguas on-line: 1996. Disponível em**:** www.leffa.pro.br.

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.

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.

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.

SHERGUE, O. **Dimensão de Variação no Discurso Médico-Acadêmico**: o Artigo de Pesquisa e a Apresentação de Trabalhos Científicos em Congressos. 2003. Dissertação (Mestrado em Letras) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2003.

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.

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.

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.

ZAVAGLIA, C. **Análise da homonímia no português**: tratamento semântico com vistas a procedimentos computacionais. 2002. Tese (Doutorado) – Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara.

ZAVAGLIA, Claudia. Extração de informações de definições de um dicionário convencional para a elaboração de uma base de conhecimento lexical: estratégias e procedimentos linguísticos. In: LONGO, Betariz N. De O.; DIAS-DA-SILVA, Bento C. (orgs.) **A construção**

de dicionários e de bases de conhecimento lexical. São Paulo: Cultura Acadêmica, 2006. p. 209-234.

ZUPPARDO, M. C. A linguagem da aviação: um estudo de manuais aeronáuticos baseado na Análise Multidimensional. **ReVEL**. v. 11, n. 21, 2013.

Artigo recebido em: 25.03.2016 Artigo aprovado em: 18.04.2016