VOCABULÁRIO CONTROLADO E REDAÇÃO DE DEFINIÇÕES EM DICIONÁRIOS DE PORTUGUÊS PARA ESTRANGEIROS: ENSAIOS PARA UMA LÉXICO-ESTATÍSTICA TEXTUAL

FINATTO, Maria José Bocorny [*]
EVERS,Aline
PASQUALINI, Bianca Franco
KUHN, Tanara Zingano
PEREIRA, Aline Maciel.

***RESUMO**:* Estudo inicial em léxico-estatística textual para colher dados que subsidiem a construção de um vocabulário controlado (VC) de referência para a redação de definições em um dicionário de português para estrangeiros. Aproveitaram-se dados do vocabulário mais frequente em jornais populares brasileiros e foram analisados três diferentes corpora. Comparadas as palavras mais frequentes das fontes, avaliou-se o uso dos VCs para elaborar definições de verbetes-teste. Evidenciou-se o bom aproveitamento dos corpora e a relevância da estatística linguística para a composição do VC.

***PALAVRAS-CHAVE**:* Vocabulário controlado; definição lexicográfica; frequência de palavras.

***ABSTRACT**:* Initial study in lexical-textual statistics that aims at collecting data to support the construction of a basic controlled vocabulary (CV) to be a reference for writing definitions in a Portuguese learner's dictionary. We used vocabulary frequency data from Brazilian popular newspapers and we also analyzed three different corpora. After comparing the most frequent words of each source, we evaluated the use of CVs to prepare a set of test entries. The results demonstrate the proper use of these corpora for the composition of a CV and the relevance of statistical linguistics for its compilation.

***KEYWORDS**:* Controlled vocabulary; lexicographic definitions; word frequency.

[*](a) Doutora em Letras pela UFRGS, pesquisadora CNPq,professor associado da UFRGS, orientadora de Doutorado. (b) Mestre em Letras (UFRGS), bolsista SEAD-UFRGS, professora de português para estrangeiros. (c): doutoranda PPG-Letras-UFRGS, bolsista CNPq. (d) Doutoranda em Linguística Aplicada, Universidade de Lisboa, Bolsista Capes Proc. 0973/13-0, professora de português para estrangeiros. (e) Acadêmica de Letras (UFRGS), bolsista Iniciação Científica PIBIC-CNPq.

INTRODUÇÃO: O PAPEL DA ESTATÍSTICA LINGUÍSTICA

Uma série de previsões e de constatações sobre o funcionamento da língua e sobre os elementos gramaticais presentes nos discursos orais ou escritos pode ser feita por meio da léxico-estatística, comprovando-se que o quantitativo é uma das propriedades do vocabulário e que a frequência é uma característica importante, típica da palavra. Essa constatação, conforme Nascimento e Isquierdo (2003, p. 2), parte de alguns dos trabalhos pioneiros de Maria Tereza Camargo Biderman (1978, 1996) para o português brasileiro. Pesquisadora e lexicógrafa de saudosa memória, Biderman apresentou, defendeu e demonstrou[1], no contexto brasileiro, já desde meados dos anos 60, o mérito da Estatística Linguística para a Lexicologia, Filologia, descrição de línguas e Lexicografia.

Como seu legado, este trabalho traz um estudo inicial na direção de uma *léxico-estatística textual*, denominação por nós aqui proposta. Esse enfoque parte não só dos trabalhos de Biderman (1978, 1996), como também do que ensinou, de modo igualmente pioneiro, Lothar Hoffmann (1998, 2007), sobre *Estatística da Linguagem*.

Nascido na Alemanha em 1928, Hoffmann foi estudioso dos léxicos técnico-científicos em diferentes idiomas. Sempre enfatizou o bom potencial das análises quantitativas dos textos especializados, das suas convencionalidades e das suas terminologias. Para ele, a Estatística da Linguagem é uma disciplina da Linguística, que trata de aspectos quantitativos do uso da linguagem e do sistema linguístico, com apoio de procedimentos estatísticos. Tais procedimentos tornam-se importantes "naquelas situações em que se trata de compreender a língua em funcionamento, ou seja, **em textos**. (...) Sua pretensão teórica consiste em **modelar a comunicação linguística como um processo de probabilidades**" (HOFFMANN, 2007, p. 61-62; grifos nossos).

Ressaltamos que essa *léxico-estatística textual* não corresponde, de modo exato, à estatística lexical que se observa em trabalhos de Processamento de Língua Natural (PLN), embora muitas técnicas sejam iguais. Por outro lado, há uma inter-relação bastante forte desse enfoque de Hoffmann com princípios, métodos e procedimentos da Linguística de Corpus (LC), tal como apresentada no Brasil por Berber Sardinha (2004). Entretanto, não se pode dizer que sejam abordagem ou metodologia idênticas.

O diferencial dessa perspectiva de *léxico-estatística textual*, conforme entendemos, centra-se no papel do texto e da ambiência discursiva nos quais se inserem os elementos lexicais. São considerados, portanto, gêneros textuais e discursivos. Tratar-se-ia, assim, de um enfoque que busca contemplar "elementos linguísticos da coerência, marcas da sintaxe do texto

[1] Conferir <seer.fclar.unesp.br/alfa/article/download/3300/3027>.

e outros elementos" (HOFFMANN, 2007, p. 62). Isso para que o objeto **texto** não seja subsumido em meio ao todo de um *corpus* – geralmente um acervo gigante – explorado em larga escala. Essa *léxico-estatística textual*, vale destacar, não se colocaria como algo melhor ou pior frente ao PLN ou à LC. Trata-se, sim, apenas, de marcar algo diferenciado, uma proposta estatística para estudo do léxico historicamente relacionada com a Linguística do Texto.

Assim inspirado, tendo tal perspectiva como um ponto de chegada desejado, este texto traz uma busca de dados. A busca segue uma práxis depreendida das indicações de Biderman e de Hoffmann e visa subsidiar a construção de um vocabulário controlado (VC). Esse VC apoiará, como uma referência, a escrita de definições em um dicionário experimental de português como língua estrangeira (doravante PLE). Tal dicionário[2], em versão protótipo, já é oferecido para acesso gratuito *on-line* e visa atender, preferencialmente, aprendizes que sejam falantes de línguas distantes do português, como, por exemplo, a língua alemã e as línguas orientais.

Encerrando esta introdução, alertamos que o enfoque estatístico da linguagem deve ser entendido como mais um ponto de referência para embasar o estudo e a descrição de elementos lexicais, e não como um fim em si mesmo. Portanto, é sempre bom lembrar que o resultado da análise estatística vale como um auxílio para o lexicólogo e para o lexicógrafo, o que é extensivo ao linguista em geral. Nessa direção, o que se obtém como um resultado matemático não pode ser interpretado como uma medida absoluta, 100% generalizável, que imponha determinadas ações ou cerceie escolhas. Afinal, o dado estatístico, bruto, descontextualizado, tende a ter pouco valor para qualquer pesquisa, em qualquer ciência.

DICIONÁRIOS, DEFINIÇÕES E LINGUAGEM FAMILIAR

Os enunciados definitórios presentes em um dicionário devem conter palavras mais fáceis de compreender do que aquelas que nele estão definidas. Essa é a afirmação de Zgusta (1971), que sustentava esse princípio básico da Lexicografia monolíngue voltada para estudantes de uma língua estrangeira (LE): o princípio de usar palavras simples ao redigirem-se paráfrases definitórias. Ratificando esse pensamento, Lew (2010), em estudo recente, afirma que "**muitas palavras pouco frequentes na definição podem criar problemas de compreensão**" (p. 293, grifo nosso). Assim, graus de frequência de determinadas palavras em uma dada língua podem ser fortes indicadores para sua maior ou menor inteligibilidade.

[2] http://www.ufrgs.br/textecc/porlexbras/

Nesse sentido, uma situação de dificuldade com a consulta de dicionários pode ser confirmada, por exemplo, entrevistando-se usuários, que afirmam que muitas vezes não conseguem entender os sentidos de uma palavra dicionarizada. Isso pode ocorrer, entre outros motivos, porque a definição é difícil de ser compreendida. É o que apontou uma pesquisa realizada por Wingate (*apud* WELKER, 2004): 80% de estudantes de francês e 95,7% de estudantes de alemão (ambos como LE) afirmaram que a principal razão para não usarem um dicionário monolíngue da língua estrangeira que estudavam é que "as definições são muito difíceis", mesmo que seus níveis de proficiência nessas LEs fossem adequados.

Conforme Ilari (1997), quando um dicionário monolíngue objetiva definições que permitam a compreensão de palavras desconhecidas, essa definição deve ser construída numa linguagem mais familiar e corrente do que a da palavra definida. Ilari tinha em mente, à época, um dicionário monolíngue de perfil escolar, dirigido para falantes nativos do português do Brasil. Sem desconsiderar diferenças entre usuário falante nativo ou um estudante de PLE com nível de proficiência razoável, vale refletir sobre a condição e a natureza dessa "linguagem mais familiar" para aprendizes, sejam nativos ou estrangeiros.

O JORNAL POPULAR COMO ACESSO A UM PORTUGUÊS POPULAR ESCRITO

No contexto do português brasileiro (PB), avançando um pouco na compreensão sobre o que viria a ser "linguagem mais familiar" em meio ao cenário do ensino de PLE, buscamos subsídios em pesquisas do projeto PorPopular[3]. Nesse projeto, o objetivo principal é descrever padrões de linguagem de textos jornalísticos no PB que têm como público-alvo leitores das classes B, C e D, com escolaridade correspondente ao Ensino Fundamental completo e pouco hábito de leitura.

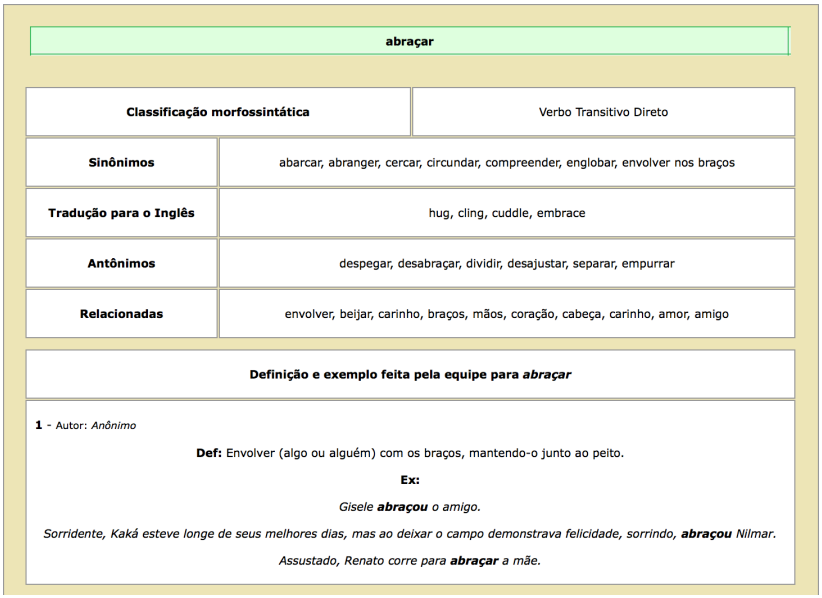
No PorPopular, buscam-se as características do que seria um Português Popular Escrito (PPE), entendido como um padrão de norma culta escrita do PB. Esse PPE estaria associado a um novo tipo de jornal, o jornal popular. Esse novo tipo jornalístico mostra-se como um tipo intermediário, situado entre o jornal tradicional ou de referência, voltado para as camadas mais letradas da população, e o jornal de estilo sensacionalista, um jornal de cunho apelativo, do tipo "espreme que sai sangue", conforme Amaral (2006, p. 15-20).

[3] Acesso ao *site* do projeto e informações sobre os *corpora* em <http://www.ufrgs.br/textecc/porlexbras/porpopular/>.

No momento, integram o *corpus* PorPopular textos dos jornais *Diário Gaúcho* e do *Massa!* de Salvador (BA). Ambos têm grande circulação e notável volume de vendagem em suas regiões. Ao escolhê-los, pressupôs-se que o sucesso de ambos esteja ancorado em uma fórmula de texto que se aproximaria de uma "linguagem familiar" ou de compreensão mais acessível para seus leitores. A propósito, quando o *Diário Gaúcho* foi lançado, seu primeiro editorial dizia que o jornal se propunha a ser "barato, completo e digno, com **linguagem clara e fácil**" (AMARAL, 2004, p. 121; grifo nosso).

O DICIONÁRIO COLABORATIVO DE PORTUGUÊS PARA ESTRANGEIROS (DCPE)

Face a demandas por materiais para o ensino de PLE, desde 2010, membros da equipe PorPopular trabalham na proposição de bases para um protótipo do DCPE[4]. Imagina-se um dicionário semibilíngue, centrado na habilidade de compreensão de leitura, voltado para estudantes com nível intermediário de proficiência em português. Esses usuários deverão ser capazes também de compreender informações auxiliares em inglês e definições em português. Trazemos, na Figura 1, uma proposta inicial de verbete-piloto do DCPE:



abraçar

Classificação morfossintática	Verbo Transitivo Direto
Sinônimos	abarcar, abranger, cercar, circundar, compreender, englobar, envolver nos braços
Tradução para o Inglês	hug, cling, cuddle, embrace
Antônimos	despegar, desabraçar, dividir, desajustar, separar, empurrar
Relacionadas	envolver, beijar, carinho, braços, mãos, coração, cabeça, carinho, amor, amigo

Definição e exemplo feita pela equipe para *abraçar*

1 - Autor: *Anônimo*

Def: Envolver (algo ou alguém) com os braços, mantendo-o junto ao peito.

Ex:

*Gisele **abraçou** o amigo.*

*Sorridente, Kaká esteve longe de seus melhores dias, mas ao deixar o campo demonstrava felicidade, sorrindo, **abraçou** Nilmar.*

*Assustado, Renato corre para **abraçar** a mãe.*

Figura 1. **Verbete do DCPE para a palavra "abraçar".**

[4] Disponível em <http://www.ufrgs.br/textecc/porlexbras/di/>.

Nos verbetes, há a classificação morfossintática da palavra, sinônimos, o equivalente em inglês, antônimos e palavras relacionadas. Há também exemplos de uso, que são buscados no *corpus* PorPopular com a ferramenta *Analisar*, disponível na aba *Enviar Definição*[5]. Nesse caso, a inteligibilidade da definição da palavra dependerá do uso de uma linguagem **clara e fácil**.

Por isso, a identificação de um conjunto lexical restrito, ou seja, a construção de um VC, torna-se uma etapa de pesquisa importante. Para ilustrar um provável problema definitório, pensando em estudantes de PLE, vejamos parte do verbete da mesma palavra antes ilustrada no DCPE, agora posta no dicionário Houaiss. Embora voltado para falantes nativos do PB, a escassez de dicionários monolíngues de PLE pode levar esses estudantes a consultar o Houaiss, entre outros dicionários:

> **abraçar** Datação: 1255 S 96|v.|verbo 1 Regionalismo: 25|t.d. e pron.|transitivo direto e pronominal envolver (algo ou alguém) com os braços, mantendo-o junto ao peito; **cingir** com os braços; dar abraços **recíprocos.** Ex.: o pai abraçou o filho|as crianças abraçaram-se ao pai.

Figura 2. Segmento do verbete para *abraçar* no dicionário Houaiss.

Os grifos nas palavras "cingir" e "recíprocos" são nossos. Essas duas unidades podem ser apontadas como possíveis complicadores da compreensão da definição. Ambas, justamente, não constaram no VC que produzimos a partir de um ensaio estatístico para detectar palavras mais frequentes e, em tese, mais acessíveis à compreensão do consulente que temos em mente.

Esse ensaio. em busca da identificação de um conjunto lexical acessível, é o que relatamos nas próximas seções.

ENSAIO ESTATÍSTICO PARA BUSCA DE UM VC: REFERENCIAIS

Em um processo inicial de tentativa e erro, buscamos uma lista de palavras que pudesse espelhar o perfil de um vocabulário básico do PB a ser usado como VC. Mas, antes do relato propriamente dito da busca, importa demarcar algumas reflexões que envolvem tentar encontrar o que se poderia perceber como um *núcleo lexical* do PB.

Primeiro, conforme bem ensinou Biderman (1996) em seu trabalho intitulado, justamente, *Léxico e vocabulário fundamental*, torna-se válida a diferenciação entre um vocabulário mais frequente multiuso (algo desejável para o nosso VC) e um *vocabulário disponível* ou de uso possível: um

[5] Basta digitar a palavra no espaço "Palavra a analisar" e clicar em "Analisar". Acesso em <http://www.ufrgs.br/textecc/porlexbras/di/enviardef.php>.

"vocabulário (...) constituído de palavras de baixa frequência e pouco estáveis, mas usuais e úteis" (BIDERMAN, 1996, p. 31). Portanto, não estamos atentando apenas para as frequências das palavras, mas para a relevância e necessidade delas no processo de escrita das definições. Além disso, com a dimensão do sentido/significado das palavras (ou lexemas ou lexias), sabemos que os campos semânticos das palavras modificam-se ao longo da história. Portanto, estamos cientes de que almejar alcançar a identificação desse vocabulário *multiuso nuclear* é uma tarefa complexa. Afinal, o núcleo da significação se desloca com o tempo e com o uso (BIDERMAN, 1978, p. 146).

Assim, sendo a língua mutável, polissêmica e variável, tentar reduzi-la a um mero repertório do tipo "lista" é uma tentativa fadada ao insucesso. Por isso os dicionários, por melhores que sejam, indo além de listas, são textos sobre línguas e sobre culturas. Ainda assim, os dicionários sempre nos oferecerão apenas um reflexo de um todo que não conseguem abarcar completamente. Mais do que listar palavras e tratar de seus sentidos ou significados, mediante paráfrases definitórias (também denominadas paráfrases explanatórias) e com o recurso de exemplos ilustrativos de uso (forjados pelo lexicógrafo ou trazidos *in natura* de textos), o dicionário terá que dar conta de fornecer as relações sintagmáticas e paradigmáticas (e pragmáticas) de uma unidade. Um VC, nesse contexto dialógico complexo, tende a ser apenas mais uma peça, coadjuvante, em meio a diferentes elementos e condições.

Partindo-se de um enfoque estatístico e de um recorte lexical X para um objetivo imediato Y, alertamos aqui, ainda, para algumas simplificações, propositais, ao longo de nossos experimentos em busca de bases para um VC. Elas dizem respeito à conceituação de elementos-chave no âmbito dos Estudos da Linguagem e dos Estudos de Lexicologia e de Lexicografia.

Haja vista tais reflexões, esclarecemos, *grosso modo*, para um melhor acompanhamento do nosso relato de pesquisa, que:

> - existem diferentes léxicos dentro do português, assim como em qualquer língua; eles são acionados pelos falantes em diferentes esferas sociais e situações. No entanto, os falantes sempre dominarão um mesmo núcleo lexical (*standard* ou padrão), que é comum a todos (BIDERMAN, 1978, p. 145). Portanto, o conjunto lexical que não varia de falante para falante, ou de lista a lista, no caso deste artigo, é o que tomamos a liberdade de chamar de núcleo;
> - o léxico é objeto da lexicologia, e o vocabulário, da lexicografia. Os itens que fazem parte do léxico de uma língua, ao serem coletados, compõem um vocabulário; neste trabalho, os itens que formam nosso vocabulário são palavras;
> - a noção de palavra é polivalente. Lidamos aqui com a face quantitativa da linguagem e trabalhamos no nível das frequências e

das ocorrências. Assim, entendemos palavras como lemas: por exemplo, o lema "querer" inclui conjugações desse verbo: "quero", "queres", "queria".
- não faremos quaisquer distinções entre sentido/significado ou entre palavra/vocábulo. No teste de uso de um VC, sublimamos o fato de que classes de palavras diferentes demandam perfis definitórios específicos.

DICIONÁRIOS PARA ESTUDANTES DE LE E VCs

Um dicionário para estudantes nativos de uma língua tem um feitio pedagógico similar, em alguns pontos, ao de um para falantes não nativos (HARTMANN e JAMES, 2001). Esse tipo de obra, denominada em metalexicografia estrangeira de *leaners'dictionary,* deve permitir mais a decodificação do que a codificação, auxiliando o usuário a sistematizar regularidades ortográficas, a perceber os exemplos de usos e as marcas de diferentes estratos socioculturais da comunidade linguística, entre outros. Sobretudo, deve conter definições que não frustrem o usuário e que o façam de fato compreender as palavras que consulta na obra.

Estudos apontam que, em diferentes línguas, as 5.000 palavras mais frequentes abarcam de 90% a 95% das palavras distintas (*tokens*) de um texto. Isso já foi mostrado em línguas como o russo, o francês, o inglês e o holandês. No entanto, o PB ainda carece de estudos nessa direção (HAZENBERG e HULSTIJN, 1996). Um dos recursos empregados, por exemplo, nos dicionários da Editora Oxford, é a seleção de um "Vocabulário Controlado de Definidores", que auxilia na redação das definições (AYTO, 1984). Isso é o que chamamos aqui de VC.

O objetivo desses VCs é tornar cada definição a mais clara possível, facilitando a compreensão e a produção das paráfrases definitórias. Para tanto, os lexicógrafos são instruídos a utilizar preferentemente as palavras ou itens que constam na listagem de um VC, que foram escolhidas pelos critérios frequência e relevância. Há abordagens diferenciadas para a conformação dessas listas, que variam de acordo com as concepções teóricas de seus autores, e o aproveitamento de uso dos VCs não é ponto pacífico.

Svensén (2009), por exemplo, aponta que definir usando vocabulário simples demanda mais espaço do que definir usando termos complexos. Por outro lado, afirma que fica evidente que, ao usar um VC, um conceito complexo está de fato sendo definido por meio de conceitos menos complexos. Por sua vez, Bogaards (2008) chamava atenção para o fato de que as palavras usadas em um VC, por estarem entre as mais frequentes, costumam ser polissêmicas.

Em que pese a importância dessas constatações, nada muda o fato de que a maioria das equipes que produzem dicionários voltados a

estudantes de LEs faz uso de VC. Isso também nos motiva a tentar produzir e a testar esse recurso e sua utilidade para um dicionário como o DCPE.

DELINEANDO UM VC PARA O PB

a) Partindo de *corpora*

Aparentemente, o número de 5.000 ou 3.000, ou mesmo 1.500 palavras, poderia basear a elaboração da lista de itens do VC. Entretanto, importa ressaltar que, conforme o que apuramos em pesquisa, não há estudos atuais, no âmbito dos Estudos da Linguagem, para o PB[6], quanto à fixação de um ponto de corte, em termos de frequência, para limites numéricos de um vocabulário básico. Biderman (1996, p. 39) endossava o critério mínimo de 40 ocorrências de cada unidade no seu *corpus* para que figurassem no seu dicionário de frequências. Essa obra, que permanece inédita até hoje, partiu de um *corpus* de 5 milhões de palavras cuidadosamente reunido pela autora. Ela, entretanto, também considerava a validade de se usar uma base de pelo menos 20 ocorrências.

Apesar da escassez de estudos linguísticos atuais a respeito, entendemos que, antes de realizar pesquisas diretas, para verificar, com sujeitos (estudantes de diferentes perfis de proficiência ou professores) seu conhecimento ou familiaridade em relação a um dado conjunto de palavras que se determine como básico ou nuclear, a exploração-piloto de alguns *corpora* do PB pode fornecer boas pistas. Afinal, dispomos, hoje, no Brasil, de recursos computacionais, de técnicas estatísticas avançadas e de ótimos acervos, cuidadosamente reunidos. Entre diferentes acervos, vale conhecer os *corpora* indicados no *site* do Projeto COMET[7] (Corpus Multilíngue para Ensino e Tradução) e o Lácio-Web[8]. O *Corpus Brasileiro*, organizado por Berber Sardinha, é o maior e mais recente disponível até 2013. Acessa-se em http://corpusbrasileiro.pucsp.br/cb/Inicial.html.

A despeito de especificidades dos atuais diferentes *corpora* disponíveis *on-line*, para utilizar esses acervos com o fim de gerar um VC para a criação de um dicionário de PLE centrado na habilidade de compreensão de leitura, acreditamos que a parte a ser examinada como amostra deva conter: (1) linguagem contemporânea e vocabulário ativo; (2) linguagem de simples compreensão; e (3) textos orais transcritos e textos escritos.

O Banco de Português[9], por exemplo, pareceu-nos um bom *corpus*

---

[6]Maiores informações podem ser encontradas nos artigos de Bacelar do Nascimento (2000, 2001), referenciados ao final deste trabalho.

[7] Disponível em <http://www.fflch.usp.br/dlm/comet/>.

[8] Disponível em <http://www.nilc.icmc.usp.br>.

[9] Disponível em <http://www2.lael.pucsp.br/corpora/bp/index.htm>.

para um experimento inicial. Pois, conforme Berber Sardinha (2004), é um *corpus* abrangente e contemporâneo, o que abarcaria os itens 1 e 3. Seus textos, em maioria, são de revistas e jornais; alguns vêm da literatura, de textos acadêmicos e de negócios; o segmento de dados obtidos de textos falados vem de conversas, reuniões, aulas, conversas telefônicas e entrevistas. Além disso, sua consulta é bastante facilitada para um trabalho exploratório inicial.

b) Partindo de dicionários e de outros acervos

Outra fonte que se pode aproveitar na busca do que seria um repertório lexical de maior inteligibilidade para o DCPE é o *Dicionário Ilustrado do Português* (2005), de Maria Tereza Camargo Biderman. Essa obra objetiva ser um dicionário para crianças no ensino básico e apresenta apenas palavras concretas. De acordo com a autora, palavras concretas seriam mais facilmente compreendidas por crianças, o que poderia indicar uma linguagem de simples compreensão, algo que se esperaria em um VC. Para sistematizar os itens lexicais desse *Dicionário Ilustrado*, contamos com a colaboração da professora Sandra Aluísio, do NILC-ICMC-USP, que já utilizou essa fonte no âmbito do Projeto Por Simples[10]. Esse dicionário, no seu todo, já serviu como uma referência para a simplificação lexical em estudos de PLN. Mais dados sobre pesquisas em PLN que envolvem simplificação textual com apoio informatizado e estatístico estão em Cândido, Oliveira e Aluísio (2009).

No quesito "linguagem de simples compreensão", a consulta a jornais populares, cujo vocabulário tenderia a ser, em tese, mais simplificado, também pode ser útil para um futuro VC. Um exemplo de fácil acesso é o *corpus* do jornal *Diário Gaúcho* (DG), já mencionado anteriormente. Com o DG, começou a nossa exploração no experimento que é a seguir relatado.

EXPERIMENTO 01: 3 listas de *corpora* diferentes

Já havíamos constatado o grau de dificuldade implicado na construção de um dicionário monolíngue de português com VC (KUHN, EVERS, FINATTO, 2012). Na ocasião, fizemos um levantamento das 3.000 palavras mais frequentes em três *corpora* distintos do PB, tomando por base a bibliografia sobre VC em dicionários de inglês como LE.

Os *corpora* para chegar à nossa lista de 3.000 itens foram: a) o *Banco de Português*; b) o *Dicionário Ilustrado do Português*, de Biderman (2005); e c) o

[10] Disponível em <http://caravelas.icmc.usp.br/wiki/index.php/Principal)>

corpus do jornal *DG*. Partindo desses dois *corpora* e da lista de entradas do dicionário de Biderman, produzimos uma lista de 3.000 palavras (entendidas palavras como *tokens*, considerando ocorrências e não formas lematizadas), que foram então cruzadas. A lista inicial, contendo os itens somados nessas 3 listas, possuía 9.250 palavras.

Já que uma lista de 9.000 itens, que trazia apenas *tokens,* tornava-se algo pouco operacional como referência para elaborar definições, pensamos em algumas soluções. Retiramos os nomes próprios, nomes de lugares, nacionalidades, verbos, substantivos e adjetivos flexionados (por exemplo, a ocorrência "acabou" foi incluída na lista de ocorrências de "acabar"; os plurais e femininos viraram singulares e masculinos). Depois disso, ainda restaram 8.625 palavras ao todo. É importante ressaltar que há, nas listas, locuções e palavras ligadas por hífen, que foram consideradas como uma palavra válida a considerar. Feito isso, cruzamos as listas entre si com a função COMPARAR COLUNAS, do Microsoft Excel. Desse cruzamento, geramos uma nova lista, que continha apenas as 1.024 palavras constantes, simultaneamente, nas três listas. Essa lista foi então analisada de acordo com critérios predefinidos pelo perfil específico do usuário do dicionário. Para tanto, duas professoras com experiência em ensino de PLE receberam a tarefa de manipular a lista e foram instruídas a eliminar palavras que considerassem muito difíceis para estudantes de português de nível intermediário, considerando estudantes cujas línguas maternas fossem distantes do português.

Acrescentamos que considerar o perfil do usuário do DCPE é fator determinante para a estruturação do dicionário – tanto no plano microestrutural quanto macroestrutural. Do mesmo modo, ter em mente o consulente parece ser decisivo para o processo de filtragem da lista de itens do VC. Além do perfil de conhecimentos e de necessidades do usuário, cruzar frequências dos itens da nossa lista em *corpora* distintos dos que tomamos como ponto de partida também trouxe uma nova visão sobre a feição de um vocabulário nuclear da língua portuguesa para esse fim específico de ensino de LE.

Em resumo, após obter a lista de 3.000 palavras mais frequentes de cada uma das fontes acima citadas, feito o cruzamento dos itens em comum, restaram **1.024** palavras (incluindo expressões ou *lexias complexas*). Delimitado esse conjunto, tentamos formular algumas definições para um conjunto-teste de verbetes usando apenas a lista desse provável VC com 1.024 unidades. Isso resultou em enunciados definitórios como o exemplificado a seguir:

APITO - Frequência no Corpus Brasileiro (fonte: WordSketch Engine): 1.004 ocorrências (0,9 por milhão)

Verbete		
Entrada	Apito	s.m.
Definição	Pequeno objeto que produz um som alto quando você o assopra	
Contextos de uso	1. O juiz já estava de *apito* na boca. 2. Não ouvi o *apito* do guarda-noturno.	

Após definir 10 entradas-teste do dicionário com apenas os itens do nosso VC de 1.024 itens, algumas dificuldades surgiram. A principal foi a falta de palavras para concluir a apresentação das informações sobre muitas das entradas selecionadas. "Assoprar", por exemplo, não constava da lista de 1.024 itens. Aqui, cabe a ponderação sobre o papel coadjuvante dessas listas frente à tarefa da redação de definições.

EXPERIMENTO 02: PARTINDO DO VC *OXFORD 3000*™

Neste experimento, traduzimos a *Oxford 3000*™, um VC utilizado pelos lexicógrafos da Editora Oxford para a redação de definições. A lista representa, de acordo com seus autores, um vocabulário básico para o estudante de inglês como LE. A tradução visava verificar se a busca de seus equivalentes em PB poderia ajudar a identificar as necessidades de comunicação de um estudante de PLE de nível intermediário.

A lista traduzida resultou em **3.853** palavras, sendo 8 delas em inglês, empréstimos que já estão presentes no dia a dia do falante brasileiro. O VC *Oxford 3000*™, na verdade, é composto por 3.354 palavras que, muitas vezes, possuem mais de um significado, de modo que não houve correspondência estrita em termos numéricos. Devido a isso, quando uma palavra na lista em inglês pudesse ser, por exemplo, um substantivo ou um verbo, colocamos essas duas possibilidades na lista em português. A lista da Oxford foi copiada e traduzida em blocos de 500 unidades para português pela ferramenta de tradução do Google. Em seguida, revisamos a tradução gerada, palavra por palavra, com o auxílio de dicionários e de buscas na internet. Esse trabalho esteve a cargo de uma acadêmica do nosso curso de Letras/Inglês, com consulta a tradutoras habilitadas.

Foram observadas, também, possíveis traduções que cada palavra da lista poderia ter, já que as palavras, como estão apenas listadas, carecem de contexto ou indicação de uso. Após revisada, a lista gerada em português saída da lista em inglês passou por uma limpeza para que dela fossem

retiradas palavras em duplicidade e palavras que parecessem muito distantes ou de um uso mais restrito. Assim, palavras – e locuções e expressões polilexicais – que pareciam "mais complexas" e "menos acessíveis" para nosso usuário foram excluídas ou substituídas por sinônimos mais "usuais". Esse processo de substituição foi feito com o auxílio de dicionários e de consulta ao *corpus* do Banco do Português, considerando opções mais frequentes. O Quadro 1 ilustra o processo:

Oxford 3000™	Tradução	Comentáriopara revisão
ahead	**à frente**	Poderia ser "adiante"
unless	**a menos que**	Poderia ser "senão"

Quadro 1. Amostra do trabalho de revisão de tradução.

EXPERIMENTO 03: COMPARAÇÃO ENTRE LISTAS DO PB E DA TRADUÇÃO DO VC *OXFORD 3000*™

A partir da tradução no Experimento 02 e da compilação das três listas do Experimento 01, realizamos uma última comparação e filtragem em um último ensaio. Utilizamos a nossa primeira lista "suja", de **9.250** palavras, e a comparamos com a lista traduzida do Experimento 02, com **3.853** palavras. Utilizamos, novamente, a ferramenta COMPARAR COLUNAS do Microsoft Excel.

LISTA DO EXPERIMENTO 02	LISTA DO EXPERIMENTO 01
a si mesmo	**abaixo**
a vapor	abaixo de
abaixo	abaixo-assinado
abandonar	abalar
abandono	**abandonar**

Quadro 2. Exemplo de comparação realizada entre as listas do Experimento 01 e Experimento 02; as palavras grifadas ocorrem nas duas listas (item 3 abaixo).

Ficamos, assim, com três listas diferentes:

1. itens que aparecem na lista do Experimento 02, mas não na do Experimento 01 (**945 palavras**);
2. itens que aparecem na lista do Experimento 01, mas não na do Experimento 02 (**5.698 palavras**);
3. itens que aparecem tanto na lista do Experimento 01 quanto na do Experimento 02 (**2.837 palavras**).

CONSIDERAÇÕES FINAIS E PERSPECTIVAS

A adoção de um VC na escrita de definições de dicionários em língua inglesa e o sucesso que esses dicionários fazem entre os estudantes de inglês LE mostram que talvez haja mais prós do que contras na adoção de um vocabulário restrito. Entretanto, chegar-se a um VC para o PB parece uma tarefa hercúlea, ainda que hoje haja, à disposição, recursos computacionais e ferramentas oriundas do PLN ou da LC.

A melhor saída para o impasse de se cobrir uma necessidade imediata, e também para subsidiar a elaboração definitória no DCPE, talvez fosse assumir uma amostra textual considerada adequada, de acordo com os tipos de textos que os usuários do DCPE mais terão de enfrentar em suas atividades de leitura. Nesse ponto, o texto do jornal diário, seja o popular, seja o tradicional, parece ser uma boa alternativa como referência e como *corpus*. Seria preciso, entretanto, assumir as peculiaridades do gênero jornalístico selecionado e ponderar com professores de PLE sobre as escolhas dos itens para um VC, tal como fizemos em nossos experimentos iniciais, aqui relatados. Assim, mostra-se importante reintroduzir o texto nessa estatística, especialmente quando se pensa em uma lista de palavras descontextualizadas, ladeada cada unidade apenas por um número ou percentual de ocorrências em um dado *corpus*.

Nos nossos experimentos, optamos pela consideração da palavra gráfica, mas inserimos, como unidades lexicais, elementos ligados por hífen, locuções e expressões. Sobre esse isolamento, vale perguntar: uma palavra como atenção, por exemplo, poderia ser apresentada, para o usuário do DCPE, sem a companhia do verbo prestar? Essa separação vale para o desenho dos itens de um VC? Nesse caminho, incluindo essas e outras perguntas, acreditamos, fica comprovada a relevância do estudo em busca de um VC, e, principalmente, a necessidade de revisitar trabalhos pioneiros sobre estatística linguística e buscar dados de levantamentos de frequências colocados em diferentes tipos de dicionários relativos ao PB.

Para explicar o significado de palavras, lexicógrafos precisam lançar mão de algumas palavras de alto nível, de alta generalização, e de uma série de itens gramaticais. É preciso lidar com polissemia, homografia, variação e pragmática. O desafio é imenso. Para enfrentá-lo, precisamos investir em pesquisas que nos apontem um norte de boas práticas e que nos permitam um lastro de critérios, para obter produtos que satisfaçam usuários e que sejam, teórica e metodologicamente, coerentes. Isso especialmente neste momento de efervescência de demandas por materiais didáticos de PB para o ensino de PLE. Por fim, entendemos que será útil, nessa busca:

a) voltar às ideias de vocabulário passivo, ativo e disponível e ao desafio de delimitar unidades lexicais para recursos como um VC ou outros

repertórios (ponderando se se deve considerar locuções e lexias complexas, por exemplo);

b) considerar uma ambiência textual do léxico que se examina e levar em conta o tipo de texto/discurso nas amostras de *corpora* que se construam ou que se tomem como referência para observações de frequências de palavras;

c) recuperar e divulgar os trabalhos de base estatística que se ocuparam perfis do léxico de *corpora*, sejam novos ou antigos, dos Estudos da Linguagem ou do PLN, aprendendo com seus limites e acertos, ponderando-os à luz de um enfoque que possa abrigar aspectos textuais e discursivos;

d) incentivar a produção de dicionários de frequências do PB e recomendar a publicação de obras atinentes, tal como o *Dicionário de Frequências do Léxico do Português Contemporâneo*, de M.T.C Biderman, infelizmente inédito até hoje.

REFERÊNCIAS

AMARAL, M. F. *Lugares de fala do leitor no Diário Gaúcho*. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2004.

AMARAL, M. F. *Jornalismo Popular*. São Paulo: Contexto, 2006.

AYTO, J. The Vocabulary of Definition. In: GOETZ, D.; HERBST, T. *Theoretische und praktische Probleme der Lexikographie*. Munique: Hueber, 1984. p. 50-62.

BACELAR DO NASCIMENTO, M. F. *Léxico Multifuncional Computarizado do Português Contemporâneo.* Lisboa: Centro Cultural Casapiano. 2001. p. Feira de Projectos, promovida pela Comissão Nacional do Ano Europeu das Línguas.

BACELAR DO NASCIMENTO, M. F. Um novo léxico de frequências do português. In: _____ *Volume de Homenagem ao Professor Herculano de Carvalho (no prelo)*. [S.l.]: [s.n.], 2001.

BACELAR DO NASCIMENTO, M. F.; PEREIRA, L. A. S.; SARAMAGO, J. *Portuguese Corpora at CLUL*. in Second International Conference on Language Resources and Evaluation. Atenas: [s.n.]. 2000. p. 1603-1607.

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

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

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

BIDERMAN, M. T. *Dicionário Ilustrado de Português*. São Paulo: Editora Ática, 2005.

BOGAARDS, P. *Frequency in learners' dictionaries*. Proceedings of 2008 Euralex. Barcelona: Universitat Pompeu Fabra. 2008. p. 1231-1236.

CÂNDIDO, J. A.; OLIVEIRA, M.; ALUISIO, S. M. *Simplifica:* um Sistema Web de Autoria de Textos Simplificados. Proceedings of WEBMEDIA. [S.l.]: [s.n.]. 2009. p. 55-58.

COWIE, A. P. *English Dictionaries for Foreign Learners:* A History. Oxford: Oxford University Press, 1999.

HARTMANN, R. R. K.; JAMES, G. *Dictionary of Lexicography*. Londres: Routledge, 2001.

HAZENBERG, S.; HULSTIJN, J. H. Defining a Minimal Receptive Second-Language Vocabulary for Non-native University Students: An Empirical Investigation. *Applied Linguistics*, Oxford, 17, n. 2, 1996. 145-163.

HOFFMANN, L. Anwendungsmöglichkeiten und bisherige Anwendung von statistichen Methoden in der Fachsprachenforschung. In: HOFFMANN, L.; KÄLVERKÄMPER, H.; WIEGAND, H. E. *Fachsprachen:* Ein internationales handbuch. Berlim: De Gruyter, 1998. p. 240-241.

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

HOUAISS. *Dicionário Eletrônico Houaiss da Língua Portuguesa*. Rio de Janeiro: Objetiva, 2009.

ILARI, R. **A** *Linguística e o Ensino da Língua Portuguesa*. São Paulo: Martins Fontes, 1997.

KUHN, T. Z.; EVERS, A.; FINATTO, M. J. B. *Uso de vocabulário controlado em dicionários de português como língua estrangeira em formato on-line:* uma experiência em andamento para uso de aprendizes coreanos. In: TEIXEIRA E SILVA, R.; YAN, Q.; ESPADINHA, M. A.; LEAL, A. V. (orgs.) *III SIMELP. A formação de novas gerações de falantes de português no mundo*. Macau: Universidade de Macau, 2012. CD-Rom.

LEW, R. Multimodal Lexicography: The Representation of Meaning in Electronic Dictionaries. *Lexikos*, v. 20, p. 290-306, 2010.

NASCIMENTO, R.; ISQUIERDO, A. Frequência de palavras: um diagnóstico do vocabulário de redações de vestibular. *Alfa*, v. 47, 2003.

PINHEIRO, G. et al. *Mapeamento de Projetos de Corpora no Brasil*. III Encontro de Corpora IEL. Campinas: UNICAMP. 2003.

SVENSÉN, B. *A Handbook of Lexicography:* The theory and Practice of Dictionary-Making. Nova York: Cambridge University Press, 2009.

WELKER, H. A. *Dicionários:* uma pequena introdução à lexicografia. Brasília: Thesaurus, 2004.