Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

A Importância dos Falsos Homógrafos para a Correção Automática de Erros Ortográficos em Português

Magali Sanches Duran, Lucas Vinícius Avanço, Maria das Graças Volpe Nunes

Núcleo Interinstitucional de Linguística Computacional (NILC) - Instituto de Ciências Matemáticas e Computação da USP – São Carlos - SP, Brasil

magali.duran@uol.com.br, avanco89@gmail.com, gracan@icmc.usp.br

Abstract. *This paper reports the analysis of 25.722 pairs of Portuguese words that differ from each other by a single diacritic, called "false homographs". Such words are relevant for spelling correction, as in these cases a misspelled word missing a diacritic is identical to a correct word, consequently preventing the identification and the correction of the misspelling. The purpose of the analysis is to identify and to exclude, from the lexicon used by a Portuguese speller, non-accented words that are relatively less frequent than their respective accented pairs. This action is specially justified when one aims to correct User-Generated Content (UGC), a kind of text characterized by missing diacritics, among other features. The result is a list of 2.052 words that fit the requirements of the aimed strategy.*

Resumo. *Este artigo relata a análise de 25.722 pares de palavras em português que só diferem por um acento. Essas palavras são denominadas aqui de "falsos homógrafos" e são relevantes para a correção de erros ortográficos, pois nesses casos uma palavra incorreta à qual falta um acento é idêntica a uma forma correta na língua, o que impede a identificação do erro e sua consequente correção. O propósito da análise é identificar pares em que a forma não acentuada tenha baixa frequência e a forma acentuada tenha alta frequência, e assim excluir, do léxico que servirá de base para o corretor ortográfico, as formas pouco frequentes. Essa proposta justifica-se especialmente quando se almeja a correção ortográfica de Conteúdo Gerado por Usuários na web (CGU), um tipo de texto caracterizado, entre outras coisas, pela falta de acentos. O resultado é uma lista de 2.052 palavras que atendem às condições da estratégia pretendida.*

1. Introdução

Os textos digitais produzidos por internautas com a finalidade de partilhar informações e opiniões apresentam uma série de características que os desviam da norma culta. Esses textos têm sido discutidos na literatura como "conteúdo gerado por usuário" (CGU). É muito comum observarmos no CGU a reprodução da pronúncia na escrita (kaza=casa, noiz=nós, vamu=vamos), bem como a total ausência de acentos e cedilhas.

No contexto em que se inserem, esses textos satisfazem sua função comunicativa. Ocorre, porém, que esses mesmos textos são uma rica fonte de informações para a sociedade em geral (empresas, governos e consumidores) e, para analisá-los em grandes lotes, é preciso utilizar o processamento automático de línguas naturais (PLN), cujas

ferramentas têm sido construídas, em geral, para processar a língua padrão, o que implica que podem não reproduzir seu melhor desempenho ao processar CGU. Uma alternativa tem sido pré-processar tais textos, normalizando-os à luz da língua padrão, antes de serem tratados pelos sistemas de PLN.

No cenário de "normalização" de CGU, os corretores ortográficos desempenham um papel de destaque. Os corretores tradicionais não estão preparados para tratar as especificidades do CGU, como reprodução da oralidade na escrita (vamu=vamos), os erros foneticamente motivados (chadres=xadrez; difíciu=difícil), as abreviações de palavras (pq=porque; q=que; nd=nada, vc=você), as repetições de letras para produzir ênfase (boooooooooooom=bom), entre outros. Uma das tarefas de um corretor ortográfico na normalização de CGU em português é colocar os acentos, que são frequentemente suprimidos nesses textos. Na maior parte dos casos, essa tarefa é simples, pois detectado um erro - por exemplo "coracao", o corretor busca uma forma que tenha as mesmas letras e contenha sinais diacríticos, produzindo a correção "coração". No entanto, ao analisar o resultado de correção automática em um corpus de CGU com o corretor Aspell[1], verificou-se que muitas palavras que deveriam ser corrigidas não o foram, como por exemplo, "obvio=óbvio". A causa disso é que existe no léxico a forma "obvio" (primeira pessoa do indicativo do verbo "obviar"). Casos como esse frustram a expectativa de quem utiliza os corretores ortográficos.

Percebeu-se, contudo, que esse tipo de problema poderia ser parcialmente superado por meio de uma adaptação do léxico utilizado pelo corretor ortográfico. No exemplo, como o verbo "obviar" é pouco frequente, se as formas "obvio", "obvia" e "obvias" fossem suprimidas do léxico do corretor ortográfico, seus falsos homógrafos "óbvio", "óbvia" e "óbvias", altamente frequentes, poderiam ser devidamente corrigidos sempre que fossem escritos sem acento. Em português, os acentos diacríticos são responsáveis por distinguir cerca de 25.000 itens lexicais e o desafio enfrentado pelo trabalho descrito neste artigo foi encontrar, nesse conjunto, pares de falsos homógrafos similares a "obvio-óbvio", em que a forma acentuada tem baixa frequência e a forma não acentuada tem frequência relativamente mais alta.

O restante deste artigo está organizado em quatro seções. Na Seção 2 fazemos uma breve revisão bibliográfica sobre o papel do léxico nos corretores gramaticais. Na Seção 3 descrevemos os materiais e métodos que utilizamos para adequar o léxico utilizado para corrigir CGU em português. Na Seção 4 discutimos os resultados e na Seção 5 tecemos nossas considerações finais e delineamos trabalhos futuros.

2. Corretores ortográficos e léxico

Os corretores ortográficos utilizam um léxico para duas tarefas: julgar se o insumo é erro ou não e, caso seja, escolher as palavras mais similares que podem ser oferecidas como candidatas à correção do erro. Para a primeira tarefa, dada uma entrada, o corretor procura-a no léxico e, caso não a encontre, aponta-a automaticamente como erro. É importante que o léxico contenha muitas palavras da língua, inclusive neologismos e estrangeirismos, pois caso contrário o corretor aponta como erro palavras que não são erros.

[1] http://aspell.net/

Um corretor ortográfico pode utilizar um léxico formal da língua ou extrair um léxico de corpus. Martins & Silva (2004) alertam, contudo, que léxicos extraídos de corpora, inclusive de corpora de língua padrão, podem conter erros ortográficos, o que pode impedir a detecção de erros. A detecção com base na comparação com o léxico pode falhar em duas situações: 1) a palavra existe na língua, mas não consta do léxico, como "backup" (estrangeirismo); 2) a palavra está errada no contexto, mas como existe uma palavra no léxico igual à palavra errada, o erro não é detectado, como em: "Eu pedalo ate cansar", onde a palavra "até" não é corrigida porque existe a forma "ate" no léxico (primeira e terceira pessoas do singular do verbo "atar" no presente do subjuntivo). Esse problema é chamado de "erro de palavra real" – *real word error* (Choudhury et al., 2007).

Já a segunda tarefa, a de sugerir palavras para corrigir a palavra errada, enfrenta vários desafios. Quanto menor a palavra, maior o número de palavras similares. Além disso, nas palavras menores, uma letra errada ou fora do lugar representa um percentual grande do número total de letras e sobram menos "pistas" para descobrir qual é a provável palavra certa. Por exemplo, a palavra "coza" não existe no léxico e, considerando-se apenas três letras como pista, há várias palavras similares que poderiam ser sugeridas para corrigi-la: cota, cola, coma, cora, copa, só para citar algumas que têm uma letra (a terceira) de distância da palavra errada. Já em uma palavra maior, como "excelente", uma letra errada representa um nono avo do total de letras da palavra, sobrando oito letras como pistas para a palavra correta ("exelente", "ecelente" são erros ortográficos comuns). Por isso, é relativamente mais fácil corrigir palavras grandes do que palavras pequenas.

Os corretores ortográficos trabalham com o que Pelizzoni (2007) chama de "otimismo", ou seja, partem do pressuposto de que apenas uma letra esteja errada ou fora do lugar (ou até duas letras, para palavras maiores), pois caso contrário qualquer palavra poderia ser corrigida para qualquer palavra. Se em uma palavra de cinco letras a distância para a palavra correta fosse de duas letras, por exemplo, poderíamos ter a palavra errada "docem" onde o usuário pretendia escrever "jovem" ou "forem", o que tornaria a tarefa de correção muito difícil até para um humano.

Para gerar uma lista de palavras candidatas à correção de uma palavra errada, utiliza-se tradicionalmente a distância de edição de Levenshtein (1966), que calcula o número de operações (substituição, inserção ou deleção de caracteres) necessárias para transformar a palavra errada na palavra candidata à correção. Tem-se, assim, um conjunto de palavras que distam da palavra errada por um caractere, por dois caracteres e assim por diante. Quanto maior o número de palavras similares, mais difícil classificá-las em ordem de probabilidade para correção. Se há interação humana, ou seja, se o corretor é usado dentro de um editor de textos, o usuário pode escolher entre as várias palavras oferecidas como candidatas para a correção; porém, se não há interação humana (a correção é automática), os critérios para decidir qual é a melhor candidata precisam ser mais eficientes ainda. Entre os critérios mais utilizados estão a distância das letras no teclado (importante para erros de digitação) e a semelhança fonética entre a palavra errada e a palavra candidata a correção, como faz o Soundex (Russel, 1918). Com o objetivo de facilitar a classificação das palavras candidatas à correção no português, Avanço et. al. (2014) desenvolveram um corretor ortográfico que incorpora regras fonéticas es-

pecíficas para o português. Por exemplo, a palavra "caza" é mais provável de ser corrigida por "casa", que tem a mesma pronúncia, do que por "cada", "cala" ou "cama".

Neste trabalho, estendemos o que Pelizzoni chama de "otimismo", pois partimos do pressuposto de que o erro mais simples que pode ser cometido por um usuário da língua seja a supressão de acento, antes de considerarmos os erros de caracteres. E é com o objetivo de facilitar a correção de acentos que propomos uma estratégia de customização do léxico usado para correção ortográfica.

3. Materiais e Métodos

Nosso objetivo é encontrar pares de falsos homógrafos compostos por uma palavra não acentuada pouco frequente e uma palavra acentuada muito frequente. Para esses pares, nossa proposta é excluir do léxico a forma não acentuada, de maneira a beneficiar a correção ortográfica de palavras muito frequentes escritas sem acento. É claro que, excluindo uma palavra dessas do léxico, corremos o risco de não a reconhecer quando ela estiver correta e de "corrigi-la" indevidamente, colocando acento. Estamos diante de um problema de custo *versus* benefício: se a forma acentuada for muito mais frequente do que a não acentuada, os ganhos em desempenho de um corretor serão maiores do que as possíveis perdas.

Utilizamos o léxico UNITEX-PB[2] (Muniz et al. 2005) como base de nosso estudo. Esse léxico contém 880.000 formas flexionadas e suas respectivas categorias gramaticais. Primeiramente, selecionamos todos os pares que só se diferenciavam por um acento. Essa seleção nos trouxe 25.722 pares de palavras. A maioria dos pares é constituída de uma forma acentuada e uma não acentuada (ex: pais, país), mas há casos em que ambas as formas são acentuadas (apôs, após).

A análise dessa lista preliminar nos mostrou que o léxico continha pares de palavras pré e pós-reforma ortográfica (ideia, idéia; voo, vôo), muitos pares constituídos pelas formas flexionadas de um mesmo verbo (amara, amará; reclamara, reclamará) e muitos pares de palavras muito raras. A fim de poder filtrar esses casos, incluímos mais dados em nossa lista: a categoria gramatical, a condição pré ou pós-reforma e a frequência de cada uma das palavras que compõem os pares de falsos homógrafos. Para pesquisar a frequência das formas, utilizamos o Corpus Brasileiro[3], que contém um bilhão de palavras e compreende diversos gêneros textuais.

O primeiro fato que observamos na lista com novos dados é que 79% dos 25.722 pares não apresentam frequência para nenhuma das duas formas. Após uma análise superficial dos 20.245 pares com frequência igual a zero, percebemos que eles realmente são constituídos de palavras pouco usuais e decidimos eliminá-los da lista para análise, ficando com 5.477 pares que possuem ocorrência no corpus em pelo menos uma das duas palavras.

Desses 5.477 pares, 616 contêm formas que perderam o acento com a reforma ortográfica e 43 são formas acentuadas criadas com a reforma, oriundas da aglutinação dos prefixos, como é o caso da forma "corresponsabilizará", do verbo "corresponsabili-

[2] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html

[3] http://corpusbrasileiro.pucsp.br/cb/Inicial.html

zar". Excluímos da nossa análise essas 616 formas, porém não cogitamos excluí-las do léxico, pois até que a reforma entre em vigor (01/01/2016), ambas as formas são consideradas corretas. A exclusão dos pares de palavras pré e pós-reforma da análise deixou nossa lista de falsos homógrafos com 4.861 pares.

Nesses pares restantes, observamos que 1.187 formas não acentuadas tinham frequência igual a zero no corpus, enquanto suas respectivas formas acentuadas apresentavam alguma frequência. Selecionamos essas 1.187 palavras para nossa estratégia de eliminação do léxico para fins de correção ortográfica.

Após essa ação, sobraram 3.674 pares para nova análise. O próximo passou foi analisar os casos de falsos homógrafos constituídos por formas flexionadas de verbos. Para isso, recortamos da lista todos os pares de palavras em que ambas as palavras eram categorizadas como verbo (V). Obtivemos 2.266 pares que atendem a essa restrição. Desses, verificamos que 2.151 pares (95%) são de verbos no passado mais-que-perfeito e no futuro do indicativo (passara, passará; ficara, ficará). Os outros 115 pares são constituídos de casos diversos, na maioria formas de verbos da terceira conjugação (diminui, diminuí; diminuis, diminuís; traia, traía; traiam, traíam), verbos "ter" e "vir" e seus derivados (contém, contêm; convém, convêm) e formas de verbos diferentes (rele, do verbo "relar", e relê, do verbo "reler"; inventariamos, do verbo "inventariar" e inventaríamos, do verbo "inventar"; seriamos, do verbo "seriar" e seríamos, do verbo "ser").

A diferença de frequência entre as formas verbais raramente é grande o suficiente para justificar a exclusão da forma não acentuada para possibilitar a correção da forma acentuada. Além disso, como a maioria dos pares é constituída de dois tempos pouco usados em CGU (passado mais-que-perfeito e futuro), decidimos excluir os pares verbais de nossa análise em busca de candidatos a serem suprimidos do léxico. Excluindo os 2.266 pares de verbos de nossa lista, restaram 1.408 pares para continuarmos nossa análise.

PAR (1)	POS TAG	FREQ.	PAR (2)	POS TAG	FREQ.
e	CONJ	22.238.879	é	V	5.325.656
a	ART	23.819.564	à	PREP	2.852.456
esta	PRON	377.018	está	V	894.867
as	ART	4.531.407	às	PREP	831.588
ate	V	2.414	até	PREP	805.937
sô	S	290	só	ADV	486.438
numero	V	4.284	número	S	419.536
pais	S	98.402	país	S	390.264
apôs	V	224	após	PREP	385.625
analise	V	4.145	análise	S	374.833
alem	V	1.575	além	ADV	295.003
historia	V	4.234	história	S	293.417
media	V	8.389	média	ADJ	277.336
inicio	V	5.600	início	S	233.844
publico	V	1.585	público	ADJ	229.410

Tabela 1. Excerto de pares de falsos homógrafos em que ambas as formas são frequentes

Quando começamos a lidar com as frequências para decidir se uma forma não acentuada poderia ser suprimida do léxico sem prejuízo para o desempenho do corretor ortográfico, percebemos que algumas frequências do corpus contradiziam o senso comum, como mostrado na Tabela 1.

Por exemplo, a forma "publico" tem 1585 ocorrências, o que consideramos um valor alto. Para verificar os contextos em que essa forma ocorre, utilizamos a ferramenta de busca ACDC[4], da Linguateca, onde o Corpus Brasileiro está disponível para consulta. A forma "publico" retornou 1555 ocorrências e, analisando as 100 primeiras, encontramos apenas uma forma que corresponde ao verbo "publicar" na primeira pessoa do singular. As demais formas correspondiam ao adjetivo "público", porém sem o acento, em sintagmas como "concurso publico", "setor publico", "serviço publico" e, em menor proporção, como substantivo, em contextos como "publico feminino", "aberto ao publico" etc.

A frequência de várias outras formas nos surpreenderam, e fizemos a mesma averiguação no corpus, que revelou que a maior parte das ocorrências da forma não acentuada correspondia à forma acentuada com falta de acento. A forma "numero", por exemplo, apresenta 4.284 ocorrências. Nas 100 primeiras concordâncias não encontramos nenhum caso do verbo "numerar" na primeira pessoa do singular: todas correspondiam ao substantivo "número" com erro de acento. Essas constatações nos permitiram concluir que:

- mesmo corpora de língua padrão contêm erros ortográficos;
- quando uma forma acentuada é muito frequente, ela tende a apresentar um número de formas com erros ortográficos, sem acento, que são confundidas com as formas corretas não acentuadas dos falsos homógrafos, inflando a frequência destas últimas;
- não podemos confiar nas frequências baixas das formas não acentuadas dos pares de falsos homógrafos, pois elas podem consistir erros ortográficos.

Decidimos verificar um a um os pares de falsos homógrafos ainda em análise, procurando identificar essas inconsistências de frequências. Para facilitar nosso trabalho, criamos uma razão entre a frequência da forma acentuada e a frequência da forma não acentuada como mostrado na Tabela 2. Classificamos a lista de pares por ordem decrescente desse novo número produzido.

[4] http://www.linguateca.pt/ACDC/

PAR (1)	POS TAG	FREQ. (1)	PAR (2)	POS TAG	FREQ (2)	FREQ (1) / FREQ (2)
leiloes	V	2	leilões	S	5088	2544
apôs	V	224	após	PREP	385625	1722
bufe	V	1	bufê	S	1261	1261
frances	NOM	36	francês	ADJ	43010	1195
camará	S	18	câmara	S	20181	1121
fabulas	V	1	fábulas	S	1015	1015

Tabela 2. Excerto das palavras selecionadas por terem baixa frequência em relação à forma acentuada.

Em 453 pares, a forma não acentuada era mais frequente que a acentuada e a nossa estratégia não se aplicava, por isso eliminamos esses pares do foco de análise. Nos 969 pares restantes, após análise manual, mantivemos 104 intactos e selecionamos 865 formas não acentuadas para serem excluídas do léxico do corretor ortográfico. A maioria dos pares que foram preservados tinha alta frequência em ambas as formas ou frequência semelhante (e, é; esta, está; da, dá). Preservamos também os pares que continham formas em primeira pessoa do singular de verbos frequentes (ex: critico, crítico), pois embora no Corpus Brasileiro elas tenham apresentado baixa frequência, é provável que em CGU elas sejam frequentes, já que o CGU consiste de textos que expressam opiniões. Somando essas 865 formas às 1.187 selecionadas previamente, obtivemos um total de 2.052 palavras não acentuadas para serem excluídas do léxico, seguindo nossa estratégia

4. Resultados

Ao final de nossa análise, obtivemos: (1) Lista de 2.052 palavras não acentuadas que são relativamente muito menos frequentes que seus respectivos falsos homógrafos, objetivo de nossa pesquisa; (2) Lista de 616 palavras, com frequência em corpus, que perderam o acento após a reforma ortográfica; (3) Lista de 2.151 pares de falsos homógrafos verbais (passado mais-que-perfeito e futuro) que são difíceis de desambiguizar, inclusive para humanos; (4) Lista de 115 pares de falsos homógrafos verbais de diferentes categorias. (5) Lista de 104 pares de falsos homógrafos em que ambas as formas são frequentes.

Cada uma destas listas poderá ser utilizada para diferentes finalidades. As palavras da lista (1) serão excluídas do léxico do corretor. As palavras da lista (2) serão excluídas do léxico do corretor assim que a reforma ortográfica estiver concluída. A lista (3) pode servir de insumo para uma investigação da frequência dos tempos passado-mais-que-perfeito e futuro em corpus de CGU: se o futuro for significativamente mais frequente que o passado, podemos excluir as formas do passado-mais-que-perfeito para beneficiar a correção das formas de futuro em que estiverem faltando acentos. As listas (4) e (5) poderão ser usadas para selecionar sentenças em corpus que contenham ambas as formas, constituindo um corpus para treinamento e teste de corretores ortográficos que levem em conta o contexto.

5. Considerações Finais e Trabalhos Futuros

O conhecimento sobre o léxico de falsos homógrafos adquirido neste estudo nos permite hipotetizar que aqueles que têm grande diferença de frequência podem ser resolvidos com a estratégia aqui apresentada; outros (em especial os pares em que cada uma das formas pertence a uma categoria gramatical diferente) podem ser resolvidos com abordagens de correção ortográfica que levem em conta o contexto; e outros, ainda, provavelmente não serão resolvidos por nenhuma das duas abordagens, porque até para um humano seria difícil decidir qual a forma correta, mesmo com informações de contexto, como é o caso dos tempos verbais passado mais-que-perfeito e futuro.

A estratégia de adaptação do léxico para a finalidade de melhorar a correção ortográfica pode ser estendida. Palavras que apresentam frequência nula em corpus, mesmo que não sejam falsos homógrafos, provavelmente podem ser suprimidas do léxico sem prejuízo para o corretor. Isso pode melhorar o tempo computacional do sistema de correção e eliminar a chance de uma palavra não frequente ser sugerida como correção.

O ideal, aliás, seria coletar as frequências das palavras no próprio gênero textual que se pretende corrigir. Porém, há de se considerar, no caso de CGU, que a alta incidência de erros ortográficos torna as frequências menos confiáveis do que as apresentadas em um corpus de língua padrão (os quais também apresentam erros, como visto no Corpus Brasileiro).

Em estudo futuro, pretendemos investigar as palavras homófonas da língua portuguesa, como "consertar-concertar", "segmento-seguimento" e "viajem-viagem", as quais também são tema de erros em CGU que não são corrigidos, até o momento, por corretores ortográficos baseados em léxico.

Agradecimentos

Parte dos resultados apresentados neste artigo foram obtidos por meio da atividade de pesquisa no projeto "Processamento Semântico de Textos em Português Brasileiro", financiado pela Samsung Eletrônica da Amazônia Ltda, sob os termos da Lei Federal 8.248/91.

Referências

Avanço, L. V., Duran, M. S.; Nunes, M. G. V. (2014) Towards a Phonetic Brazilian Portuguese Spell Checker. TorPorEsp - Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish. Available at: http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Evento?id=755.).

Choudhury, M.; Thomas, M.; Mukherjee, A.; Basu, A.; Ganguly, N. (2007) How Difficult is it to Develop a Perfect Spell-checker? A Cross-linguistic Analysis through Complex Network Approach. In: TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pages 81–88. Rochester: Association for Computational Linguistics.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 1966.

Martins, B.; Silva, M. J. Spelling correction for search engine queries. Advances in Natural Language Processing. Springer Berlin Heidelberg, 2004. 372-383.

Muniz, M.C.M.; Nunes, M.G.V.; Laporte. E. (2005) "UNITEX-PB, a set of flexible language resources for Brazilian Portuguese", Proceedings of the Workshop on Technology of Information and Human Language (TIL), São Leopoldo (Brazil): Unisinos.

Pelizzoni, J. M. (2007). Preâmbulo ao aconselhamento ortográfico para o português do Brasil: uma releitura baseada em utilidade e conhecimento linguístico. (Tese de doutorado). PPG em Ciências da Computação. Universidade de São Paulo.

Russel, R. C. (1918) SOUNDEX. US patent 1261167 issued 1918-04-02.