O Acordo Ortográfico aplicado aos grafos e aos dicionários do Unitex[1]

(The Orthographic Agreement applied to Unitex graphs and dictionaries)

Nathalia Perussi Calcia

[1]Programa de Pós-Graduação em Linguística (PPGL) – Universidade Federal de São Carlos (UFSCar)

nathalia.perussi@gmail.com

Abstract: The Orthographic Agreement of 1990 proposes unifying the orthography of Portuguese, the language of the CPLP members (Community of Portuguese Language Countrie), namely, Brazil, Portugal, Angola, Cape Verde, São Tomé and Príncipe, East Timor, Guinea-Bissau, and Mozambique. Since part of the Portuguese lexicon was modified, its dictionaries need revising and adapting to the new rules. Our work contributes to the development of NLP (Natural Language Processing) enhanced tools, like Unitex. Unitex is a software developed by the Linguistics and Computing team of the Université Paris-Est Marne-la-Vallée (PAUMIER, 2002), which allows the research of regular expressions within large *corpora*, besides other functionalities. This software is characterized by its embedded large coverage electronic dictionaries that need constant revision for maintenance of such lexicons.

Keywords: Orthographic Agreement; Orthography; Computational Linguistics.

Resumo: O Acordo Ortográfico de 1990 propõe unificar a ortografia das línguas dos países membros da CPLP (Comunidade dos Países de Língua Portuguesa), ou seja, Brasil, Portugal, Angola, Cabo Verde, São Tomé e Príncipe, Timor-Leste, Guiné Bissau e Moçambique. Devido a isso, uma parte do léxico do português sofreu modificações e por isso necessita de uma revisão e adequação às novas regras. O presente trabalho se inscreve nos esforços para a constituição de ferramentas de PLN (Processamento de Linguagem Natural) mais aprimoradas, como o Unitex, um *software* desenvolvido pela equipe de Linguística e Informática da Université Paris-Est Marne-la-Vallée (PAUMIER, 2002), que permite, entre outras funcionalidades, a busca por expressões regulares em grandes *corpora*. Esse *software* se caracteriza por possuir dicionários eletrônicos de grande cobertura incorporados, que quando construídos, precisam ser constantemente revisados.

Palavras-chave: Acordo Ortográfico; Ortografia; Linguística Computacional.

Introdução

A construção de recursos para o PLN (Processamento de Linguagem Natural) tem crescido significantemente nos últimos anos e sua necessidade é reconhecida mundialmente tanto para pesquisadores em Linguística quanto para a Computação. Dentre esses recursos, destaca-se a construção de léxicos de grande envergadura tanto para a língua de um modo geral como para as chamadas linguagens de especialidade.

O papel do léxico nas tarefas de processamento automático do português se dá a partir da necessidade de sua descrição, seja a mais dependente de seu conhecimento ou ainda menos dependente, porém fundamental. Essa descrição deve levar em conta o pré-conhecimento de quem a utilizará, por exemplo, no caso de falantes nativos de uma língua é possível descrever a noção de gênero a partir das palavras *casa* e *cômodo*, res-

[1] Parte deste trabalho foi financiada com uma bolsa de Iniciação Tecnológica (PIBITI) do CNPq e pelo Dicionário Informal.

pectivamente feminino e masculino. Desse modo, o falante saberá distinguir e listar todas as palavras de gênero feminino e masculino. Quando passamos a tratar esse conteúdo em uma máquina, o exemplo anterior não se dá, pois não existe um pré-conhecimento que possa ser buscado para realizar essa tarefa, por esse motivo o léxico precisa ser descrito exaustivamente, ou seja, além das formas lematizadas é preciso levar em conta as formas flexionadas dos itens lexicais. Essa descrição é pertinente na elaboração de um dicionário eletrônico, chamado também por léxico computacional. Os dicionários eletrônicos são feitos especificamente para integrarem programas computacionais por apresentarem características como a abrangência, a flexibilidade e a sistematicidade. Esses léxicos são abrangentes pelo grande número de entradas lexicais que possibilitam; são flexíveis por serem adaptáveis em qualquer programa em que forem utilizados e recebem a característica de sistemáticos devido ao fato de serem feitos de maneira coerente.

Existem recursos computacionais que são caracterizados por apresentarem dicionários eletrônicos incorporados, o Unitex[2] (PAUMIER, 2002) é um deles. Trata-se de um ambiente de desenvolvimento linguístico que pode ser usado como um processador de *corpus* que permite, entre outras funcionalidades, a busca por expressões regulares em grande *corpora* de milhões de palavras em tempo real. Como já mencionado, esse *software* utiliza recursos linguísticos na forma de dicionários e gramáticas eletrônicas de grande cobertura para várias línguas (Espanhol, Inglês, Francês, Português, etc.) que descrevem as palavras simples e compostas, associando-as a um lema e a uma série de códigos gramaticais, semânticos e flexionais. Esses dicionários podem ser aplicados aos textos para a localização de padrões morfológicos, lexicais e sintáticos, remoção de ambiguidades e até mesmo para a etiquetação de palavras simples e compostas, possibilitando aos usuários a busca de ocorrências por categoria gramatical ou ainda pelo lema das entradas lexicais.

Uma vez construídos, os dicionários eletrônicos precisam de uma constante manutenção, seja por causa da introdução de neologismos, seja pela constante evolução da língua, ou, ainda, pelo recente caso da reforma da ortografia. Com o Acordo Ortográfico, uma parte do léxico do português sofreu modificações, consequentemente, uma parte do léxico dos dicionários eletrônicos do Unitex, também, devido a isso houve uma motivação para a realização deste trabalho, que visa ao aprimoramento e à adequação desse recurso.

Para contextualizar, o Acordo Ortográfico da Língua Portuguesa, assinado em Lisboa em 16 de dezembro de 1990, de fato entrou em vigor no Brasil a partir de 1 de janeiro de 2009, e tem sido objeto de várias adaptações e polêmicas. Segundo o governo brasileiro, sua utilização passará a ser obrigatória a partir de janeiro de 2016, porém neste momento ambas as grafias (antes e depois do Acordo Ortográfico) estão em uso no Brasil. Embora ele não elimine todas as diferenças ortográficas observadas nos países que possuem a Língua Portuguesa como idioma oficial, como Brasil, Portugal, Angola, São Tomé e Príncipe, Cabo Verde, Moçambique e Timor Leste, é um importante passo em direção à unificação ortográfica desses países.

Este artigo apresenta o processo de aprimoramento do dicionário eletrônico do português desenvolvido por Muniz (2004), que adaptou o léxico[3] construído para o

[2] Disponível em: <http://www-igm.univ-mlv.fr/~unitex/>. Acesso em: 4 mar. 2013.

[3] Esse léxico é o que dá suporte ao corretor ortográfico do Microsoft Word, versão brasileira.

ReGra[4] (JESUS; NUNES, 2000) – revisor gramatical desenvolvido no NILC (Núcleo Interinstitucional de Linguística Computacional) em parceria com a ITAUTEC – para o *software* Unitex. Para tanto, cabe destacar as principais mudanças ortográficas que ocorreram no Português Brasileiro, as quais serão a base para a adequação dos grafos de flexão nominal e adjetival e do dicionário de formas compostas flexionadas. Este artigo então traz informações a respeito das regras que regem o Acordo Ortográfico de 1990 e o modo em que serão aplicadas em informática, mais especificamente, nos dicionários e nos grafos da ferramenta linguístico-computacional em questão.

Metodologia

Além da dissertação de mestrado de Muniz (2004) e do *Manual de utilização do Unitex* (PAUMIER, 2002), foi realizada uma leitura minuciosa do texto oficial do Acordo Ortográfico de 1990, destinado aos países que possuem a língua portuguesa como idioma oficial (países membros da CPLP) e também das apresentações do Volp (Vocabulário Ortográfico da Língua Portuguesa), elaborado pela Academia Brasileira de Letras e do VOP (Vocabulário Ortográfico do Português), elaborado pelo Portal da Língua Portuguesa, que apresentam o vocabulário de mudança, essenciais para atestar os resultados obtidos durante a verificação das entradas revisadas.

Após a revisão teórica, partiu-se para a prática com a elaboração de um dicionário-piloto contendo as entradas típicas de cada grafo de flexão (nominal e adjetival) a partir das etiquetas que cada um recebeu em um trabalho feito anteriormente por Calcia, Carneiro e Rufo (2011).[5] Essas etiquetas servem como uma palavra-exemplo, ou seja, um modelo de entrada que flexiona da forma do grafo, no qual será explicitado no decorrer do texto. Em seguida, essas entradas foram flexionadas para o procedimento de verificação das formas recém-geradas. Após essa revisão, também foi levantado os problemas de flexão nos grafos, observados nessa fase. O último material a ser revisado foi o dicionário de formas compostas flexionadas (DELACF), o que sofreu mais modificações em decorrência do Acordo Ortográfico, uma vez que é o mais extenso em número de entradas.

Após realizadas todas as etapas, os resultados foram processados em forma de lista pelo *software* Lince, uma ferramenta que converte o conteúdo de textos e ficheiros para a grafia a ser introduzida após a reforma ortográfica. O resultado desse processo serviu para confirmar a veracidade dos resultados originados da revisão manual. A conclusão do trabalho se deu com a implementação dos novos grafos e das novas entradas ao *software* Unitex.

Este trabalho tem como princípio a proposta metodológica para o tratamento de língua natural adotada por Dias-da-Silva *et al.* (2006, p. 121), para quem a construção de um sistema, ferramenta ou recurso linguístico-computacional é composta de um conjunto de atividades agrupadas em três fases: a fase linguística, a fase representacional e a fase implementacional. As atividades da primeira fase são a descrição dos fatos linguísticos segundo um modelo teórico dado. Na fase representacional são estudados os modelos formais de representação para aquilo que foi descrito na fase anterior. Já a fase implemen-

[4] Disponível em: < http://nilc.icmc.usp.br/nilc/projects/regra.htm>. Acesso em: 25 jun. 2013.

[5] Resumo disponível em: http://gel.org.br/detalheResumo.php?trabalho=7637. Acesso em: 4 abr. 2013.

tacional trata da implementação do conhecimento linguístico aos sistemas de PLN. Cada uma dessas fases pode realimentar outra, de acordo com as necessidades que ocorram.

Análise e resultados

O Unitex possui modelos de flexão tanto para os substantivos quanto para os adjetivos, que são representados por meio de grafos de Autômatos Finitos, comuns na construção de compiladores. Os Autômatos Finitos (AF) são estruturas usadas para representar grandes dicionários eletrônicos. Por meio deles é possível ter acesso à palavra e aos atributos que pertencem a ela, como por exemplo, gênero, número e grau. Os AF são caracterizados por apresentarem um começo e um fim, como mostra o exemplo a seguir:

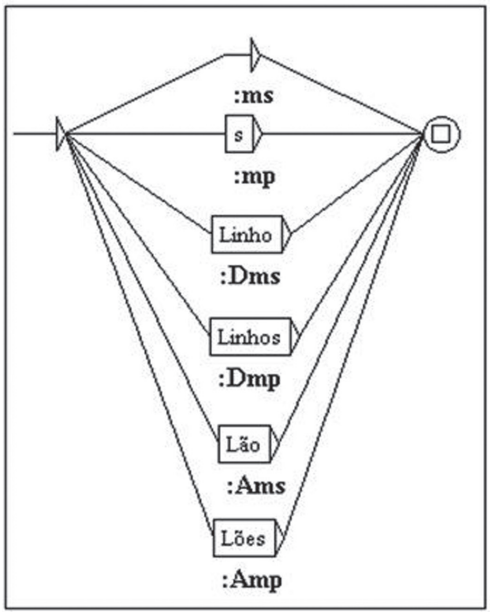

Figura 1. Grafo N001D026A01 para o Português do Brasil

A Figura 1 acima mostra um grafo que representa um modelo de entrada nominal do dicionário de formas lematizadas do Unitex. Ao ler uma forma canônica, o grafo gera as respectivas formas flexionadas. Dessa maneira, este grafo gerou as formas flexionadas em número (singular e plural) e as formas derivadas em grau (aumentativo e diminutivo). Para entender melhor seu funcionamento, é necessário realizar uma leitura da esquerda para a direita e de cima para baixo, desse modo, os caminhos percorridos descrevem os sufixos que podem ser acrescentados ou subtraídos à forma lematizada da entrada; e as saídas fornecem os códigos flexionais que são acrescentados à essa entrada, dependendo da sua flexão. Na Figura 1, seis caminhos são possíveis para esse tipo de entrada, assim, a partir do dicionário de formas lematizadas, o Autômato parte de uma entrada simples e produz as seguintes entradas no dicionário de formas flexionadas.

Entrada simples:

```
carro,N001D026A01
```

Entradas flexionadas:

```
carro,carro.N:ms
carros,carro.N:mp
carrinho,carro.N:Dms
carrinhos,carro.N:Dmp
carrão,carro.N:Mas
carrões,carro.N:Amp
```

Os grafos desenvolvidos por Muniz (2004) apresentavam uma dificuldade suplementar para os novos usuários do *software*, pois para a introdução de novas entradas em um dicionário era necessário percorrer a totalidade dos grafos de flexões nominais e adjetivais para encontrar a mais adequada àquela entrada. A partir disso, Calcia, Carneiro e Rufo (2011) estabeleceram um ponto de partida para facilitar essa tarefa, eliminando assim, essa dificuldade de utilização do *software*. A solução encontrada foi a inserção de uma entrada correspondente em cada um dos grafos, ou seja, uma etiqueta que contém a palavra-exemplo. Durante esse processo 634 grafos receberam uma palavra exemplo, 392 eram nominais e 242 adjetivais.

Adequação dos grafos de flexão nominal

Nesta etapa foi realizado um levantamento para verificar o alcance das novas regras ortográficas nos modelos de flexão nominal do Unitex. Em consequência do trabalho anterior, alguns grafos foram excluídos, restando ao todo 380 grafos referentes à flexão nominal. Esses grafos apresentam informações flexionais, ou seja, informações que descrevem o gênero (feminino e masculino) e o número (singular e plural), e eventualmente também descrevem informações de cunho derivacional, isto é, o grau (diminutivo e aumentativo).

Primeiramente, foi elaborado um dicionário-piloto com a entrada típica de cada grafo, ou seja, com a etiqueta fruto do trabalho realizado anteriormente. Em seguida, essas entradas foram flexionadas para o procedimento de verificação das formas que foram geradas. A partir do dicionário-piloto foram obtidas 2231 formas, que foram organizadas em uma tabela da seguinte maneira: forma flexionada, lema, classe gramatical, forma após o Acordo. O dicionário contendo todas as formas flexionadas está disponibilizado anexado a este trabalho.

Após a realização de uma revisão manual das 2231 entradas do dicionário-piloto, a lista obtida foi processada pelo *software* Lince, que converteu as formas da antiga para a atual ortografia do português brasileiro. Tal processamento foi pertinente, pois, em caso de algum erro ocasionado pela revisão manual das entradas, seria possível corrigi-lo, além disso, esse *software* atestou a veracidade das entradas corrigidas. As listas obtidas pelo Lince foram sistematicamente comparadas com as listas originais, procurando identificar as possíveis mudanças ocorridas.

Como consequência desse processo, o quadro abaixo mostra os resultados obtidos:

Quadro 1. Entradas nominais adequadas ao Acordo Ortográfico de 1990

FORMA FLEXIONADA	LEMA	CLASSE GRAMATICAL	FORMA APÓS O ACORDO
agüinha	água	N:Dfs	aguinha
agüinhas	água	N:Dfp	aguinhas
atéia	ateu	N:fs	ateia
atéias	ateu	N:fp	ateias
lingüinha	língua	N:Dfs	linguinha
lingüinhas	língua	N:Dfp	linguinhas
pêra	pêra	N:fs	pera

De acordo com o Quadro 1, as formas flexionadas adequadas ao Acordo Ortográfico foram: *aguinha, aguinhas, ateia, ateias, linguinha, linguinhas* e *pera*, respectivamente sete formas. E as regras aplicadas foram:

I) Eliminação do trema, sinal colocado sobre a letra *u* para indicar que ela deve ser pronunciada como *gue, gui, que, qui*;

II) Eliminação do acento dos ditongos abertos *éi* e *ói* das palavras paroxítonas, ou seja, que possuíam o acento tônico na penúltima sílaba;

III) Eliminação do acento que diferenciava os pares *pára/para, pêlo(s)/pelo(s), pólo(s)/polo(s)* e *pêra(s)/pera(s)*.

No Unitex os grafos são nomeados por meio de códigos. Esses códigos estão dispostos no programa de flexão *inflect*, que explora todos os caminhos da gramática de flexão e gera todas as formas flexionadas do dicionário Delas-PB. Então, os grafos que receberam modificações provenientes do Acordo Ortográfico foram:

(01) O grafo N101D114 é referente às entradas que flexionam como entrada *água*, ou seja, é possível usar este grafo para flexionar outras entradas que possuem as mesmas características flexionais desta, porém apenas a entrada *água* está presente no dicionário de formas simples do Unitex. Neste grafo, a principal mudança foi a eliminação do trema na forma diminutiva da entrada. Como o grafo apresentava as duas formas, com e sem o trema, retiramos apenas a forma ultrapassada.

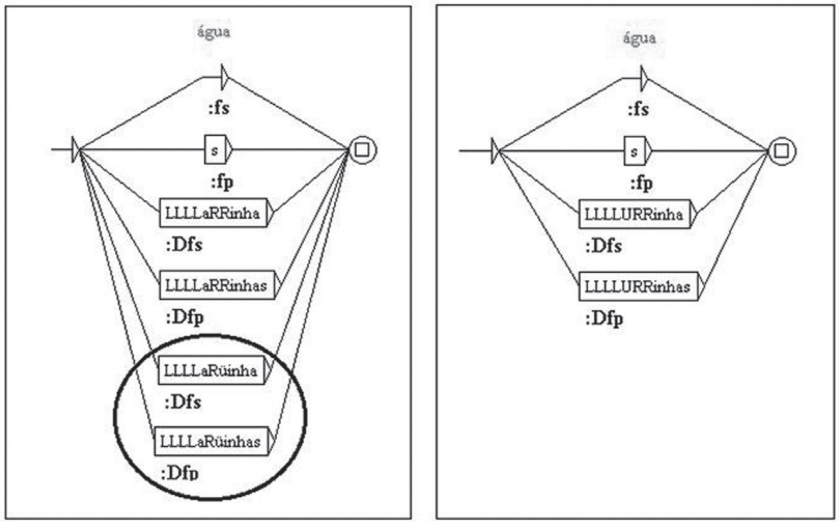

Figura 2. Antes e depois do grafo N001D026A01 para o Português do Brasil

(02) O grafo N114 é referente às entradas que flexionam como a entrada *pera,* porém apenas a entrada *pera* está presente no dicionário de formas simples do Unitex. Neste caso, foi preciso retirar o acento circunflexo que diferenciava os pares pera/pêra em sua forma lematizada e em sua forma flexionada em número.

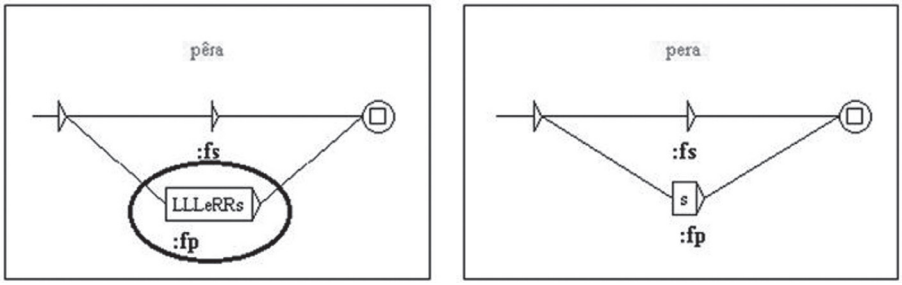

Figura 3. Antes e depois do grafo N114 para o Português do Brasil

(03) O grafo N211 é referente às entradas que flexionam como a entrada *ateu*, assim como *aqueu, caldeu, epigeu, europeu, filisteu, hebreu, pigmeu, pireneu* e *plebeu*. A alteração neste grafo foi feita ao retirar o acento agudo da forma em gênero feminino em singular e em plural.

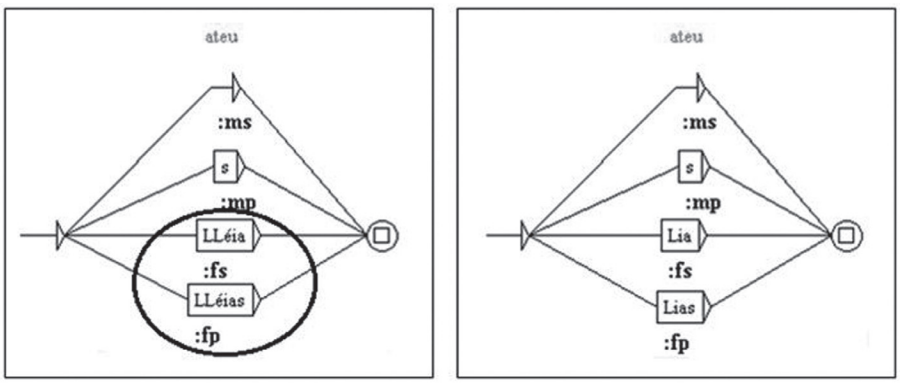

Figura 4. Antes e depois do grafo N211 para o Português do Brasil

(04) O grafo N301D063 é referente às entradas que flexionam como a entrada *língua*, porém apenas a entrada *língua* consta no dicionário de formas simples do Unitex. Nesse caso, foi retirado o trema que pertencia à forma diminutiva dos gêneros feminino e masculino e de número singular e plural.

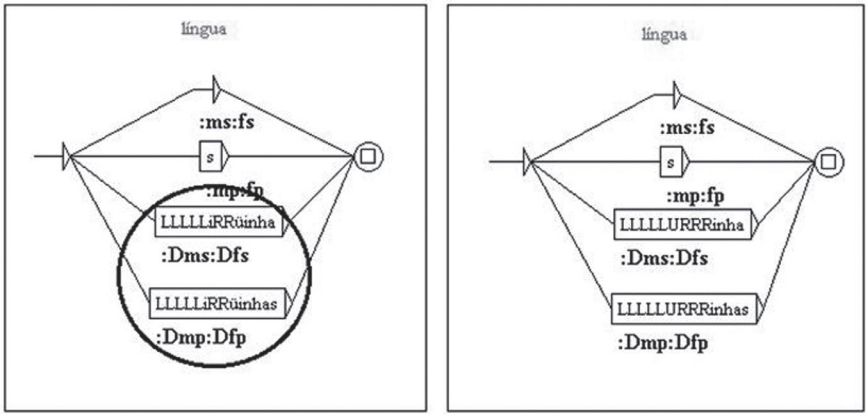

Figura 5. Antes e depois do grafo N301D063 para o Português do Brasil

Durante esse processo foram observados alguns equívocos: as formas diminutivas das palavras *avô e avó,* que são, respectivamente, *avozinho* e *avozinha*, aparecem no dicionário com o acento que provém de sua forma não flexionada*, *avôzinho* e **avózinha,* ou a *asteroide* como forma diminutiva da entrada *astro*.

Adequação dos grafos de flexão adjetival

Como já mencionado, os grafos nominais foram reduzidos devido ao trabalho feito por Calcia, Carneiro e Rufo (2011), o mesmo acontece com os grafos adjetivais, que passaram de 242 grafos para 233 grafos de flexão. O funcionamento dos grafos adjetivais é um pouco diferente dos nominais, isso se dá pelo fato dos adjetivos, além de descreverem o gênero, o número e o grau diminutivo e aumentativo, também podem descrever o grau em superlativo.

O procedimento para a adequação dos grafos adjetivais foi padrão, do mesmo modo feito para os grafos nominais. O dicionário-piloto possuía 1950 formas flexionadas e forma adequadas ao Acordo ortográfico 10 formas, que são referentes a 3 grafos de flexão, como mostra o quadro abaixo:

Quadro 2. Entradas adjetivais adequadas ao Acordo Ortográfico de 1**990**

FORMA FLEXIONADA	**LEMA**	**CLASSE GRAMATICAL**	FORMA APÓS O ACORDO
antiquíssima	antigo	A:Sfs	antiquíssima
antiqüíssimas	antigo	A:Sfp	antiquíssimas
antiqüíssimo	antigo	A:Sms	antiquíssimo
antiqüíssimos	antigo	A:Smp	antiquíssimos
européia	europeu	A:fs	europeia
européias	europeu	A:fp	europeias
brandiloqüentíssima	brandíloquo	A:Sfs	brandiloquentíssima
brandiloqüentíssimas	brandíloquo	A:Sfp	brandiloquentíssimas
brandiloqüentíssimo	brandíloquo	A:Sms	brandiloquentíssimo
brandiloqüentíssimos	brandíloquo	A:Smp	brandiloquentíssimos

Diferentemente da etapa anterior, foram usadas somente duas regras ortográficas na adequação dessas formas:

I) Eliminação do trema, sinal colocado sobre a letra *u* para indicar que ela deve ser pronunciada como *gue, gui, que, qui*;

II) Eliminação do acento dos ditongos abertos *éi* e *ói* das palavras paroxítonas, ou seja, palavras que possuíam o acento tônico na penúltima sílaba.

Diante das mudanças ocasionadas pelo atual Acordo Ortográfico, foram modificados os seguintes grafos de flexão adjetival:

(05) O grafo A201S61 é referente às formas que flexionam como *antigo*. No dicionário Delas consta apenas essa forma para o grafo. Nesse grafo estão representadas as formas masculino e feminino; singular e plural e também as formas derivadas no superlativo absoluto sintético. Para o superlativo o grafo apresenta duas possibilidades de derivação: *antiguíssimo e antiquíssimo*. Nesse caso, a mudança ortográfica está na forma superlativa à direita do grafo, onde foi retirado o trema.

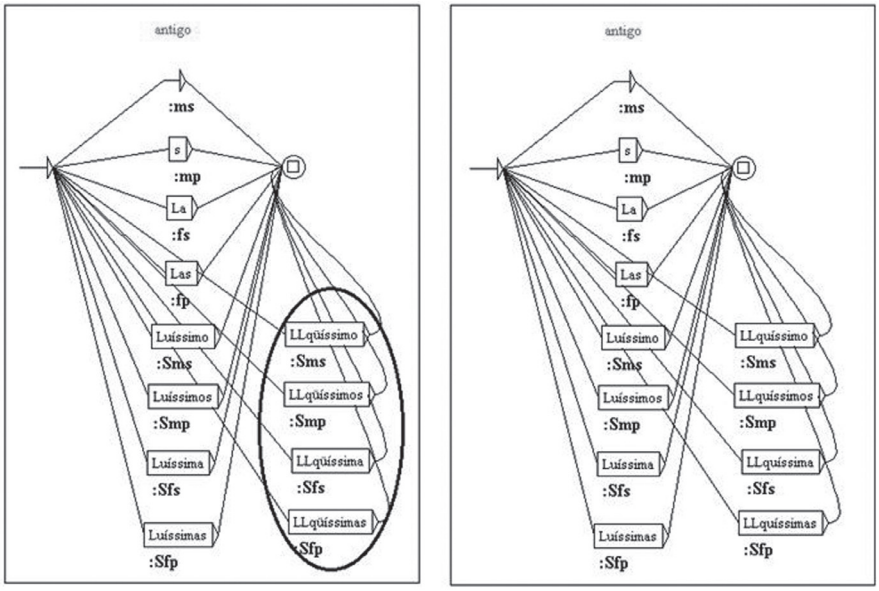

Figura 6. Antes e depois do grafo. A201S61 para o Português do Brasil

(06) O próximo grafo modificado é o que recebe o código A211, referente às formas que flexionam como *europeu*, dessa maneira as formas que também podem ser flexionadas por esse grafo são: *pigmeu, ateu, caldeu, epigeu, filisteu, hebreu, plebeu* e *pireneu*. Neste caso, a mudança ocorreu nas formas em feminino, tanto no singular como no plural, pois segundo o Acordo Ortográfico, não é mais usado acento nos ditongos abertos *ei*, como em *europeia*, *plebeia*, *etc*.

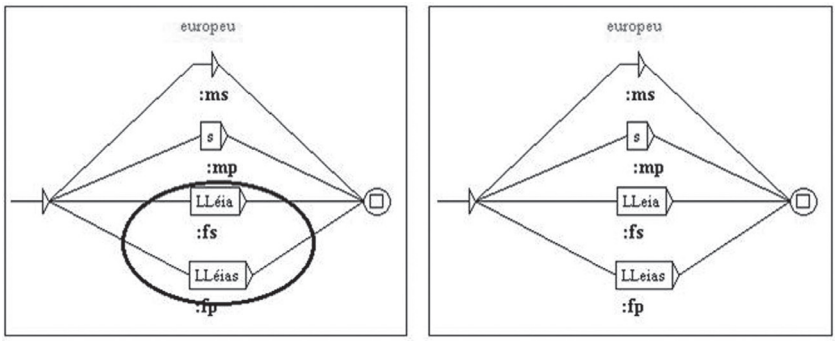

Figura 7. Antes e depois do grafo A211 para o Português do Brasil

(07) O grafo A201S17 é referente às formas que flexionam como *brandíloquo,* assim como *blandíloquo, grandíloquo* e *magníloquo.* E a modificação está também nas formas superlativas, uma vez que é necessário subtrair o trema nessas formas.

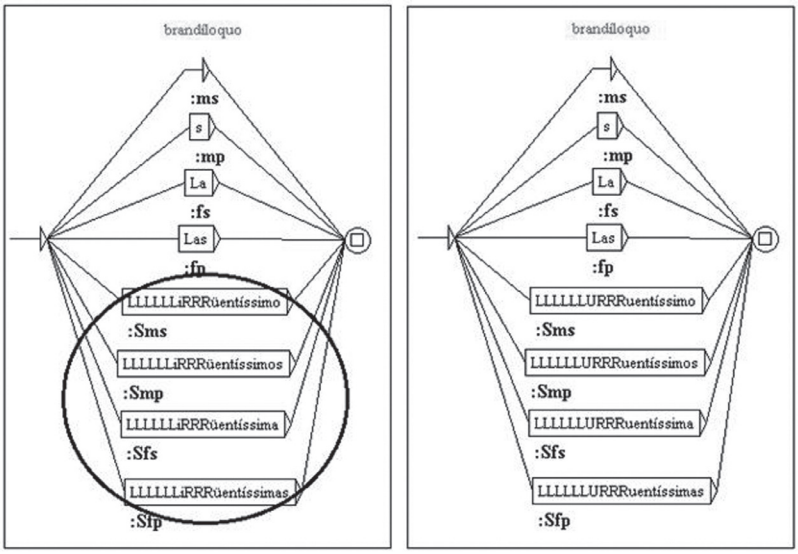

Figura 8. Antes e depois do grafo A201S17 para o Português do Brasil

Adequação do dicionário de formas compostas flexionadas – DELACF

As palavras compostas são caracterizadas por possuírem mais de um radical, mais precisamente são unidades lexicais formadas da combinação de palavras simples, possíveis com qualquer significante de qualquer língua. Como as palavras simples, as palavras compostas também podem receber o acréscimo de desinências, de modo que exprima a categoria gramatical de flexão. Os nomes, adjetivos e pronomes podem receber as flexões de gênero e número e os verbos podem receber a flexão de número, pessoa, tempo e modo.

O dicionário de formas compostas flexionadas (DELACF) foi resultado da flexão gerada a partir do dicionário de formas compostas (DELAC) que é um tipo de dicionário DELA de palavras compostas (SILBERZTEIN, 1990). Vale ressaltar que as palavras compostas em português são formadas a partir de dois ou mais radicais que podem se juntar por meio de três processos: juntos por aglutinação, juntos por justaposição e separados com ou sem hífen.

O DELACF, além da informação da classe gramatical a que a palavra pertence, necessita da informação da classificação dos constituintes da palavra composta. Por exemplo, em *lateral esquerdo,lateral esquerdo.N+NA:ms* os constituintes são um substantivo (*N*) e um adjetivo (*A*), onde *NA* representa substantivo + adjetivo. Já em *rabos-de-tatu,rabo-de-tatu.N+NDN:mp* os constituintes são um substantivo, uma preposição e outro substantivo, onde *NDN* representa substantivo + preposição (*de*) + substantivo. Na época em que o dicionário foi confeccionado para o Português Brasileiro, a informação dos constituintes das palavras compostas não estava presente no dicionário do ReGra, dessa forma, foi feita uma classificação manual por linguistas. Segundo Muniz (2004), durante essa fase verificou-se se as palavras compostas eram realmente separadas ou não pelo hífen, uma vez que o DELACF é basicamente composto por formas hifenizadas. Após a adequação dessas formas, é pertinente a criação de um dicionário de formas compostas não hifenizadas, para que assim, possa contemplar todas as palavras compostas do Português Brasileiro.

O conteúdo do DELACF foi disposto em um modelo semelhante ao do dicionário-piloto confeccionado para os substantivos e adjetivos, contendo o lema das entradas compostas, a forma flexionada em número, a classe gramatical e seus constituintes e a forma após o Acordo Ortográfico de 1990. O dicionário contém 4078 entradas compostas flexionadas e o procedimento para a revisão foi o mesmo adotado nas etapas anteriores. Desse modo, foram verificadas manualmente cada entrada e adequadas àquelas que sofreram modificações. Após esse processo, 791 entradas flexionadas foram adequadas à atual ortografia do português.

Devido ao grande número de entradas modificadas, apenas os principais casos de palavras formadas por prefixação e palavras formadas por composição, serão apresentados a seguir:

(08) Quando o prefixo termina em vogal e o segundo elemento começa pelas letras *r* ou *s*, não se usa mais o hífen e essas letras são duplicadas: *anterrosto, antessala, autorretrato, autosserviço, suprassumo, ultrassom,* entre outros;

(09) Palavras que são formadas pelos prefixos auto, co, contra, extra, infra, intra, pseudo, semi, supra, não associados ao exemplo anterior, são escritas sem hífen: autoacusação, autoadesivo, coautor, contraindicação, extraescolar, infraestrutura, intraocular, pseudoesfera, semioculto, supraexcitação, entre outros;

(10) Palavras formadas por substantivo + determinante + substantivo não são mais escritas com hífen: açoite de rio, água de cana, baba de moça, balaio de gato, dor de cotovelo, jardim de inverno, jogo da velha, entre outros. Há algumas exceções como: *açaí-do-pará, água-de-colônia, cachorro-do-mato, canário-da-terra, castanha-do-pará,* entre outros*;*

(11) Palavras formadas por verbo + conjunção + verbo não recebem mais o hífen: *come e dorme, ir e vir, leva e traz, vai e vem*. Exceção: abre-e-fecha;

(12) Palavras formadas por verbo + substantivo não são mais hifenizadas: *mandachuva, micareme, paraquedas, paraquedista*. Exceções: *para-brisa, para-choque, para-lama, para-raios;*

Há palavras compostas que se diferem semanticamente pela sua grafia, ou seja, quando escritas com hífen possuem um significado e quando escritas sem hífen possuem outro significado, como em:

(13) Bico de papagaio (nome de doença) e bico-de-papagaio (tipo de planta);

Olho de boi (selo postal) e olho-de-boi (espécie de peixe);

Pomo de adão (tireoide acentuada) e pomo-de-adão (tipo de árvore);

Vassoura de bruxa (doença dos cacaueiros) e vassoura-de-bruxa (tipo de fungo);

Entre outros.

Além das modificações ortográficas, constatou-se a presença de palavras compostas de origem estrangeira no dicionário, por exemplo: *banana-split*, *bang-bang*, *best-seller*, *blank verse*, entre outras. Essas palavras não foram adequadas ao Acordo Ortográfico por não serem palavras pertencentes ao léxico do Português.

Considerações finais

Conforme apresentado no início deste artigo, a adequação do léxico nas ferramentas linguístico-computacionais é de extrema importância para manter esses recursos sempre atualizados. Pelo fato de o *software* Unitex possuir dicionários in*corpora*dos a sua escolha para ser objeto de análise deste trabalho foi pertinente. Trabalhos como este são um importante passo em direção à unificação ortográfica que o Acordo de 1990 propõe.

Sabemos que as palavras apresentadas neste trabalho são apenas uma pequena parcela do léxico do português, mas é considerada uma parte significativa, pois abrange as principais bases contidas no texto oficial do Acordo Ortográfico. A partir disso, é possível perceber que a adequação dos grafos e dicionários não foi uma tarefa fácil, pois há muitas regras que apresentam exceções que não estão atestadas nos vocabulários ortográficos existentes, ou seja, que não contemplam o léxico contido no Volp e no VOP. Nesses casos, foi necessário percorrer todas as regras e ver qual era a mais adequada àquela entrada. Além disso, foi possível perceber que as palavras mais afetadas pelo Acordo, de fato, foram as palavras compostas, tendo em vista o grande número de modificações que o dicionário de formas compostas flexionas recebeu.

É importante ressaltar que não foram tratados neste trabalho a adequação dos verbos implementados na ferramenta Unitex. Mas é relevante mencionar que o emprego dos verbos não foi pelo léxico do ReGra, que continha 6.672 verbos, mas sim pelos verbos obtidos por Oto Vale (1990) (14.284 verbos), verbos estes que já estavam associados às regras de flexão. A revisão dos modelos de flexão verbal é objeto de um trabalho em curso (KUCINSKAS; VALE, 2014).

De modo geral, com os resultados obtidos foi possível gerar um dicionário atualizado de formas simples flexionadas e um novo dicionário de formas compostas flexionadas, que serão implementados na ferramenta Unitex.

Pode-se dizer, ainda, que este artigo apresentou um trabalho de cunho tecnológico, mas refletiu as dificuldades pelas quais o sistema de escrita adotado pela língua portuguesa passa, sofrendo constantes modificações impostas, em grande parte, por uma política de línguas. Ainda, mostrou as dificuldades técnicas em "aprender" e aplicar uma nova grafia que não foi ensinada para os adolescentes e adultos de hoje, fato que também pode explicar os inúmeros adiamentos da obrigatoriedade do Acordo Ortográfico de 1990.

Este estudo, portanto, é mais um passo importante para manter o *software* atualizado, visando à utilização por novos pesquisadores da área de PLN para a análise de fenômenos em vários níveis da língua e para a motivação de construção de novos recursos linguístico-computacionais.

REFERÊNCIAS

ALMEIDA, G. M. B.; FERREIRA, J. P.; CORREIA, M.; OLIVEIRA, G. M. Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa. *Estudos Linguísticos*, São Paulo, n. 42, v. 1, p. 204-215, 2013.

BRASIL. Decreto n. 12.605, de 3 de abril de 2012. Determina o emprego obrigatório da flexão de gênero para nomear profissão ou grau em diplomas. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12605.htm> Acesso em: 20 ago. 2013.

CALCIA, N. P.; CARNEIRO, A. S.; RUFO, A. Revisão dos modelos de flexão do unitex: passos iniciais. In: SEMINÁRIO DO GEL, 59, 2011, Programação... Bauru (SP): GEL, 2011. Disponível em: <http://gel.org.br/detalheResumo.php?trabalho=7637> Acesso em: 4 abr. 2013.

DIAS-DA-SILVA, B. C.; MONTILHA, G.; RINO, L. H. M.; SPECIA, L.; NUNES, M. G. V.; OLIVEIRA JR., O. N.; MARTINS, R. T.; PARDO, T. A. S. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações. Série de Relatórios Técnicos do NILC, NILC-TR-07-10. São Carlos-SP, Agosto, 2006. 121p

JESUS, M.A.C.; NUNES, M.G.V. *Representação de léxicos através de autômatos finitos.* Relatórios Técnicos do ICMC-USP, 110 (NILC-TR-00-5). 2000.

KUCINSKAS, A. B.; VALE, O. A. Revisão dos modelos de flexão verbal do Unitex-PB. Relatório Técnico, 2014 (no prelo).

MUNIZ, M. C. M. A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. 72f. 2004. Dissertação (Mestrado) – Instituto de Ciências Matemáticas de São Carlos, USP, 2004

PAUMIER, S. Unitex 3.0 User Manual. Université Paris-Est Marne-La-Vallée, 2002. Disponível em: < http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.0.pdf> Acesso em: 4 mar. 2013.

SILBERZTEIN, M. Le dictionnaire électronique des mots composés, langue française. *Dictionnaires électroniques du français*, n. 87, p. 71-83, 1990.

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