FINATTO, M. J. B. (2012). Projeto PorPopular, frequência de verbos em português e no jornal popular popular brasileiro In: **As Ciências do Léxico: lexicologia, lexicografia, terminologia, volume VI**.1 ed.Campo Grande, MS : Editora da UFMS/Laboratório de Edição da FALE-UFMG, 2012, v.VI, p. 227-244.

PROJETO PORPOPULAR, FREQUÊNCIA DE VERBOS NO PORTUGUÊS E NO JORNAL POPULAR BRASILEIRO

Maria José Bocorny Finatto*

Introdução

> *Todos nós temos a impressão de que a liberdade na utilização das palavras é absoluta, que as restrições na escolha das palavras são inexistentes e que as nossas possibilidades são ilimitadas. André Clas, 1994.*
>
> *As palavras não ocorrem aleatoriamente em um texto. R. Harald Baayen, 1996.*

As palavras que perfazem o todo de um texto escrito, sob um olhar ingênuo, parecem ter sido ali colocadas apenas em função de um livre arbítrio do seu redator. Entretanto, descontando-se, talvez, alguns usos obrigatórios de palavras associadas a um dado tema sobre o qual se escreva e as palavras que se precisa usar para manter a língua minimamente reconhecível em suas regras gramaticais mais básicas, a seleção das palavras acaba mostrando-se como resultado de toda uma série de obrigatoriedades. Desse modo, a autonomia de uso de um dado vocabulário, em um texto escrito qualquer, parece que será sempre relativa. Aqui vemos a fronteira entre o léxico e a gramática, pois há faixas de escolhas mais e menos livres e faixas de obrigatoriedades, se quisermos nos fazer entender por nosso leitor.

Supondo um *continuum* do mais obrigatório ao mais livre quanto à escolha de palavras para um texto escrito, podem ser colocados vários questionamentos sobre a configuração, natureza e quantidades das palavras empregadas em diferentes situações de escrita e de comunicação. Por exemplo, haveria como afirmar que, via de regra, há uma dada proporção, mais ou menos constante, de uso de verbos, de adjetivos ou de preposições num dado tipo de texto? Haveria como depreender, verificado um dado padrão de uso de uma palavra X ou de uma classe de palavra determinada, que conjunto de condições estabeleceria a sua presença maior ou menor em um dado universo textual?

Nesse questionamento, que tem embutido em si o desejo de alguma previsibilidade para além de uma descrição, há a necessidade, antes de qualquer coisa, de confirmar se existe, mesmo, algum padrão ou pelo menos alguma frequência reiterada de uso de uma dada palavra ou expressão (o que incluiria reconhecer padrões de não uso) em um dado *corpus* ou *corpora*. Ao partir dessas idéias, este capítulo visa justamente: a) apresentar um pesquisa, em andamento, sobre o texto do jornal popular brasileiro no que tange ao vocabulário empregado, o Projeto *PorPopular;* a) trazer, a título de ilustração, pequena amostra do perfil quantitativo e qualitativo de uso de verbos no jornal popular frente ao jornal tradicional e frente a outros tipos de texto.

Algumas questões nos guiam – e assombram – uma é, por exemplo, a seguinte: haveria uma proporção de verbos que diferenciasse os textos escritos em português de acordo com diferentes tipos de texto? Ao tentar enfrentar a pergunta, estamos cientes de que há toda uma série de fatores complexos envolvidos no processo nada singelo de se escolher dizer ou escrever uma palavra e não outra, sem contar as condições para uma maior "verbosidade" ou maior "adjetivação", sendo

* Universidade Federal do Rio Grande do Sul, Depto. de Lingüística, Filologia e Teoria Literária, Porto Alegre – RS – pesquisador CNPq, *e-mail*: mfinatto@terra.com.br

importante não perder de vista que há toda uma exterioridade constitutiva do texto envolvida aqui. Permeando as escolhas das palavras a usar, estão as condições de produção do texto, elementos pragmáticos, finalidade do texto, suposições e pressuposições de conhecimento do leitor, etc. Ademais, o que está mobilizado em termos de palavras e de classes de palavras escritas importa tanto quanto o como essa mobilização está posta, quando e como as palavras foram arranjadas ou organizadas, incluindo-se aqui a sua ordem em frases, ênfases, preferência por voz ativa ou passiva, modalizações diversas, tempos verbais, etc. Enfim, há aqui toda uma amplitude envolvida, e nosso propósito, frisamos, é apenas o de estimular a curiosidade de quem pretenda aventurar-se a explorar o universo lexical de um texto tal como o do texto do jornal popular brasileiro.

Dito isso, explicitamos a organização que segue este texto. Primeiro, na seção 1, há uma brevíssima apresentação do Projeto PorPopular. Depois, alguns trabalhos pré-existentes sobre frequência de verbos em *corpora* e uma pequena amostra de dados do nosso *corpus*. Por fim, considerações sobre os indicativos e limites da observação sobre verbos, concluindo-se com perspectivas para estudos futuros a partir do que já se oferece no *website* da pesquisa.

1- Sobre a pesquisa PorPopular

O projeto PorPopular (http://www6.ufrgs.br/textecc/porlexbras/porpopular/) envolve a descrição e estudo de padrões do vocabulário de textos de jornais diários populares brasileiros voltados para públicos de menor poder aquisitivo e menor hábito de leitura, tomando como objeto de estudo o jornal *Diário Gaúcho* (*DG*), apenas em sua versão impressa[1].

Esse tipo de jornal, aparentemente, constitui uma nova variedade, intermediária entre o jornal popular sensacionalisa ou de cunho jocoso e o jornal tradicional, dirigido às camadas mais letradas e abastadas da população. Segundo Amaral (AMARAL, 2006), a partir de 2000, um novo conceito de jornal popular busca proximidade e empatia com o público-alvo, caso do *DG,* o que lhe confere um "tom" diferenciado (SILVA; FINATTO, 2009).

A descrição e os estudos do Projeto foram feitos à luz de referenciais teórico-metodológicos da Lingüística de Corpus (BERBER SARDINHA, 2004), havendo uma tendência bastante forte para o tratamento estatístico do vocabulário dos textos. A investigação, em diferentes etapas, contou com apoio do CNPq, do programa PIBIC-UFRGS CNPq e com auxílio da Secretaria de Ensino a Distância (SEAD-UFRGS).

Para começar, partimos do pressuposto que o texto desse tipo de jornal integraria um uso específico da língua portuguesa, que denominamos, provisoriamente, *Português Popular Escrito*. Por vocabulário, entendemos o conjunto geral de palavras que conforma um dado uso da língua, sem distinguir entre itens gramaticais e itens lexicais ou vocabulário ativo ou passivo, tipos de palavras e formas/flexões de palavras. Esse vocabulário foi observado em termos de freqüências e de distribuição e uso de formas ao longo de um *corpus* de estudo, *corpora* de contraste e de *corpora* de referência. Quanto ao nosso entendimento sobre padrões, consideramos distribuições de uso recorrentes e também as construções e combinatórias de palavras reiteradas ao longo do corpus, contrastando-se frequências e distribuições com outros tipos de *corpora*.

Considerando que a grande maioria das pesquisas em *corpora* sobre vocabulário, sobre neologismos, outros elementos mórficos ou gramaticais da língua portuguesa, feitas no Brasil até hoje, não utilizaram materiais do jornalismo popular, identificamos uma lacuna de conhecimento. Assim, objetivamos também o oferecimento gratuito e *on-line* de dados desse *corpus*, incluindo o acesso a ele mesmo. A figura 1 a seguir traz a tela de abertura do *site* em que se oferecem vários desses dados. Na guia EXPERIMENTE, por exemplo, é possível explorar as edições do jornal

---

[1] Pelo que pudemos apurar, a versão *on-line* do DG tem várias diferenças em relação à versão impressa, inclusive pela seleção lexical. Como a circulação da versão impressa chega a atingir 150 mil exemplares dia, preferimos essa versão para estudo, apesar da maior dificuldade para a geração do *corpus*.

mediante palavras de busca, observar sua inserção em frases e obter listas de palavras de uma edição completa ordenadas alfabeticamente ou por frequência.

Figura 1 – tela inicial do PorPopular em www6.ufrgs.br/textecc/porlexbras/porpopular

O *corpus* inicial foi uma amostra seriada de edições diárias dos 12 meses do ano de 2008 do jornal. Mais adiante, esse *corpus* acabou sendo complementado por um semestre de amostra do ano de 2009 e por alguns meses do ano de 2010. O material textual do jornal foi cedido, com a devida autorização, para armazenamento, compartilhamento e publicação *on-line*. Com o auxílio de sistemas informatizados e das ferramentas gratuitamente disponíveis no *site* do grupo TEXTQUIM (http://www6.ufrgs.br/textquim/index.php), foram observadas as palavras mais freqüentes, mês a mês, em amostras que incluíram, a cada mês, dez diferentes dias da semana. Quando pronta, a lista de freqüências (*wordlist*) trouxe informações sobre a quantidade de palavras (*tokens*) e as diferentes formas como cada palavra se repete (*types*).

A partir das *wordlists*, foi feita a comparação do *corpus* DG com dados de padrões de vocabulário colhidos do Banco do Português (http://www2.lael.pucsp.br/corpora/bp/) um *corpus* de linguagem geral do Brasil, que possui mais de 120 milhões de palavras (tomando como base o ano 2000). No contraste feito, percebemos que as palavras mais freqüentes (DE, A, O, E, QUE) se repetiam em ambos os *corpora*, variando apenas a sua posição no *ranking* das mais empregadas.

Além disso, foram feitas contrastes com padrões de vocabulário do jornal *Zero Hora* (ZH), publicado pela mesma empresa do DG e dirigido a um público de maior poder aquisitivo. O contraste teve três etapas: primeiro, a partir das listas de palavras de cada *corpus*, em seguida, com as combinações de palavras mais freqüentes (*n-gramas*) em cada um e, por fim, com textos sobre um mesmo assunto publicados nos dois jornais. O paralelo entre as *wordlists* do DG *e* ZH mostrou que, tal como na comparação com o Banco do Português, não houve diferença significativa entre as palavras gramaticais mais freqüentes (DE, A, O, E, QUE).

Assim, em termos de unidades gramaticais, no que se refere à presença mais básicos, vimos que o texto escrito do DG – descontada a parte de imagens e ilustrações da publicação - não diferia muito de outros textos. As diferenças ficavam mais por conta da seleção lexical e palavras lexicais mais empregadas. No caso do DG, a palavra mais presente é VAGA(S), principalmente no sentido de EMPREGO/TRABALHO. Além disso, o tamanho de frase tende a ser bastante curto no DG, se comparado, por exemplo, ao jornal ZH.

2- Alguns resultados – frequências de palavras no jornal Diário Gaúcho

Antes de apresentar alguns dos nossos dados, para haver um termo de comparação, é conveniente alguma perspectiva de trabalhos pré-existentes. Sem poder trazer vários, resgatamos dois importantes.

2.1. M.T.C. Biderman – de 1978 a 2001

A partir do trabalho pioneiro de Maria Tereza C. Biderman sobre perfis frequenciais do vocabulário do português do Brasil realizado em 1978 (BIDERMAN, 1978), percebe-se um quadro inicial de investigações sistemáticas, especialmente quantitativas, sobre o léxico. Com esse trabalho, também é possível depreender como se colocavam, no panorama da Linguística brasileira dos anos 70, a Estatística Lexical, em particular, e a Lexicologia, em geral. São dessa obra as seguintes constatações:

> Considerando a língua como entidade abstrata ou conjunto de possibilidades, então os dados qualitativos se superpõem aos quantitativos porque o sistema de inter-relações e oposições preexiste a seu emprego. Daí dizer-se que a freqüência das unidades nada tem a ver, em princípio, com a estrutura da língua, que se sustenta unicamente por sua coesão interna. Nesse sentido a Matemática não consegue descobrir estruturas ou apreender significações, mas proporcionar magnitudes, relações de magnitudes e distribuições.
> (...) Ora, se a freqüência não influi no sistema enquanto tal, pode influir no seu dinamismo e no seu vir-a-ser – já se provou que há estreita relação entre freqüência e desgaste de formas. Por isso o estudo quantitativo é desejável não só para compreensão da deriva da língua como também do próprio equilíbrio do sistema e da dinâmica inerente a ele.
> (...) Não é propriamente uma teoria da linguagem, mas uma seleção daqueles aspectos teóricos necessários para o tratamento estatístico e computacional do léxico. Por exemplo, um dos problemas Teórico-práticos cruciais do lexicólogo que quer utilizar o computador é saber decidir-se pela unidade léxica básica – aquela entidade conhecida como *palavra* no âmbito das línguas indo-européia. Este é um dos tópicos a que a Professora Biderman deu bastante atenção.
> (BORBA, 1978, apresentação in BIDERMAN 1978, p. IX-XI)

> (...) nos discursos individuais, orais ou escritos, ou em qualquer corpus lingüístico, notam-se, portanto certas constantes na distribuição dos signos. De maneira geral qualquer texto evidencia o mesmo tipo de distribuição léxica, para falar apenas nesse nível dos signos lingüísticos. Assim, seguindo a ordem de distribuição das palavras das mais altas às mais baixas freqüências, verifica-se que as formas usadas mais frequentemente são os vocábulos gramaticais;(...)seguidos de palavras de uso menos e menos freqüente, em escala descendente. Chega-se, enfim, a palavras específicas, cuja freqüência diferirá de texto para texto, até atingirmos os "hapax legomena" que caracterizam cada corpus linguístico em particular, só havendo coincidência entre eles esporadicamente."(BIDERMAN, 1978, p.9)

Em 2001, seu livro de 1978 tem nova edição, com reformulações e omitida a apresentação do Prof. Francisco da Silva Borba (BIDERMAN, 2001). Mas, antes disso, em 1998, a autora (BIDERMAN, 1998) já relatava como se deu o processo de composição do seu *Dicionário de Frequências do Léxico do Português Brasileiro Contemporâneo*, obra infelizmente, nunca publicada na íntegra.

Tendo organizado um *corpus* de 5 milhões de palavras com textos publicados entre os anos de 1950 e 1990, Biderman relata que apenas 42.212 palavras diferentes compunham o *corpus*, tendo sido excluídos nomes de pessoas e de lugares. Além disso, a maciça maioria de palavras era composta por palavras gramaticais ou instrumentais, verbos auxiliares e modalizadores. Nesse universo de 5 milhões de palavras, feito de textos escritos de diferentes tipos, temos notícia de que havia 6 mil verbos. Não fica claro se seriam 6 mil ocorrências de verbos ou se seriam 6 mil tipos de verbos diferentes, considerando a sua forma infinitiva canônica. Valendo a primeira opção, os verbos representariam, no seu conjunto, uma fatia de 12% do *corpus*. Além disso, Biderman informa que:

a) cerca de 42% do total de ocorrências do *corpus* é constituído por pouco mais de mil palavras, sendo essas as mais frequentes na língua; b) 80% de qualquer texto é constituído por essas

mil palavras mais frequentes; e, c) **os verbos têm estatuto diferenciado nesse *corpus*** (BIDERMAN, 1998, p.166-171, grifo nosso).

Quanto aos 20 verbos mais utilizados, há a seguinte informação que vemos aqui na tabela 1:

Ranking	Verbo – forma lematizada infinitivo	Número de ocorrências
1º	ser:	50.222
2º	ter:	34.586
3º	ir:	28.965
4º	estar:	27.746
5º	poder:	16.593
6º	dizer:	15.445
7º	haver:	15.004
8º	fazer:	14.279
9º	dar:	10.792
10º	ver:	10.391
11º	saber:	10.247
12º	querer:	9.986
13º	ficar:	8.605
14º	achar:	7.980
15º	dever:	7.758
16º	falar:	5.259
17º	chegar:	4.628
18º	precisar:	4.039
19º	começar:	3.596
20°	olhar:	3.383

Tabela 1 – Os 20 verbos mais frequentes no *corpus* de 1950-1990. (Fonte: BIDERMAN, 1998, p. 172.)

Embora alerte que não tenha feito experimentos comprobatórios, a autora acredita que, seguindo um trabalho para o francês que organizou *corpus* semelhante ao seu em termos de fontes textuais, é muito provável que esses 20 verbos sejam os mais freqüentes no português do Brasil, independentemente do tipo de texto e até do fato de se tratar de linguagem oral ou escrita.

Conforme a autora,

> A lista dos verbos mais freqüentes é encabeçada pelos auxiliares **ser**, **estar**, **ter**. Até o verbo **ir** registrou um elevado número de valores modais e aspectuais, razão para estar também nos primeiros lugares da hierarquia dos verbos usuais. Constam dessa lista ainda verbos modalizadores como **poder**, ou vicários, e/ou suportes como **fazer**, **dar**; entre os de significação plena apenas **dizer**, **falar**, **olhar** e **ver**. (BIDERMAN, 1998, p. 174)

Além disso, Biderman comenta que

> "O curioso a respeito desses verbos é que os totais de ocorrências são determinados apenas por umas tantas formas do verbo, como já constatamos em um verbo de altíssima freqüência como *querer*. Essas formas são sempre as mesmas flexões de tempo, modo e pessoa: o infinitivo, o gerúndio, as 3as. Pessoas do singular do presente e do pretérito perfeito e imperfeito; a seguir, são mais freqüentes: as 3as. Pessoas do plural dos mesmos tempos e na mesma seqüência. Em uns raros verbos a primeira pessoa do singular do presente e do pretérito perfeito ocorre muitas vezes. Todas as demais formas do paradigma verbal têm freqüência muito baixa (1,2) ou nula. Pode-se concluir que a virtual possibilidade de 74 formas para os 6 mil verbos da língua portuguesa registrados nesse corpus não passa de virtualidade. Essa potencialidade não ocorre jamais nem mesmo com aqueles vinte verbos de altíssima freqüência. **Essa constatação permite asseverar que é preciso rever integralmente a questão do ensino de conjugações verbais nas escolas primárias e secundárias para falantes nativos e também o ensino do verbo para estrangeiros.** (BIDERMAN, 1998,p. 174, grifo nosso).

2.2. Nascimento – 2001

No trabalho de Nascimento (2001), cujo título é *O vocabulário dos estudantes universitários: um estudo com base em redações de vestibular*, ainda que não haja uma pesquisa centrada em verbos ou mesmo em *corpora* especializados, há uma exemplar incursão sobre aspectos quantitativos do léxico do português brasileiro. Ademais, são feitos contrapontos com os trabalhos de Biderman antes citados. A abertura do trabalho, bastante afinada com o que está na introdução deste texto, resume muito bem o tipo de enfoque a ser empreendido, tanto que é aqui reproduzida:

> Quando se trata de falar, ou de escrever, quem já parou, por exemplo, para pensar no número de palavras que utiliza diariamente? Em como este número difere de acordo com as pessoas, situações, lugares, atividades, ou mesmo normas de conduta a que os indivíduos estão sujeitos? Será verdade o que propaga o senso comum sobre as mulheres serem mais tagarelas e que, por isso, seus textos escritos tendem à prolixidade? Ou então, que os jovens da atualidade possuem, de fato, um repertório vocabular bem reduzido e não dominam o padrão escrito culto da língua? (NASCIMENTO, 2001, p.1)

Lembrando que, em 2001, já havia instrumentos computacionais muito facilitados para auxiliar a observar quantas e quais palavras há num texto, a autora frisa que, mesmo quando os recursos ainda não eram acessíveis, desde a época de Ulmann (1964), reconhecia-se a validade de enfoques estatísticos do vocabulário. Entretanto, conforme salienta, também esse autor reconhecia, em 1964, que poucos linguistas estariam em condições de acompanhar uma matematização mais estrita acoplada às suas investigações. Não obstante, a informatização e o desenvolvimento de enfoques combinados entre estudos do léxico e de Estatística já tinham rendido vários frutos até os anos 2000. Por isso, a autora declarava que a

> Lexicologia, a Lexicografia e a Terminologia, três áreas do saber ligadas ao léxico, foram as que mais se beneficiaram dos refinamentos tecnológicos de que atualmente dispomos, fazendo que fossem abertas novas portas a pesquisas que contam com grandes bases de dados lingüísticos. (NASCIMENTO, 2001, p .2)

Ao trazer um quadro sobre trabalhos na linha tecnológica e estatística, Nascimento (2001), com justeza, também assinala o pioneirismo de Biderman e registra que é dela o ensinamento de que "mediante resultados da Estatística Léxica ou Léxico-estatística, podem-se fazer uma série de previsões e constatações sobre o funcionamento da língua e sobre os elementos gramaticais presentes nos discursos orais ou escritos."

Para realizar a sua pesquisa com redações de vestibulandos, Nascimento (2001) apoiou-se nos trabalhos desenvolvidos em Lisboa pelo VPF (Vocabulário do Português Fundamental), em 1987, e nas pesquisas desenvolvidas por Biderman, no Brasil, por ocasião da elaboração do *Dicionário de Freqüências do léxico do português contemporâneo*, publicadas em alguns artigos e em livro.

Nascimento (2001, p. 61) buscava uma resposta científica para uma constatação impressionística - mas recorrentemente referida: a de que o vocabulário utilizado nas redações de vestibulandos seria muito pobre para o nível de cultura que se esperava de redatores com escolaridade equivalente ao Ensino Médio completo. Ao procurar refutar ou confirmar tal impressão, acabou reunindo e descrevendo um rico *corpus* e realizando comparações entre seus achados e os registros do Dicionário de Biderman, o qual, como já citado, abarcava um *corpus* de 5 milhões de palavras.

Nascimento, assim, explorou 450 redações de vestibulandos aprovados para ingresso no ensino superior em 1999. O conjunto de textos corresponde a um universo de 113.638 palavras, sendo: 53.238 palavras de redações de candidatos com ingresso em instituição particular e 60.400 em

uma universidade pública. A tabela a seguir oferece uma boa noção geral sobre os dados do vocabulário manifestado pelos universitários em seus textos, considerando faixas de frequência e tipos de palavras:

Tabela 2 - Distribuição por intervalos de frequência: alta, média e baixa (Fonte: NASCIMENTO, 2001, p. 84)

Frequência Alta **F>=20**	**Frequência Média** **F 10 a 19**	**Frequência Baixa** **F 1 a 9**
691 unidades **(470 plenas e 221 instrumentais)**	**605 unidades** **(537 plenas e 68 instrumentais)**	**9.855 unidades** **(9.778 plenas e 77 instrumentais)**
84. 879 ocorrências	**8.083 ocorrências**	**20.676 ocorrências**

As cinco palavras mais frequentes são DE, QUE, E , A, e O e seu emprego representa, sozinho, 17% do total do *corpus*. Uma representação que bem sintetiza a diferença de observações entre Nascimento e Biderman é a que se pode conferir na próxima tabela:

Tabela 3 - Comparativo da Distribuição Geral das Palavras nos dois *Corpora* (Fonte*:* NASCIMENTO, 2001)

DICIONÁRIO DE FREQUÊNCIAS (DIF) *Corpus 5 milhões*	***VOCABULÁRIO DOS ESTUDANTES (VEU)*** *Corpus 113.638*
Frequência • 1078 palavras mais frequentes: 42% • outras palavras: 58%	***Frequência*** • 100 palavras mais frequentes: 51,3% • 359 palavras mais frequentes: 66,7%
Número de unidades • 42.212 unidades léxicas diferentes • 10.452 unidades de frequência 1 (25%)	***Número de unidades*** • 11.151 unidades léxicas diferentes • 5.813 unidades de frequência 1 (52%)

Quanto aos 20 verbos mais utilizados, a autora nos traz a seguinte comparação:

Tabela 4 - Comparativo dos vinte primeiros verbos

20 verbos mais frequentes **Dicionário Biderman**	**20 verbos mais frequentes** **Vocabulário redações**
1º:ser [50.222], 2º:ter [34.586], 3º: ir [28.965], 4º:estar [27.746], 5º: poder [16.593], 6º: dizer [15.445], 7º:haver [15.004], 8º:fazer [14.279], 9º: dar [10.792], 10º: ver [10.391] , 11º: saber [10.247], 12º: querer [9.986], 13º: ficar [8.605], 14º: achar [7.980], 15º: dever [7.758}, 16º:falar [5.259], 17º:chegar [4.628}, 18º: precisar [4.039], 19º: começar [3.596], 20º: olhar [3.383].	1º: ser [3.770], 2º: ter [1.006], 3º: estar [702], 4º: poder [567], 5º: fazer [414], 6º: haver [330], 7º: ver [293], 8º: vir [254], 9º: saber [251], 10º: viver [239], 11º: dever [238], 12º: ir [231], 13º: dizer [214], 14º: mostrar [183], 15º: possuir [159], 16º: existir [148], 17º: chegar [147], 18º: comemorar [139], 19º: descobrir [134], 20º: dar [130].

Os verbos em comum são 13: *SER, TER, ESTAR, PODER, FAZER, HAVER, VER, SABER, DEVER, IR, DIZER, CHEGAR* e *DAR*. Conforme alerta Nascimento, o conjunto desses 20 verbos, em suas flexões, foi lematizado e, portanto, o número de ocorrências de cada uma das formas inclui todas as flexões existentes no *corpus*. O total de frequência desses 20 verbos é de 9.549 ocorrências,

o que corresponde a 8,4% do *corpus* (113.638). Infelizmente, a autora não indica o quanto percentual corresponde o total do uso de verbos em meio ao todo do *corpus*.

Por fim, no que tange a uma observação mais geral sobre a incidência de verbos ao longo do *corpus* de Biderman e o de Nascimento, vale registrar a seguinte observação (NASCIMENTO, ISQUERDO, 2003, p. 76):

> "todos" os verbos cuja ocorrência é maior na literatura tecnocientífica e na literatura jornalística - *ser, ter, ir, poder* e *dever* - também são os mais freqüentes no vocabulário dos estudantes. Isso faz supor que esses cinco últimos verbos provenientes das literaturas tecnocientífica e jornalística sejam indispensáveis a qualquer tipo de texto, justificando-se, pois, a sua presença significativa no vocabulário dos ingressantes universitários.

2.3. O verbo no Diário Gaúcho

Com apoio do listador de palavras TEXTQUIM (www.ufrgs.br/textquim, Caixa de Ferramentas), do programa Wordsmith Tools 3.0 (SCOTT, 2001) e do etiquetador morfossintático MXPOST, criado para o português do Brasil por pesquisadores do NILC (AIRES et. al. 2000), foram verificadas as freqüências e percentuais de verbos em relação ao número total de palavras no *Corpus PorPopular* - Diário Gaúcho –amostra 06 meses - edições completas de 2008 – versão impressa, esse *corpus* atinge 974.672 palavras.

O etiquetador MXPOST, originalmente concebido para o inglês (RATNAPARKHI, 1996), foi adaptado para lidar com o português do Brasil por Aires et al. (2000). Segundo o *site* do projeto Lácio Web, um repositório de *corpora* e de recursos para sua exploração (http://www.nilc.icmc.usp.br/lacioweb/), a melhor precisão para esse etiquetador, ao lidar com textos dos diferentes cadernos do jornal *Folha de São Paulo*, foi de 96.98% no Caderno Agrofolha, que é bastante padronizado e com vocabulário restrito; a pior é foi de 94.39% no caderno MAIS, caderno com textos literários e analíticos e vocabulário diversificado. Escolhemos esse sistema em função de sua precisão previamente testada e do acesso gratuito no *site* do NILC, Núcleo Interinstitucional de Lingüística Computacional da USP (http://www.nilc.icmc.usp.br/nilc/index.html)

Abaixo, está um exemplo das etiquetas atribuídas pelo MXPOST como verbo (VERB) a um trecho de uma notícia do DG. Depois do trecho, aparece o que o MXPOST retornou da marcação, com destaque em negrito para os verbos – pois todas as classes estão identificadas (ADV= advérbios, ART= artigos, etc).

> Nossa intenção **é trabalhar** mais forte nessas áreas. De dez dias para cá, **começaram** a **ocorrer** mais crimes nessa região. **Vamos remanejar** policiais e, em alguns momentos, **haverá** ações específicas nos ônibus - **prometeu** o comandante do 31^{0} BPM, major Marcelo Mello. Ontem, às 6h40min, dois homens **ingressaram** no micro-ônibus da linha Guaíba-Porto Alegre, **pediram** para os passageiros **abaixarem** a cabeça e **executaram** Everton Dias de Azambuja, 19 anos. (12 verbos)
>
> Nossa_NP intenção_NP é_**VERB** trabalhar_**VERB** mais_ADV forte_ADJ nessas_PREP+PD áreas._N De_PREP dez_NUME dias_N para_PREP cá,_N começaram_**VERB** a_PREP ocorrer_**VERB** mais_ADV crimes_ADJ nessa_PREP+PD região._N Vamos_**VERB** remanejar_**VERB** policiais_N e,_, em_PREP alguns_ADJ momentos,_N haverá_**VERB** ações_N específicas_ADJ nos_PREP+ART ônibus_ADJ -_N prometeu_**VERB** o_ART comandante_N do_PREP+ART 31∫_NUME BPM,_ART major_N Marcelo_NP Mello._NP
> Ontem,_ADV às_PREP+ART 6h40min,_NUME dois_NUME homens_N ingressaram_**VERB** no_PREP+ART micro-ônibus_N da_PREP+ART linha_N Guaíba-Porto_NP Alegre,_NP pediram_**VERB** para_PREP os_ART passageiros_N abaixarem_**VERB** a_ART cabeça_N e_CONJ executaram_**VERB** Everton_NP Dias_NP de_PREP Azambuja,_NP 19_NUME anos._N (12 verbos)

Pelos trechos acima, vê-se que, dos **12** verbos que havia no segmento, o sistema MXPOST acertou **12**, o que nos dá uma precisão de 100%. De modal igual, a cobertura dos itens assinalados foi de 100%. *F-Measure* (a média entre essas duas medidas), neste caso, também é 100%. Aqui tem-se uma situação ideal. Esse desempenho tão bom, conforme registros dos criadores do sistema para o

português (AIRES et al., 2000) deve-se ao fato que ele foi "treinado" justamente com um *corpus* de textos de jornalísticos. Esse treinamento, em Processamento da Linguagem Natural (PLN), é denominado *Aprendizado de Máquina* e envolve a percepção de vários padrões associados aos usos de palavras em um *corpus*, que precisará ser grande o suficiente para que variações repetidas sejam percebidas automaticamente.

No começo deste texto, perguntávamos se haveria uma quantidade mais ou menos fixa para a o uso de verbos em um texto. A tabela seguir, apresenta os dados de três *corpora*, o tamanho de palavras de cada *corpus* (Tamanho), a ocorrência bruta de verbos em cada um (Verbos) e sua proporção de verbos em relação ao número de palavras (%verb/pal). Como contraste, um pequeno conjunto de textos da revista *SuperInteressante* que tratam de algum assunto relacionado à Química e o *corpus* PLN-Gold, formado por 1.024 textos de notícias do Jornal Folha de São Paulo (BRUCKSCHEN et al., 2008) e o nosso *corpus* DG de 2008 (Dgimpresso):

Tabela 5 – Ocorrências de verbos e percentuais nos diferentes *corpora*

Textos	Tamanho	Verbos	%verb/pal
Dgimpresso	974.672	155.266	15,93
RevSuperInt	59.585	9.224	15,48
PLNGold	338.441	51.306	15,16

Nesse conjunto de textos jornalísticos, nos três *corpora*, temos uma média percentual de 15,43% de verbos em relação ao número de palavras dos textos. Ao que parece, há uma tendência para ser este o tamanho da "fatia" de verbos em um texto do tipo jornalístico nesse universo em foco. E, um dado interessante, é que vemos uma tendência para que esse percentual seja mais ou menos estável em outros grupos de textos de jornais que já examinamos preliminarmente.

De outro lado, o DG, embora se coloque como algo *a priori* diferenciado, visto que compõe um novo gênero de jornal, segue o mesmo padrão quantitativo do jornal tradicional. Aqui vale destacar que temos, nessa comparação, o jornal *Folha de São Paulo*, bastante associado com um público de alta escolaridade e elevado poder aquisitivo (classes A e B), um jornal dirigido a leitores de menor poder aquisitivo, o DG (classes B, C e D), e textos da revista *SuperInteressante*. Esses últimos são bastante curtos, têm como público jovens leitores, especialmente estudantes do Ensino Médio e tendem a um perfil de leitores de classe média (classes A e B). Além disso, os verbos É, SER, SÃO, ESTÁ, TEM e PODE, nesta ordem, foram os verbos mais utilizados em todos os três *corpora* com distribuições diferenciadas.

4- Considerações finais e perspectivas

Ao longo do estudo com verbos no DG, e também com adjetivos[2] e outros elementos (inclusive elipses) (FINATTO, SCARTON, ROCHA, ALUÍSIO, 2011), encontramos vários trabalhos relacionados à prospecção lexical em *corpora*. Um trabalho que deve ser mencionado, na linha de comparação entre jornais populares e tradicionais, é o de Oliveira (2009). No entanto, ainda são poucos aqueles específicos sobre aspectos quantitativos/estatítiscos do uso de verbos. É a essa lacuna que pretendemos nos dedicar, especialmente no território das construções recorrentes com verbos. Além disso, vemos a necessidade de também categorizar os verbos – e também adjetivos e outros elementos - identificados ao longo dos diferentes tipos de texto desse jornal, tal como em reportagens, notícias e colunas assinadas. Do mesmo modo, fica a curiosidade em verificar se esse percentual de uso de verbos , em torno de 15%, manter-se-ia quando examinarmos o texto de jornais populares de linha jocosa, como, caso do jornal *Diarinho* (http://www.diarinho.com.br/).

---

2 No trabalho acima citado Finatto, Scarton, Rocha e Scarton (2011), vemos, por exemplo, que o DG exibe menos adjetivos que um jornal tradicional, o ZH, o que contraria expectativas sobre um estilo, em tese, mais "emotivo" do jornal popular. Esse trabalho está disponível no *site* PorPopular.

De todo modo, acreditamos que o projeto PorPopular, no que já reúne e oferece no seu *site*, cumpre um importante papel. Esperamos que os colegas dos estudos do léxico se interessem por ele e que se sintam estimulados a outras explorações com os textos de jornais populares brasileiros. Um produto do PorPopular é o protótipo do Dicionário de Português como Língua Estrangeira, alimentado com o *corpus* DG (http://www6.ufrgs.br/letras/dicionarioportuguesle/), cobrindo, por hora, o recorte temático "futebol". Por fim, reiteramos o mérito dos trabalhos da saudosa colega M.T.C. Biderman aqui citados, pioneira da estatística lexical e do trabalho com *corpora*, que muito deixou para ser comparado sobre o uso de verbos no português do Brasil. Esperamos que, algum dia, seu *Dicionário de Freqüências do Português* possa ser publicado.

5 – Agradecimentos

Ao CNPq pelo apoio e bolsas e às colegas do NILC-ICMC-USP, Carolina E. Scarton e Sandra Aluísio pelo apoio estatístico no estudo sobre verbos no DG e em outros *corpora.*

Referências

AIRES, Raquel V. X.; ALUÍSIO, Sandra M.; KUHN, Denise C.S.; ANDREETA, Maria de Lourdes B.; OLIVEIRA JR., Osvaldo. N. Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. In: *Proceedings of the Brazilian AI Symposium* (SBIA'2000), p. 20-22, 2000.

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

BAAYEN, R. Harald. *The Effects of Lexical Specialization on the Growth Curve of the Vocabulary.* Computacional Linguistics, Volume 22, Issue 4, December, MIT Press Cambridge, MA, USA, 1996.

BERBER SARDINHA, Tony. *Lingüística de Corpus***.** Barueri-SP: Manole, 2004.

BIDERMAN, Maria Tereza Camargo. *Teoria lingüística:* lingüística quantitativa e computacional. Rio de Janeiro: Livros Técnicos e Científicos, 1978.

_____. *Teoria lingüística:* lingüística quantitativa e computacional. Rio de Janeiro: Livros Técnicos e Científicos, 2001.

_____ Léxico e vocabulário fundamental. In: *Alfa*, n. 40, 1996.

_____. A face quantitativa da linguagem: um dicionário de freqüências do português. In: *Alfa*, n. 42, 1998.

BRUCKSCHEN, Mírian; SACCOL, Daniela. MUNIZ, Fernando; SOUZA, José Guilherme C.; FUCHS, Juliana. T.; INFANTE, Kleber; MUNIZ, Marcelo; GONÇALVES, Patrícia. N.; VIEIRA, Renata; ALUÍSIO, Sandra M. (2008). *Anotação Lingüística em XML do Corpus PLN-BR*. Série de Relatórios do NILC (NILC-TR-09-08). São Carlos - SP, Junho 2008, 39 p.

CLAS, Andre. Collocations et Langues de Spécialité. In: *Meta,* vol. XXXIX, 1994.

FINATTO, Maria José B.; SCARTON, Carolina. E.; ROCHA, Amanda; ALUISIO, Sandra M. (2011) Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In: *VIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 2011, Cuiabá - MT. Anais do STIL 2011. Cuiabá: Sociedade Brasileira de Computação, 2011. v. 01. p. 30-39.

NASCIMENTO, Roseli; ISQUIERDO, Aparecida Negri. Freqüência de palavras: um diagnóstico do vocabulário de redações de vestibular. In: *Alfa*, n. 47, 2003.

NASCIMENTO, Roseli Imbernom. *O vocabulário dos estudantes universitários: um estudo com base em redações de vestibular.* Dissertação (Mestrado em Letras) - UFMS. *Orientador:* Aparecida Negri Isquerdo, 2001.

OLIVEIRA**,** Márcia Regina A. R. Jornal Popular X Jornal Tradicional*:* Análise léxico-gramatical da notícia a partir da Linguística de Corpus Um estudo de casos dos jornais cariocas "O Globo" e "O Dia". In: *Veredas On-Line* – Linguística de Corpus e Computacional – vol. 2/2009, p. 07-19, 2009.

RATNAPARKHI, Adwait. A maximum entropy model for part-of-speech tagging. In: *Proceedings of the First Empirical Methods in NLP Conference*, 1996.

SILVA, Bruna. R.; FINATTO, Maria José. B. Português popular escrito: o vocabulário do jornal Diário Gaúcho. In: *Anais do X Salão de Iniciação Científica da PUCRS*. Porto Alegre: EDIPUCRS, 2009. p. 3332-3334, 2009.