BUSCA DE UM VOCABULÁRIO BÁSICO DO PORTUGUÊS DO BRASIL ESCRITO: LISTAS DE FREQUÊNCIA DE PALAVRAS – JORNAIS POPULARES, LINGUAGEM GERAL E TRADUÇÃO

Material de pesquisa para consulta *on-line*

elaborado por Érica Spagnolo e Maria José B. Finatto em fevereiro de 2014.
revisado por Gerônimo Loss Bergmann e Maria José B. Finatto em janeiro de 2018.

Para citar: FINATTO, M.J.B; SPAGNOLO, É. BERGMANN, G.L. (2018). **BUSCA DE UM VOCABULÁRIO BÁSICO DO PORTUGUÊS DO BRASIL ESCRITO: LISTAS DE FREQUÊNCIA DE PALAVRAS – JORNAIS POPULARES, LINGUAGEM GERAL E TRADUÇÃO**. Material de pesquisa para consulta on-line. Inédito. Porto Alegre. Janeiro de 2018, 06p. Disponível em: ww.ufrgs.br/textecc/porlexbras/porpopular/index.php

INTRODUÇÃO

Os dicionários da editora Oxford para aprendizes de inglês como língua estrangeira utilizam, como referência, uma lista de 3 mil palavras criada por essa editora. A "lista Oxford" representa, assim, um vocabulário básico (**VB**), com as palavras mais importantes e úteis para o estudante que têm em mente. Esse vocabulário, inclusive, baliza a escolha de palavras que fazem parte de uma definição em um verbete de dicionário Oxford, constituindo uma espécie de "vocabulário controlado". Sua proposta é a de ser um conjunto de palavras mais acessível para o entendimento do aprendiz. Veja um estudo sobre esse tipo de vocabulário em dicionários didáticos em:

Nunes, Paula Ávila; Finatto, Maria José Bocorny. Dicionários monolínguues para aprendizes de inglês como língua estrangeira:alguns elementos para o professor. In: **Horizontes de Linguística Aplicada**. Brasília Vol. 6, n. 2 (dez. 2007), p. 33-54. (para ler, acesse a nossa BIBLIOTECA VIRTUAL do PorPopular)

Nessa lista Oxford, oferece-se um conjunto básico das palavras mais frequentes na escrita no **inglês britânico**, independentemente do tipo de texto em que ocorrem[1] .Esse universo de palavras foi obtido a partir da seleção, por uma equipe de especialistas em línguas e ensino de inglês para estrangeiros, das palavras mais frequentes no *Bristish National Corpus* (**BNC**). O **BNC** que é um enorme acervo de textos em inglês disponível na internet.

Uma vez que ainda não encontramos um lista desse tipo para o português brasileiro (**PB**), realizamos um EXPERIMENTO INICIAL E EXPLORATÓRIO para tentar identificar um conjunto de palavras que pudesse corresponder a um vocabulário básico do **PB**, mais frequentemente empregado na modalidade escrita. Buscamos também um *corpus* brasileiro de textos que pudéssemos aproximar ao **BNC**.

[1]Para saber mais detalhes sobre como a lista Oxford foi feita, veja em: https://www.oxfordlearnersdictionaries.com/about/oxford3000

EXPERIMENTO INICIAL

No Brasil, contamos também com um enorme acervo de textos que representa o **PB**, o *Corpus Brasileiro*[2]. Esse *corpus* tem um conjunto de textos que soma um bilhão de palavras, o que seria o equivalente do **BNC**. Com ele, poderíamos, em tese, repetir o procedimento da lista Oxford. Esses acervos de textos são chamados de *corpus*, palavra em latim, plural *corpora.* Mas, no Brasil, há quem escreva corpus, sem precisar fazer o plural.

Dado que a produção da lista Oxford em inglês não estava detalhadamente descrita pela editora, resolvemos iniciar nosso trabalho voltando-nos para a lista em inglês e para um **teste inicial**. Outros experimentos nossos, na busca de um **Vocabulário Controlado** partindo de *corpora* do **PB***,* podem ser conferidos em um artigo nosso publicado[3] em 2014, na revista Trama. Também disponível na nossa BIBLIOTECA VIRTUAL do PorPopular.

Para tentar, **preliminarmente**, delimitar um conjunto de palavras que servisse como referência de um vocabulário básico do português escrito, partimos da lista de 3 mil palavras da Editora Oxford, traduzindo-a para o português. Esse procedimento, a tradução, naturalmente, pode ser (duramente) criticado, dado que envolve uma série de diferenças e a natureza peculiar da lista inglesa. Mesmo assim, resolvemos testar (e vivenciar) a opção de sua tradução, para então criticá-la. Como saber sem tentar?

Depois de traduzida, a nossa **lista Oxford em Português** seria comparada com uma outra, baseada em *corpus*, que corresponde a um conjunto de 3 mil palavras em português. Em vez de buscar o Corpus Brasileiro, optamos por um acervo menor. Escolhemos fazer um contraste com as 3 mil palavras mais usadas em uma amostra aleatória de textos de um jornal popular brasileiro, o jornal Massa!, publicado em Salvador/BA. Esse seria, para um teste inicial, nosso pequeno *corpus* do **PB**. O quadro a seguir sintetiza os objetos (listas) em contraste.

MATERIAIS CONTRASTADOS – LISTA 1 e LISTA 2

	LISTA 1	**LISTA 2**
Nº DE PALAVRAS	**3000**	**3000**
FONTE	**DICIONÁRIO OXFORD****	**70 NOTÍCIAS DO JORNAL MASSA***
LÍNGUA ORIGINAL	**INGLÊS BRITÂNICO**	**PORTUGUÊS BRASILEIRO**
TRADUÇÃO	**PORTUGUÊS*****	**ORIGINAL em PB**

*As 70 notícias do jornal Massa! foram retiradas da versão *on-line* do jornal e reunidas num *corpus* pelo Projeto PorPopular. Disponível em: http://www.ufrgs.br/textecc/porlexbras/porpopular/

**A lista do Dicionário Oxford está disponível *on-line* no *site*: http://oald8.oxfordlearnersdictionaries.com/oxford3000/

[2] http://corpusbrasileiro.pucsp.br/cb/Inicial.html
[3] http://e-revista.unioeste.br/index.php/trama/article/view/10345/7464

***A tradução da lista por nós produzida – e a descrição do processo de TRADUÇÃO E DE REVISÃO DE ITENS DA LISTA EM PORTUGUÊS estão descritas no nosso artigo de 2014 publicado na revista Trama antes citado. A lista em português e alguns comentários sobre sua produção encontram-se em: http://www.ufrgs.br/textecc/porlexbras/porpopular/files/Material_OXFORD-rev2017MJ.pdf

PASSOS DO EXPERIMENTO COMPARATIVO: SÍNTESE

1- Reunimos um *corpus* de amostra com 70 textos do jornal Massa! de 03 de janeiro a 06 de fevereiro de 2013. Esses textos tratam de temas bem variados, o que representa uma amostra de português popular escrito heterogênea. Os 70 textos, reunidos, resultam em 21.352 *tokens* (total de palavras) e 5.518 *types* (palavras diferentes). Todos os textos foram agrupados num arquivo único, que foi editado por Érica Spagnolo para ser submetido a ferramentas computacionais simples que observam e listam as palavras empregadas. No *site* do Projeto PorPopular, esse acervo de 70 textos que utilizamos pode ser conferido, acessado e percorrido[4] com algumas ferramentas *on-line*.

2- A amostra de 70 textos foi submetida a uma ferramenta computacional grátis *on-line* (disponível em: http://www.ufrgs.br/textecc/textquim/ferramentas.php) que gera lista de palavras por frequência ou por ordem alfabética de um texto ou de vários textos. Consideramos as 3 mil palavras mais frequentes nessa amostra do Massa! e com elas fizemos uma lista em ordem alfabética.

3- Traduzimos livremente[5] a lista de 3 mil palavras da Editora Oxford para o português. Esse processo foi bem complexo, pois se traduz uma lista descontextualizada de itens. Para a tarefa, tivemos a colaboração de Aline Maciel Pereira, então bolsista de Iniciação Científica (PIBIC-CNPq), e de duas tradutoras habilitadas nesse par de línguas.

4- Comparamos as duas listas, Oxford em Português e Massa, imaginando que **os elementos em comum entre ambas** poderiam, talvez, funcionar como um conjunto de palavras equivalente a um vocabulário básico (**VB**). Naturalmente, é preciso considerar que partimos de uma lista de palavras soltas/descontextualizadas em inglês – traduzida para o português – e de um conjunto de palavras cuja origem é um jornal popular da Bahia (BA).

RESULTADOS INICIAIS

Das 3 mil palavras que cada lista contém, Oxford traduzida e Massa!, **apenas 916 palavras são comuns** ou bem aproximadas entre as duas listas. Isto é, verificamos **apenas 30,53%** de semelhança entre as duas listas de vocabulário. Assim, 70% das palavras não

[4] Veja em:
http://www.ufrgs.br/textecc/porlexbras/porpopular/caixaferramentas_3.php

[5] O processo de tradução, seus procedimentos e avaliação, está descrito no nosso artigo da revista Trama de 2014 e também no material antes citado.

coincidem! Isso nos levou a observar que, provavelmente, a noção de um vocabulário básico (VB) de uma língua para a outra pode ser MUITO divergente. É possível também pensar que a comparação em foco seja do tipo "alho com bugalhos", isto é, que estamos colocando em contraste elementos por natureza profundamente distintos.

A equiparação dos dados que coletamos evidencia que esse vocabulário básico, "clonado do inglês britânico", seria BEM diferente da nossa pequena amostra de vocabulário em Português Brasileiro (PB) que escolhemos. Apesar de serem visivelmente de fácil compreensão no PB, conforme nosso julgamento pessoal, as 916 palavras comuns entre as duas listas carecem de avaliação cuidadosa.

A seguir, veja uma amostra aleatória, com algumas palavras em comum entre as duas listas, do universo de 916 itens. O destaque de cores visou gerar apenas conforto de leitura, sem qualquer relação com maior ou menor frequência das palavras. A impressão que temos, olhando a figura, é de que são palavras bem acessíveis e bastante frequentes na nossa língua.

advogado popular trabalho ensino ocasião dança paz jornal revolução de vez liberdade obrigado palavra rádio saúde vencer qualquer usar eletrônico bairro justiça líder mulher necessidade ajuda região executar opinião segurança fato bom melhorar país voto zona afastar social médico origem universidade não sair morte isso quando álcool fazer pagar bandeira falar lavar mãe tipo número poder bem cabelo cada fácil dar índice já passar receber encontrar talvez então bater local nenhum quem último valor feliz investigação ir viver cair

Imagem 1 – itens comuns da lista Oxford traduzida e da lista do jornal Massa

A seguir, na Imagem 2, veja uma amostra, também aleatória, de algumas palavras que só apareceram no levantamento da Editora Oxford (Lista 1). Notamos que há algumas palavras mais "eruditas", tais como VORACIDADE e HESITAR e PROBABILIDADE.

histórico parecer abandonado duro efeito igreja perdoar recordar zombaria romance usuário ameaçar hesitar suco queimar noivo quintal bagunça tímido explosão fábrica raramente voracidade filmar xingar forno junção unha levemente tempestade vantagem orgulhoso mentir atrair humor padre seriamente barulho nervoso jantar brincar ganhar beber ovelha data voar década dever motor lábio notar zunido vencedor imaginar rachar adivinhar obter parabéns manusear julgar sacudir brilhar olho caixa grosseiro sofrer músculo câncer tranquilamente hábil carregar urbano nadar tablete sujeira ausente selvagem xícara gorduroso construir economizar zero probabilidade nascido incomum esboçar louco maçã etiqueta investir mapa qualidade odiar voz tropical

Imagem 2- amostra de palavras que apenas ocorrem na lista Oxford traduzida

Agora, vejamos, na Imagem 3, algumas palavras que só aparecem na nossa amostra de 70 notícias do Jornal Massa! (Lista 2). Essas palavras, provavelmente, seriam mais específicas, mais relacionadas com o tipo de texto em foco e com seus temas. Neste ponto, é importante considerar que o texto-fonte desse vocabulário é uma notícia de um jornal popular, dirigido para leitores de menor escolaridade e que se trata de um jornal de uma determinada região do Brasil. Por isso, talvez, encontremos apenas aqui, nesse cenário ou *corpus*, palavras como ARRASTÃO, ZAGUEIRO, MACONHA e XIXI. Esse universo de palavras também oferece um interessante material de estudo.

votação agressão cachoeira decreto seguranças empresa zagueiro estragos tributos famílias atualizar fãs xerife flagrante bastidores busca hino arrastão galeria humano identificados eleições urnas dançarinas roubado jogadores ingressos comparecer torcedores anunciado secretarias judicial líderes valorização gravado localidades mandado balada medicamentos maconha policial utilizados camarote negociado motivos suspeitos novidades tentativas gastos nordeste trabalhadores ocorrências consumidores oportunidades dengue pacientes órgão homicídio paralisação imediações partido fiscalização petróleo incêndio jornalistas quadrilha quedas escolar quartel rádio feijão realizado divulgado recursos gente salários inflação regiões juntamente estádios socorrido tiros homenagem cidadãos traficando utilização grupos vestibulares irregularidades lágrimas vítimas falsas xixi aleluia polêmica

Imagem 3 – Amostra de palavras que ocorrem apenas na lista de palavras do *corpus* Massa!

PERSPECTIVAS: NOVOS ESTUDOS PARA O VOCABULÁRIO BÁSICO DO PORTUGUÊS BRASILEIRO

Este estudo piloto, aqui resumido, é uma TENTATIVA INICIAL DE EXPLORAÇÃO DE UM TEMA MUITO COMPLEXO, a identificação de vocabulário básico de uma língua. Igualmente complexo foi o procedimento de traduzir uma lista de palavras em inglês e tentar observar se o universo de palavras obtido seria próximo ou distante de uma lista original, "nativa" do português escrito, conforme vemos em jornais populares. A nossa lista Oxford traduzida e a lista Massa! não se correspondem em vários aspectos. Essa não-correspondência merece alguma reflexão, dada a oportunidade de aprendizagem que um trabalho de pesquisa nos oferece, mesmo que seja inicial.

O que sabemos é que a lista Oxford original foi produzida a partir de um grande *corpus* do inglês, o **BNC**, tendo sido dele "filtrada" por especialistas e professores. Afinal, conforme explicam, apenas a frequência da palavra no **BNC** não foi fator suficiente para que uma dada palavra figurasse nessa lista, qualificada como uma lista "***keywords***". Mais detalhes sobre como a Editora Oxford selecionou os itens estão em "*How are the keywords selected*?, no *site* https://www.oxfordlearnersdictionaries.com/about/oxford3000 . Nós também fizemos uma filtragem partindo de uma lista de frequências de um grande *corpus* do PB (conferir o artigo da

revista Trama de 2014). Vale verificar os resultados a que chegamos e relatamos neste artigo e quem foram os "filtradores" no nosso caso – professores de português para estrangeiros.

De qualquer modo, apenas o experimento comparativo entre a lista traduzida Oxford e a lista de palavras mais frequentes na nossa amostra do jornal baiano mostrou indicativos bem interessantes no que tange ao estudo de vocabulário/léxico com apoio estatístico/quantitativo. Para aprofundar a exploração e os questionamentos dela advindos, valeria voltar às duas listas, ponderá-las com mais vagar, e realizar outros contrastes. Por exemplo, poderíamos considerar as palavras mais usadas no jornal popular Diário Gaúcho (**DG**) – tendo-se com o **DG** um *corpus* com mais de um milhão de palavras – e/ou com um levantamento de frequências de um grande *corpus* do português do Brasil escrito. Para conhecer um pouco do **DG**, clique em EXPERIMENTE, em http://www.ufrgs.br/textecc/porlexbras/porpopular/.

Pesquisar na área da léxico-estatística é isso – tentativas, erros, acertos, aprendizagens e reflexão sobre dados que as frequências e distribuições das palavras nos revelam. Desafios não faltam, sobretudo no que se refere ao tema do que seria um vocabulário básico do português escrito do Brasil, especialmente se pensarmos no que poderia vir a ser um *corpus* de referência com um perfil de "linguagem simples". Esperamos que, em breve, possamos contar com alguma referência ou acervo sobre o que seria um universo de palavras do tipo básico ou simples do português escrito. Essa referência poderá ser útil, por exemplo, para fins de ensino de português para estrangeiros, assim como também para situações em que se precise promover a **acessibilidade da informação escrita – e do vocabulário** - para leitores de escolaridade limitada.

UFRGS, Porto Alegre, 03/01/2018.