RECONHECIMENTO DO VOCABULÁRIO DE JORNAIS POPULARES BRASILEIROS POR UM DICIONÁRIO COMPUTACIONAL DE ACESSO LIVRE

Maria José Bocorny FINATTO[*]
Oto Araújo VALE[**]
Éric LAPORTE[***]

▪ RESUMO: Relata-se um experimento de verificação da identificação de um universo de palavras do português popular escrito por duas versões de um dicionário computacional do português brasileiro (PB), DELAF PB 2004 e DELAF PB 2015. Esse dicionário computacional é gratuitamente acessível para ser utilizado em análises linguísticas do Português do Brasil e em outras pesquisas, o que justifica um estudo crítico. O universo vocabular provém do *corpus* PorPopular, composto por jornais populares, o *Diário Gaúcho (DG)* e o jornal baiano *Massa*! (MA). Do DG, partiu-se de um conjunto de textos com 984.465 palavras (*tokens)*, publicados em 2008, com ortografia desatualizada frente ao Acordo Ortográfico da Língua Portuguesa adotado em 2009. Do MA, examinou-se um universo com 215.776 palavras (*tokens)*, em publicações de 2012, 2014 e 2015, com todo o material na nova ortografia. A verificação envolveu: a) gerar listas de palavras diferentes empregadas em DG e MA; b) comparar essas listas com as listas de entradas das duas versões do DELAF PB; c) avaliar a cobertura desse vocabulário; d) propor modos de inclusão de itens não cobertos. Os resultados do trabalho mostraram, no DG, uma média de 19% de palavras diferentes (*types)* desconhecidas pelos DELAF PB 2004 e 2015. No MA, essa média ficou em 13%. A versão do dicionário repercutiu ligeiramente sobre o desempenho do reconhecimento de itens.

▪ PALAVRAS-CHAVE: Jornais populares. Léxico. Vocabulário. Dicionário computacional. Cobertura lexical. Reconhecimento de palavras. Português brasileiro.

Introdução

O texto escrito de jornais populares do Brasil é ainda pouco estudado no âmbito dos estudos da linguagem e um tanto desprezado como fonte de dados para o desenvolvimento de recursos computacionais ou mesmo para alimentar os grandes

[*] Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre - RS – Brasil. Doutora pelo Programa de Pós-Graduação em Letras. maria.finatto@ufrgs.br. ORCID: 0000-0002-6022-8408

[**] Universidade Federal de São Carlos (UFSCar), Centro de Educação e Ciências Humanas, São Carlos - SP - Brasil. Professor do Departamento de Letras. otovale@ufscar.br. ORCID: 0000-0002-0091-8079

[***] Université Paris-Est (UPE), LIGM, UPEM/CNRS/ESIEE/ENPC, Champs-sur-Marne - França. Institut d'électronique et d'informatique Gaspard-Monge. eric.laporte@univ-paris-est.fr. ORCID: 0000-0002-0984-0781

corpora que representam o português atual. Por outro lado, esse tipo de fonte, considerado um novo e importante tipo de veículo de comunicação (AMARAL, 2006), já tem gerado dados para algumas aplicações computacionais promissoras, associadas à descrição de usos do português, como vemos, por exemplo, no trabalho de Zilio (2015). Os diferentes jornais populares brasileiros também têm subsidiado estudos sobre novos formatos dos gêneros jornalísticos na área da Comunicação Social (TRISTÃO; MUSSE, 2013; SELIGMAN, 2008), assim como, entre linguistas, têm servido como fonte para o abastecimento de materiais de ensino de português como língua materna ou estrangeira (FINATTO; PEREIRA, 2014) ou como objeto de estudos críticos de discurso (SOARES, 2017).

O vocabulário presente nesse novo tipo de jornal, entretanto, representa um desafio para a cobertura de dicionários computacionais. Também conhecidos como léxicos computacionais, esses dicionários são concebidos com o fim específico de serem consultados por sistemas informatizados, funcionando como partes de programas específicos (cf. ZAVAGLIA, 2006). Dicionários ou léxicos computacionais, denominados em inglês *NLP dictionaries* (CHISHMAN, 2016), são os dicionários construídos especialmente para uso no processamento da linguagem e exigem a incorporação cuidadosa de toda uma série de dados de descrição linguística (cf. LAPORTE, 2013) para o seu bom funcionamento. Essas incorporações são processos de melhoria ao longo de diferentes edições ou versões desses dicionários. Entre diferentes dicionários computacionais, destacamos aqui o DELAF PB, gratuitamente acessível para ser utilizado em diferentes pesquisas, em Linguística e em Ciência da Computação, o que justifica um estudo crítico de suas diferentes versões.

Com essas prováveis lacunas em mente, no que se refere ao estudo do léxico de jornais populares do Brasil e a aplicações informatizadas que se dedicam ao tratamento do léxico escrito do português brasileiro, relatamos aqui um experimento de identificação das palavras[1] utilizadas nesse tipo de jornal, observando sua cobertura ou reconhecimento por duas versões de um dicionário computacional de acesso livre. Os jornais populares brasileiros que forneceram um universo de vocabulário para exame foram o *Diário Gaúcho* (DG) e o jornal baiano *Massa!* (MA). Os textos utilizados podem ser consultados, mediante palavras de busca, no *site* do Projeto PorPopular[2]. Mais informações sobre esses veículos, seus textos e vocabulário mais frequente encontram-se em Amaral (2006) e em Finatto (2012). Em Oliveira (2009), encontram-se referências sobre jornais populares de outras regiões do Brasil.

Conforme já citado, o recurso linguístico selecionado para o nosso teste de reconhecimento de vocabulário dos jornais populares foi o DELAF PB, um dicionário computacional do português brasileiro (PB) do sistema UNITEX (PAUMIER, 2016).

---

[1] A noção "palavra", no âmbito dos Estudos da Linguagem, é bastante controversa conforme já ensinou Biderman (1999,1998). O que seja uma "palavra" pode ser entendido de vários modos, daí haver as denominações "lexia", "lexema", "lema" ou "unidade/item lexical", "vocábulo", entre outras, para corresponder a diferentes facetas desse conceito.

[2] Os textos que utilizamos neste experimento podem ser consultados via ferramenta "gerador de contextos" em: http://www.ufrgs.br/textecc/porlexbras/porpopular/experimente.php

Foram utilizadas duas versões desse dicionário, o DELAF PB 2004 (MUNIZ, 2004) e o DELAF PB 2015 (VALE; BAPTISTA, 2015; CALCIA *et al.*, 2014). Embora este léxico computacional ainda seja pouco utilizado por linguistas brasileiros, há, de longa data, publicações no Brasil sobre este recurso (MUNIZ, 2004), como também sua aplicabilidade em diferentes pesquisas (como, por ex.: ALMEIDA; FERREIRA, 2007; VICENTINI, 2010), sejam elas no âmbito do Processamento da Linguagem Natural (PLN) e/ou dos estudos de Linguística Aplicada, o que também justifica o seu exame neste trabalho. Além disso, vale frisar, seu acesso é gratuito e favorece, inclusive, adaptações e inserções pelos usuários, conforme suas necessidades.

Os dicionários DELAF PB, assim, são recursos colaborativamente construídos e atualizados. Têm sido bastante úteis em diferentes aplicações, especialmente em produtos de PLN. Um exemplo de aplicação muito recente e bem-sucedida no Brasil dos dicionários DELAF PB encontra-se no trabalho de Paiva, Barbosa, Faria e Martino (2017). A pesquisa premiada desses autores, tendo reunido linguistas e cientistas da Computação, produziu um tradutor automático do português escrito do Brasil para a Língua Brasileira de Sinais (LIBRAS)[3].

Com o dicionário DELAF PB nas suas versões de 2004 e 2015, quisemos também mensurar a sua atualização. Apenas a versão mais recente está preparada para lidar com a aplicação do *Novo Acordo Ortográfico da Língua Portuguesa*. Os textos dos jornais populares que utilizamos foram produzidos antes (DG) e depois (MA) da nova ortografia. Assim, pudemos considerar também o impacto da presença de dois padrões de escrita sobre o desempenho do recurso computacional em foco.

Além do exposto até aqui, os intuitos gerais deste nosso trabalho são:

a) difundir os jornais populares brasileiros como fonte de pesquisa sobre o léxico na modalidade escrita;

b) contribuir com dados úteis para futuras ampliações e melhorias do dicionário DELAF PB.

Este texto, na sequência, traz: a) dados do *corpus* de jornais populares utilizado no experimento; b) uma revisão sobre os dicionários e sobre a operação de reconhecimento ou de identificação de palavras de um dado texto ou *corpus*; c) as etapas do trabalho e seus resultados com comentários e uma visão geral dos resultados; d) uma caracterização do perfil majoritário dos itens desconhecidos pelos dicionários e uma ponderação sobre como seu desempenho poderia ser melhorado.

[3] Essa investigação e seu produto, que utilizaram o dicionário DELAF PB, foram reconhecidos como o melhor trabalho – categoria Mérito Científico, durante edição de 2017 do evento "Encontro de Linguística de Corpus e Escola Brasileira de Linguística Computacional" (http://www.ufrgs.br/elc-ebralc2017).

C*orpus* textual sob exame e amostras utilizadas

Do DG, utilizamos, como ponto de partida, apenas uma parte do acervo disponível no *corpus* PorPopular. Desse *corpus* DG, selecionamos um universo de textos publicados em 2008 que representa 984.465 *tokens* (total de ocorrências). Tem-se, assim, um *corpus* com quase um milhão de palavras, o que confere um tamanho relevante nesse tipo de estudo. Os textos sob exame na amostra para verificação do vocabulário do DG são de tipos variados e correspondem ao que está publicado nas edições completas do jornal diário, conforme o seu formato impresso. Têm-se notícias de tipo geral, colunas de horóscopo, colunas esportivas, editoria policial e assuntos variados. Outras partes do *corpus* DG já foram utilizadas em estudos anteriores (ZILIO, 2015; FINATTO *et al*., 2011), para finalidades diversas.

Por outro lado, no *corpus* do jornal MA, partimos de um universo amostral menor e bem mais específico em termos de tipologias textuais: um conjunto de 724 textos composto apenas por notícias publicadas na versão *on-line* do jornal. Essas notícias tratam de temas variados[4]. Nosso ponto de partida no MA foi um conjunto de 215.776 *tokens* (total de ocorrências), empregados em textos publicados em 2012, 2014 e 2015.

Os textos do DG – de 2008 – conforme já citado, foram publicados **antes** da mais recente alteração ortográfica do português do Brasil. Mais detalhes sobre a repercussão dessa alteração, em termos quantitativos, sobre a variação ortográfica verificada em jornais, podem ser vistos em Flores e Finatto (2009). Para uma visão, em termos gerais, dos impactos dessa alteração ortográfica, vale conferir a coletânea organizada por Moreira, Smith e Bocchese (2009).

Como os textos do MA foram produzidos na vigência dessa última reforma ortográfica e os textos do DG não, tivemos de lidar com dois padrões de ortografia no nosso experimento. Assim, as palavras da amostra do DG de 2008 ainda trazem hífens, acentos e tremas (ex.: *agüentar*), eliminados na amostra do MA (ex.: *aguentar*). Essas diferenças nos levaram a utilizar duas versões dos dicionários, uma a cada padrão ortográfico. Entretanto, cabe frisar, o uso das diferentes versões do dicionário DELAF PB, de 2004 e de 2015, também serviu ao propósito da verificação da abrangência de atualização, em termos da cobertura/reconhecimento dos itens empregados nesses jornais populares, com e sem ortografia modificada.

Assim, realizamos a verificação com duas versões diferentes do dicionário: a versão DELAF PB 2004 (MUNIZ, 2004) e a versão DELAF PB 2015 (VALE; BAPTISTA, 2015; CALCIA *et al.*, 2014).

Os dicionários DELAF

O DELAF é, na verdade, um formato de dicionários computacionais que se originou de trabalhos realizados para a língua francesa pela equipe liderada pelo linguista

[4] Esta é a "Amostra 2" do jornal Massa no *corpus* PorPopular. Disponível para *download* em – http://www.ufrgs.br/textecc/porlexbras/porpopular/caixaferramentas.php#dadosCorpus

Maurice Gross no "Laboratório de Automação Documental e Linguística" (LADL[5]). Tais trabalhos foram, posteriormente, estendidos a outras línguas através da rede de laboratórios RELEX[6].

Os dicionários DELAF, assim, descrevem as palavras simples e também as palavras compostas de uma dada língua associando cada forma tanto a um lema quanto a uma série de códigos gramaticais, semânticos e flexionais.

Os dicionários DELAF foram elaborados por equipes de linguistas para várias línguas (francês, inglês, grego, italiano, espanhol, alemão, tailandês, coreano, polonês, norueguês, português, árabe, entre outras) e são utilizados hoje em projetos acadêmicos e industriais. Vários, entre eles o DELAF do francês e o do português do Brasil, são de acesso livre e atualizados colaborativamente.

Funcionamento dos dicionários no sistema UNITEX

O UNITEX é um analisador de *corpus* de acesso livre que permite processar textos em línguas naturais utilizando recursos linguísticos. Esses recursos apresentam-se na forma de dicionários computacionais no formato DELAF, de gramáticas e de tabelas de léxico-gramática integradas ao sistema. Alguns recursos linguísticos são distribuídos junto com o UNITEX e outros podem ser desenvolvidos por usuários. O sistema UNITEX, como um todo, é gratuitamente acessível e segue sendo utilizado para dar suporte a estudos sobre diferentes idiomas (PAJIĆ *et al.*, 2018), incluindo línguas antigas (KINDT, 2018).

Conforme mostrou a síntese de Almeida e Ferreira (2007), o sistema UNITEX permite o processamento de qualquer conjunto de textos – para que sejam localizadas e categorizadas as expressões linguísticas que o integram. Essa localização ou identificação de expressões de um dado texto ou *corpus* funcionará "desde que tais expressões integrem o dicionário a ele acoplado" ou que o usuário as descreva numa fórmula de busca. O UNITEX permite que se identifiquem palavras por classes.

Uma vez selecionado um texto determinado, como os textos dos nossos jornais populares, o UNITEX propõe-se a pré-processá-lo. Esse pré-processamento consiste em aplicar ao texto as seguintes operações: normalização de separadores, segmentação em unidades lexicais, normalização de formas não ambíguas, segmentação em frases e, por fim, a aplicação dos dicionários computacionais presentes no computador. A presença desses dicionários constitui um diferencial do sistema UNITEX em relação a outras ferramentas usuais de busca por padrões de palavras em *corpora*, pois pode-se localizar amplas classes de palavras com padrões simples.

Quando se processa um dado *corpus* ou texto em uma dada língua, o funcionamento interno do UNITEX consiste na construção de um subconjunto dos dicionários que

[5] Para maiores informações sobre o LADL, ver: http://infolingu.univ-mlv.fr/LADL/Historique.html

[6] Cf. http://unitexgramlab.org/pt/relex-network

contém somente as formas presentes no *corpus* sob exame. Assim, por exemplo, a aplicação do dicionário DELAF sobre um texto como *O time de Neymar corria atrás do prejuízo* produzirá o seguinte subconjunto do dicionário de palavras simples:

atrás,.ADV
corria,correr.V:I1s
corria,correr.V:I3s
de,.PREP
do,.PREPXDET+Art+Def:ms
do,.PREPXPRO+Dem:ms
o,.DET+Art+Def:ms
o,.N:ms
o,.PRO+Dem:ms
o,ele.PRO+Pes:A3ms
prejuízo,.N:ms
time,.N:ms

O nome *Neymar*, sendo descrito em um dicionário de nomes próprios distinto do DELAF do português, será considerada uma '*palavra desconhecida'*.

A aplicação dos dicionários ao texto é realizada pelo sistema UNITEX com o seu programa denominado DICO e gera subconjuntos – "subdicionários" – assim denominados: *palavras simples; palavras compostas; palavras desconhecidas*. Neste trabalho, ocupamo-nos apenas do último grupo.

O sistema UNITEX tem uma segunda funcionalidade: permite que o próprio usuário insira em seus dicionários computacionais novas palavras e informações gramaticais, assim adaptando esses recursos para diferentes finalidades. Os dicionários computacionais do UNITEX utilizam o formalismo dos DELA (Dicionários Eletrônicos do LADL). Esse formalismo permite descrever as entradas lexicais simples e compostas de uma língua associando-lhes, de modo opcional, informações gramaticais, semânticas e flexionais. Distinguem-se, dentro desse formalismo, dois tipos de dicionários eletrônicos. O utilizado com maior frequência é o dicionário de formas flexionadas, no formato DELAF (DELA de formas Flexionadas). O segundo tipo é o dicionário de lemas, nos formatos DELAS (DELA de formas Simples) ou DELAC (DELA de formas Compostas), que gera os demais dicionários. Neste trabalho, nos ocuparemos apenas do DELAF. Uma entrada do DELAF PB tem a seguinte forma:

sambou, sambar.V:J3s

Nela, "sambou" é a forma encontrada no texto, "sambar" é o lema, "V" é a classe gramatical – no caso, um verbo – e "J3s" é o código flexional – no caso, uma forma da terceira pessoa do singular do pretérito perfeito. A lista completa dos códigos gramaticais e flexionais para o português do Brasil pode ser encontrada em Muniz (2004).

O sistema UNITEX, para o português do Brasil, traz integrados os seguintes dicionários computacionais, em duas versões:

- na versão 2005: a partir de um DELAS de 61.335 palavras, gerava-se um DELAF de 878.095 palavras flexionadas simples e 4.100 palavras compostas. Esses recursos foram criados a partir do léxico do ReGra – base do corretor ortográfico do sistema *Word for Windows* - para os substantivos, adjetivos e advérbios (MARTINS *et al.*, 1998) e dos 102 modelos de flexão verbal de Vale (1990).
- na versão 2015: foram incorporadas as formas da nova ortografia resultante do Acordo Ortográfico de 1990. Foram introduzidas 7.900 novas entradas de formas simples (entre substantivos, adjetivos e advérbios) além das formas verbais enclíticas e mesoclíticas que não se encontravam na primeira versão. Com isso, o DELAF 2015 do português do Brasil conta hoje com 10.954.724 entradas, das quais 7.632.498 são formas distintas umas das outras.

Etapas do trabalho e resultados

O experimento de verificação, em linhas gerais, envolveu:

a) gerar a lista de palavras diferentes (*types*) empregadas no jornal DG – sem proceder a qualquer alteração da grafia;
b) gerar a lista de palavras do jornal MA;
c) comparar essas listas com as listas de entradas das versões 2004 e 2015 do dicionário computacional DELAF PB;
d) avaliar a cobertura das palavras - em termos de *tokens* (total de ocorrências) e de *types* (número de palavras distintas) - de cada amostra pelo DELAF PB, em cada versão do dicionário;
e) propor modos de inclusão de itens não identificados pelos dicionários.

Nessas etapas, a verificação gerou duas listas de palavras desconhecidas pelo DELAF-PB de cada jornal (lista DG_{04}, lista DG_{15}, lista MA_{04} e lista MA_{15}), com e sem a vigência da nova ortografia. Em seguida, a diferença entre maiúsculas e minúsculas foi ignorada.

Ao iniciar o processo de comparação, uma das primeiras constatações foi a de que o DELAF PB 2015 não continha a lista de abreviaturas e siglas – etiquetadas como ABREV (abreviaturas) e SIGL (siglas) – que constava na versão de 2004 do dicionário. Com efeito, durante a revisão efetuada por Calcia *et al.* (2014), as formas das abreviaturas e siglas (como *ABS* ou *ABNT* no jornal DG, *UFBA* ou *UFC* no jornal MA), que não eram objeto de trabalho na atualização do dicionário, tinham sido separadas num dicionário distinto. Para padronizar as condições experimentais, as listas DG_{15} e MA_{15} foram geradas novamente, utilizando o DELAF PB 2015 junto com o dicionário de abreviações e siglas. No que segue, as estatísticas referentes ao DELAF PB 2015 são o resultado desta segunda geração.

Os **Quadros 1** e **2**, a seguir, reproduzem partes de cada uma dessas listas. Destacamos os 30 primeiros itens desconhecidos pelos dicionários do UNITEX na letra **U** e, em seguida, os 30 primeiros itens iniciados pela letra **A**:

Quadro 1 – Amostra da lista de itens DG e MA desconhecidos no DELAF PB, iniciados pela letra U.

Itens iniciados por U	DG_{04} DELAF 2004	DG_{15} DELAF 2015	MA_{04} DELAF 2004	MA_{15} DELAF 2015
1.	uai	uai	ualex	ualex
2.	uau	uau	uanderson	uanderson
3.	ubial	ubial	ubandista	ubandista
4.	ubs	ubs	ubang	ubang
5.	udesca	udesca	ubatã	ubatã
6.	udi	udi	ubiracê	ubiracê
7.	údiche	údiche	ucla	ucla
8.	udine	udine	uefs	uefs
9.	udinese	udinese	uellinton	uellinton
10.	uebel	uebel	uelliton	uelliton
11.	uefa	uefa	uenf	uenf
12.	ufa	ufa	ueslei	ueslei
13.	ufcspa	ufcspa	uezo	uezo
14.	uflacker	uflacker	ufc	ufc
15.	ufsm	ufsm	uff	uff
16.	ugapoci	ugapoci	ufrrj	ufrrj
17.	ughini	ughini	uhu	uhu
18.	ugowski	ugowski	uibaí	uibaí
19.	uilson	uilson	ulício	ulício
20.	ulalá	ulalá	umidificador	unasul
21.	ulbra	ulbra	unasul	under
22.	uli	uli	under	undime
23.	ulmen	ulmen	undime	uneb
24.	ulsan	ulsan	uneb	unifacs
25.	ultramen	ultramen	unifacs	unifcas
26.	ultrasom	ultrasom	unifcas	unirio
27.	ultrassonografias	umbom	unirio	unit
28.	umbom	umchorão	unit	united
29.	umchorão	umespa	united	universitario
30.	umespa	unasul	universitario	uol

Fonte: Elaboração própria.

Quadro 2 – Amostra da lista de itens DG e MA desconhecidos no DELAF PB, iniciados por A.

Itens iniciados por A	DG_{04} DELAF 2004	DG_{15} DELAF 2015	MA_{04} DELAF 2004	MA_{15} DELAF 2015
1.	aabb	aabb	abadá	abadá
2.	aaliyah	aaliyah	abadábraço	abadábraço
3.	aas	aas	abadás	abadás
4.	abachilov	abachilov	abaralhau	abaralhau
5.	abadía	abadía	abdelmassih	abdelmassih
6.	abandon	abandon	abdulá	abdulá
7.	abatê	abbey	abefin	abefin
8.	abbey	abbott	aberbach	aberbach
9.	abbott	abdel	abisson	abisson
10.	abdel	abdômem	abla	abla
11.	abdômem	abdul	abordá	aboubacar
12.	abdominoplastia	abdulla	aboubacar	abravanel
13.	abdul	abebe	abravanel	academiagf
14.	abdulla	abech	academiagf	accosta
15.	abebe	abelão	accosta	acessando
16.	abech	abelhocídio	acessando	acessar
17.	abelão	abenício	acessar	acesse
18.	abelhocídio	ablo	acesse	acm
19.	abenício	about	acm	adab
20.	ablo	abp	adab	adailson
21.	abordá	abração	adailson	adailton
22.	aborígenes	abraciclo	adailton	adan
23.	about	abramet	adan	adanascimento
24.	abp	abramovich	adanascimento	adecir
25.	abraçá	abrh	adecir	adelmário
26.	abração	abrhrs	adelmário	adelmo
27.	abraciclo	abrigagem	adelmo	ademi
28.	abramet	abrilina	ademi	ademilson
29.	abramovich	abrito	ademilson	adenilton
30.	abrh	abs	adenilton	aderam

Fonte: Elaboração própria.

Após a geração das listas de palavras desconhecidas pelos dicionários DELAF 2004 e 2015, desprezando-se a diferença maiúsculas/minúsculas, as listas de itens desconhecidos foram comparadas entre si. Os itens desconhecidos de cada lista foram estudados conforme exame de seu uso nos textos e levando em conta as informações registradas em dois dicionários convencionais do português do Brasil, o dicionário Aurélio (FERREIRA, 1999) e o Dicionário Houaiss (HOUAISS; VILLAR, 2009), e foram tentativamente agrupadas em categorias tais como:

(1) **Erro de digitação** (umchorão, ubandista);

(2) **Grafia antiga** (idéia);

(3) **Nomes próprios** (uilson, uanderson);

(4) **Abreviações/siglas** (abs, ufrrj);

(5) **Expressões diversas/gírias/estrangeirismos** (ulalá, university, united);

(6) **Outros substantivos** (umidificador);

(7) **Outros** (abadábraço, aboubacar).

Essas categorias, naturalmente, foram uma referência provisória e inicial para um enquadramento dos itens desconhecidos que se apresentavam nas listas e podem ser refinadas em trabalhos futuros. Há casos de palavras desconhecidas que podem ser classificadas como neologismos ou regionalismos, por exemplo. Há também aquelas que também são, ao mesmo tempo, um substantivo e um neologismo (cf. *abelhocídio*). Praticamente não foram encontrados adjetivos ou verbos como itens desconhecidos, de modo que as categorias 'verbo' e 'adjetivo', nessa aproximação inicial, não foram propostas. Uma categorização multifatorial dos itens desconhecidos renderia, por si, um outro trabalho.

Resultados – visão geral e resumida

Sintetizamos, a seguir, no Quadro 3, os principais resultados obtidos nas duas amostras de jornais populares examinados. Esses resultados serão comentados na próxima seção.

Quadro 3 – Resultados obtidos no DG e no MA.

Jornal	*Diário Gaúcho* (DG) – amostra de textos variados	*MASSA!* (MA) – amostra de notícias
Ortografia	antiga	atual
Formas distintas (***types***)	53.966	22.414
Formas simples (***tokens***)	984.465	215.776
Com DELAF 2004:		
formas desconhecidas	10.512	3.048
% do total de ***types***	19,48%	13,60%
ocorrências desconhecidas	36.190	11.624
% do total de ***tokens***	3,68%	5,39%
Com DELAF 2015:		
formas desconhecidas	9.967	2.769
% do total de ***types***	18,47%	12,35%
ocorrências desconhecidas	34.611	10.870
% do total de ***tokens***	3,52%	5,04%

Fonte: Elaboração própria.

Considerações sobre os resultados obtidos – identificação do universo vocabular

O DELAF PB possui uma ampla cobertura lexical de jornais do século XX e de textos literários do século XIX (1,9% de *types* desconhecidos em *A senhora* de José de Alencar). Em comparação, as percentagens de *types* desconhecidos no quadro 3 (de 12% até 19%) são sensivelmente maiores. Portanto, o universo vocabular do tipo de jornal em foco pode ser visto, sim, como um empecilho importante para a identificação de palavras pelos dicionários DELAF-PB. Esse é um aspecto relevante para os pesquisadores do PB interessados na sua utilização futura.

Para que se possa contextualizar os resultados sintetizados na seção anterior, importa lembrar os fatores envolvidos no desempenho da identificação de itens do vocabulário dos jornais populares pelos dicionários DELAF PB:

a) os itens do DG têm palavras na ortografia antiga – antes do Acordo;

b) os itens do MA têm palavras na ortografia atual;

c) apenas o DELAF PB 2015 está preparado para a nova ortografia;

d) o DELAF PB 2004 não inclui a nova ortografia.

A lista de palavras distintas (*types*) empregadas no DG, em um universo de 53.966 itens diferentes, inclui 19,48% de itens desconhecidos pelo DELAF 2004 e 18,47% de itens não contemplados pelo DELAF 2015. Assim, observa-se uma pequena redução nesse percentual de **palavras desconhecidas** entre as duas versões do dicionário.

Salienta-se que há, nos itens do DG, um padrão antigo de ortografia. No entanto, o fato de que as grafias antigas, como "*agüentar*", não constam mais no DELAF 2015 parece ter repercutido sobre o desempenho menos do que a inserção de palavras novas, como "*umidificador*", e de formas verbais enclíticas, como "*abordá*" em "*abordá-lo*". Assim, o percentual de palavras desconhecidas diminuiu.

No jornal MA, tivemos 22.414 formas diferentes, sendo 13,60% dos itens desconhecidos no DELAF 2004 contra 12,35% no DELAF 2015. Neste caso, a ortografia está apenas no formato atualizado. Novamente, a cobertura do texto pelo DELAF melhorou de 2004 para 2015. Desta vez, o efeito da adaptação do dicionário ao acordo ortográfico se adicionou ao efeito da inserção de palavras novas e formas enclíticas.

Conforme verificamos, tanto no *corpus* MA quanto no *corpus* DG, no dicionário DELAF de 2004 para o DELAF 2015, o desempenho melhorou ligeiramente no que se refere ao reconhecimento de itens oriundos do jornal popular. O reconhecimento de itens, de 2004 para 2015, ampliou-se, em média, em 1,13 ponto percentual em termos de *types* (palavras diferentes) reconhecidos.

Em termos de números de ocorrências (*tokens*), o vocabulário do DG mostra-se sensivelmente mais coberto (em média 96,4%) do que o vocabulário do jornal MA (em média 94,8%), com pouca repercussão da atualização do dicionário incidindo nesse processo. As razões para isso podem ser, naturalmente, de várias ordens. Entretanto, vale ponderar que o jornal MA é da região Nordeste do Brasil e o DG, da região Sul, o que pode repercutir sobre o perfil vocabular observado. Em termos de palavras distintas (*types*), a diferença de cobertura parece ser inversa, mas esta comparação não é significativa, por causa da diferença de tamanho das amostras: numa amostra maior, como a do jornal MA, estatísticas sobre o número de *types* levam a uma super-representação das palavras pouco repetidas, que têm menos chance de constar no dicionário.

De qualquer modo, o jornal popular mostra-se como uma fonte de pesquisa interessante e desafiadora. Entretanto, em diferentes *corpora* atuais do português do Brasil, encontram-se ainda, via de regra, apenas dados coletados de grandes jornais tradicionais, tais como, por exemplo, a *Folha de S. Paulo* (cf. ALUISIO; ALMEIDA, 2006).

Como vimos, para os dicionários DELAF PB, os jornais populares representam uma fonte, aparentemente, um tanto "estranha". Assim, fica como ideia para um trabalho futuro contrastar esse percentual de palavras desconhecidas entre jornais populares e jornais tradicionais produzidos num mesmo período.

Outro fator interessante relacionado ao vocabulário desconhecido é a sua vinculação com uma temática potencialmente regionalista ou local (como o item '*cacetinho*' no DG, equivalente sulista de "pão francês") e com nomes próprios de criação recente (como o item '*Abadábraço*'do MA).

Na próxima seção, apresentamos, em linhas gerais, um perfil das palavras mais desconhecidas entre as duas versões dos dicionários, em diferentes categorias de palavras, a partir da amostra de itens aqui trazida. Em seguida, ponderamos como alguns itens desconhecidos poderiam ser incorporados ao dicionário DELAF PB 2015.

Perfil das palavras desconhecidas e opções para enriquecimento do DELAF PB 2015

Conforme já foi notado, o exame das listas de palavras dos nossos quadros demonstra que a versão de 2015 do DELAF PB obteve um modesto ganho de cobertura nesse tipo específico de jornal. Assim, a listagem das 30 primeiras palavras iniciadas por A é praticamente idêntica nas duas colunas que se referem ao *corpus* MA. Pode-se notar que existe algum vocabulário regional (*abadá*, *abaralhau* - MA) e uma série de nomes próprios. Embora o DELAF PB contenha um bom número de nomes próprios (*Aldemário*, *Abramovich*) que normalmente não constariam de um dicionário convencional, há ainda um bom trabalho a ser feito no sentido de listar e caracterizar esses nomes próprios, sobretudo por conta de sua grande variação no Brasil (cf. *Uanderson*, *Uellinton*, *Ueslei*).

Um exame mais apurado das listas demonstra também que a maioria dos itens não reconhecidos é de substantivos (no DG: *abrigagem, abdômem, abelhocídio*), incluídos aí os nomes próprios (*Abelão*). Os dois únicos verbos identificados nas listas aqui apresentadas foram o verbo '*acessar*', com diversas formas, e a forma '*abordá*' do verbo '*abordar*', uma das formas enclíticas não identificadas na versão de 2004 do DELAF PB (*abordá+lo*).

Outra questão diz respeito às abreviações e siglas presentes nos textos. Note-se que o domínio da construção de dicionários computacionais demanda um esforço especial, integrando linguistas e cientistas da Computação. A construção de recursos computacionais mais abrangentes, que incluam processos e fenômenos recorrentes do léxico do português atual é um desafio bastante complexo. Nessa direção, as abreviaturas/siglas/acrônimos e os nomes próprios constituem fenômenos recorrentes na língua escrita e já demandam trabalhos específicos para preservar e melhorar o funcionamento desses tipos de sistemas computacionais. Em trabalhos como Vale *et al.* (2008), já se apontou essa necessidade para o tratamento para *corpora* históricos do português, o que pode ser estendido aos acervos da atualidade. De fato, a construção de dicionários computacionais específicos de abreviaturas, de siglas e de entidades nomeadas poderia ser um caminho adequado para tratar do problema que colocamos para o sistema UNITEX com os nossos jornais populares. Naturalmente, conforme podemos verificar nos seguintes trechos de duas notícias do nosso *corpus*, há muito mais a ser explorado:

TRECHO 1:
"A primeira delas é o lançamento do Abadábraço, um bloco que desfilará sem cordas, mas com os foliões. Desfilará sem cordas, mas com os foliões devidamente trajados com abadás. A proposta aqui é incluir. É ter mais pessoas brincando nas ruas e com direito a usar o seu abadá."

TRECHO 2:
"Chaleira, César Oliveira & Rogério Melo, Bochincho, Os Quatro Gaudérios, Portal Gaúcho e Eco do Minuano & Bonitinho. Foi grande a integração entre as invernadas adulta e xiru na Sociedade Gaúcha de Lomba Grande, em Novo Hamburgo, que comemorou 70 anos na noite de terça-feira.

Conclusão

O problema de pesquisa enfrentado neste trabalho foi, assim, a descrição e a ponderação sobre o desempenho de um dicionário computacional de ampla cobertura, gratuitamente acessível para ser utilizado na pesquisa sobre o português do Brasil e na indústria de processamento computacional de línguas.

O dicionário DELAF sob exame, ainda que extremamente importante para subsidiar diferentes tarefas linguísticas, poderia ser incrementado com a alimentação de *corpora* de jornais populares brasileiros. Afinal, conforme já mencionado, esse gênero jornalístico tem sido ainda pouco contemplado como fonte de dados para o estudo do português escrito culto. Tal "estranhamento" lexical, em termos de números de ocorrências (*tokens*) desconhecidas, mostra-se sensivelmente menor no vocabulário do jornal gaúcho (em média 96,4%) do que do baiano (em média 94,8%), fato que enseja um aprofundamento da pesquisa. Portanto, ao examinar o desempenho de diferentes versões do dicionário, frente ao tratamento do léxico de jornais populares brasileiros, demonstramos a validade de tomar ambos como objetos e como fontes de dados para pesquisas sobre a linguagem em que cooperam linguistas e cientistas da Computação.

AGRADECIMENTOS

Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento do Pessoal de Nível Superior – Brasil (CAPES) no âmbito do Programa CAPES-STIC-AMSud (proj.047/14), pelo CNPq – Conselho Nacional de Desenvolvimento Científico (bolsa PQ – proc. 305625/2016-0 e Auxílio APV proc. 453058/2015-9), e pela Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP (proc. 2016/24670-3).

FINATTO, M.; VALE, O.; LAPORTE, E. Recognition of the vocabulary of popular Brazilian newspapers with a freely available computational dictionary. **Alfa**, São Paulo, v.63, n.1, p.63-80, 2019.

- *ABSTRACT: We report an experiment of checking the identification of a set of words in popular Portuguese written text with two versions of a computational dictionary of Brazilian Portuguese, DELAF PB 2004 and DELAF PB 2015. This computational dictionary is freely available for use in linguistic analyses of Brazilian Portuguese and other research, which gives reasons for undertaking a critical study. The set of words comes from the PorPopular corpus, composed of popular newspapers, the Diário Gaúcho (DG) and the Bahian newspaper Massa! (MA). From DG, we studied a set of texts with 984,465 words (tokens), published in 2008, in the spelling used before the Orthographic Agreement of the Portuguese Language adopted in 2009. From MA, we examined a vocabulary of 215,776 words (tokens), from papers published in 2012, 2014 and 2015 in the new spelling. The verification involved: a) generating lists of unique words used in DG and MA; b) comparing these lists with the entry lists of the two versions of DELAF PB; c) assessing the coverage of this vocabulary; d) proposing ways of including the items not covered. The results showed that an average of 19% of the types in the DG corpus were unknown by the DELAF PB 2004 and 2015. In the MA sample, this average was 13%. The version of the dictionary impacted slightly on item recognition performance.*

- *KEYWORDS: Popular newspapers. Lexic. Vocabulary. Computational dictionary. Lexical coverage. Recognition of words. Brazilian Portuguese.*

REFERÊNCIAS

ALMEIDA, M. L. L.; FERREIRA, R. G. Ferramentas eletrônicas de busca e a pesquisa linguística: o caso do angulador “um tipo de”. **Estudos Linguísticos**, São Paulo, v. 35, n.1, p. 188-196, 2007.
Disponível em: http://www.gel.hospedagemdesites.ws/estudoslinguisticos/edicoes anteriores/4publica-estudos-2007/sistema06/20.PDF. Acesso em: 17 jan. 2018.

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um corpus? lições aprendidas na compilação de vários corpora para a pesquisa linguística. **Calidoscópio**, São Leopoldo, v.4, n. 3, p. 155-177, set./dez. 2006.

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

BIDERMAN, M. T. C. Conceito linguístico de palavra. **Revista Palavra**, Rio de Janeiro, n. 5, p. 81-97,1999.

BIDERMAN, M. T. C. Dimensões da palavra. **Filologia e Linguística Portuguesa**, Lisboa, n. 2, p. 81-118, 1998.
Disponível em: http://dlcv.fflch.usp.br/sites/dlcv.fflch.usp.br/files/Biderman1998_0.pdf. Acesso em: 17 jan. 2018.

CALCIA, N. P.; KUCINSKAS, A. B.; MUNIZ, M.; NUNES, M. G. V.; VALE, O. A. **Révision et adaptation des dictionnaires et graphes de flexion d'Unitex-PB à la nouvelle orthographe du portugais**. 2014. Trabalho apresentado no 3rd. UNITEX/ GRAMLAB WORKSHOP, Tours, 2014.

CHISHMAN, R. Convergências entre semântica de frames e lexicografia. **Linguagem em (Dis)curso – LemD**, Tubarão, v. 16, n. 3, p. 547-559, set./dez. 2016. Disponível em: http://www.scielo.br/pdf/ld/v16n3/1518-7632-ld-16-03-00547.pdf. Acesso em: 17 jan. 2018.

FINATTO, M. J. B. Projeto PorPopular, frequência de verbos em português e no jornal popular popular brasileiro. *In*: ISQUERDO, A. N.; SEABRA, M. C. T. da C. (org.). **As Ciências do Léxico:** lexicologia, lexicografia, terminologia. Campo Grande: Ed. da UFMS, 2012. v.6, p. 227-244.

FINATTO, M. J. B; PEREIRA, A. Frequências de verbos em *corpora* de jornais populares: dados para atividades ensino com os jornais "Diário Gaúcho" e o "The Sun". **Linguagem Estudos e Pesquisas,** Catalão, v.18, n. 2, p. 149-165, 2014.
Disponível em: https://www.revistas.ufg.br/lep/article/view/39730/21165. Acesso em: 30 jan. 2018.

FINATTO, M. J. B.; SCARTON, C.; ROCHA, A.; ALUÍSIO, S. M. Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 8., 2011, Cuiabá. **Anais do STIL 2011**. Cuiabá: Sociedade Brasileira de Computação, 2011. v. 1, p. 30-39.

FERREIRA, A. B. de H. **Dicionário Eletrônico Aurélio Século XXI**. Versão 3.0. Rio de Janeiro: Nova Fronteira: Lexikon Informática, 1999. 1 CD-ROM.

FLORES V.; FINATTO, M. J. B. Quantificação e argumento de autoridade no Acordo Ortográfico de 2009: aspectos enunciativos e estatísticos. *In*: MOREIRA, M. E.; SMITH, M. M.; BOCCHESE, J. da C. (org.). **Novo acordo ortográfico da língua portuguesa:** questões para além da escrita. Porto Alegre: EDIPUCRS, 2009. p. 109-136.

HOUAISS, A.; VILLAR, M. S. **Dicionário Houaiss de Língua Portuguesa**. Elaborado pelo Instituto Antônio Houaiss de Lexicografia e Banco de. Dados da Língua Portuguesa S/C Ltda. Rio de Janeiro: Objetiva, 2009.

KINDT, B. Processing tools for Greek and other languages of the Christian Middle East. **Journal of Data Mining and Digital Humanities**, Cedex, 2018. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages. Disponível em: https://jdmdh.episciences.org/4184. Acesso em: 27 jun. 2018.

LAPORTE, E. Dictionaries for language processing: readability and organization of information. *In*: LAPORTE, E.; SMARSARO, A.; VALE, O. A. (org.) **Dialogar é preciso:** linguística para o processamento de línguas. Vitória: PPGEL/UFES, 2013. p. 119-132.

MARTINS, R. T.; HASEGAWA, R.; NUNES, M. D. G. V.; MONTILHA, G.; OLIVEIRA, O. N. Linguistic issues in the development of REGRA: a grammar checker for Brazilian Portuguese. **Natural Language Engineering**, Cambridge, v. 4, n. 4, p.287-307, 1998.

MOREIRA, M. E.; SMITH, M. M.; BOCCHESE, J. da C. (org.). **Novo acordo ortográfico da língua portuguesa:** questões para além da escrita. Porto Alegre: EDIPUCRS, 2009.

MUNIZ, M. C. M. **A construção de recursos lingüístico-computacionais para o português do Brasil**: o projeto de Unitex-PB. 2004. 92f. Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) - Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, 2004.
Disponível em: http://ladl.univ-mlv.fr/brasil/bibliografia/oto/DissMuniz2004.pdf. Acesso em: 22 jan. 2018.

OLIVEIRA, M. R. 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". **Veredas:** Revista de Estudos Linguísticos**,** Juiz de Fora, v. 13, n. 2, p. 07-19, 2009.

PAIVA, F. A.; BARBOSA, P.; FARIA, P.; MARTINO, J. M. de. Towards machine translation from Brazilian Portuguese to libras: a corpus-based, morphosyntactic analysis *In*: EVERS, A.; NARDES, A.; BRANGEL, L.; FINATTO, M. J. B.; CHISHMAN, R. L. (org.). **Caderno de Resumos do ELC-EBRALC 2017**. São Leopoldo: UNISINOS, 2017. p.24-28.
Disponível em: http://www.ufrgs.br/elc-ebralc2017/caderno-de-resumos/Cadernode Resumos_ELC2017_16out17.pdf. Acesso em: 22 jan. 2018.

PAJIĆ, V.; VUJIČIĆ STANKOVIĆ, S.; STANKOVIĆ, R.; PAJIĆ, M. Semi-automatic extraction of multiword terms from domain-specific corpora. **The Electronic Library**, Oxford, v.36, n.3, p.550-567, 2018. DOI: https://doi.org/10.1108/EL-06-2017-0128.

PAUMIER, S. **Unitex 3.1:** user Manual. Paris: Université Paris-Est Marne-la-Vallée, 2016.

SELIGMAN, L. Jornais populares de qualidade - ética e sensacionalismo em um novo fenômeno no mercado de jornalismo impresso. *In*: ENCONTRO NACIONAL DE PESQUISADORES EM JORNALISMO - SBPJor, 6., 2008, São Bernardo do Campo. **SBPJOR - a construção do campo do jornalismo no Brasil**. São Bernardo do Campo: Sociedade Brasileira de Pesquisa em Jornalismo, 2008.
Disponível em: https://bjr.sbpjor.org.br/bjr/article/viewFile/199/198. Acesso em: 23 jan. 2018.

SOARES, L. A. Análise do jornal popular super notícia sob enfoque crítico e multimodal. **Alfa,** São Paulo, v. 61, n.3, p. 575-597, 2017.

TRISTÃO, M. B.; MUSSE, C. F. O direito à informação e o (ainda restrito) espaço cidadão no Jornalismo Popular impresso. **Intercom:** Rev. Bras. Ciênc. Comun., São Paulo, v. 36, n. 1, p. 39-59, 2013. DOI: http://dx.doi.org/10.1590/S1809-58442013000100003.

VALE, O. A. **Dictionnaire électronique des conjugaisons des verbes du portugais du Brésil.** Paris: Université Paris 7, 1990. (Rapport Technique du LADL, n.27).

VALE, O. A.; BAPTISTA, J. Novo dicionário de formas flexionadas do Unitex-PB: avaliação da flexão verbal. *In*: STIL 2015; BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 10., 2015, NATAL. **Proceedings of the Conference**. Natal: Pontifícia Universidade Católica do Rio de Janeiro, 2015. v. 1, p. 171-180.

VALE, O. A.; CANDIDO JUNIOR, A.; MUNIZ, M.; BENGTSON, C.; CUCATTO, L.; ALMEIDA, G. M. B.; BATISTA, A.; PARREIRA, M. C.; BIDERMAN, M. T. C ; ALUÍSIO, S. M. Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION; WORKSHOP LANGUAGE TECHNOLOGY FOR CULTURAL HERITAGE DATA, 2008, Marrakech. **Proceedings** [...]. Marrakech: [*s. n.*], 2008. p. 47-54.

VICENTINI, M. Construção de *corpora* de mensagens eletrônicas para conversão automática em fala. **Língua, literatura e ensino**, Campinas, v. 5, p. 229-237, out. 2010. Disponível em: http://revistas.iel.unicamp.br/index.php/lle/article/view/1151. Acesso em: 28 jan. 2018.

ZAVAGLIA, C. Extração de informações de definições de um dicionário convencional para a elaboração de uma base de conhecimento léxica: estratégias e procedimentos linguísticos. *In*: LONGO, B. N. de O.; DIAS DA SILVA, B. C. (Org.) **A construção de dicionários e de bases de conhecimento lexical.** Araraquara: Laboratório Editorial FCL/UNESP; São Paulo: Cultura Acadêmica Editora, 2006. p. 209-234.

ZILIO, L. **Verblexpor**: um recurso léxico com anotação de papéis semânticos para o português. 196 f. 2015. Tese (Doutorado em Letras) – Universidade Federal do Rio Grande do Sul, 2015.

Recebido em 23 de março de 2018

Aprovado em 17 de junho de 2018